STRUCTURAL PROGNOSTIC EVENT MODELING FOR MULTIMODAL CANCER SURVIVAL ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

The integration of histology images and gene profiles has shown great promise for improving survival prediction in cancer. However, current approaches often struggle to model intra- and inter-modal interactions efficiently and effectively due to the high dimensionality and complexity of the inputs. A major challenge is capturing critical prognostic events that, though few, underlie the complexity of the observed inputs and largely determine patient outcomes. These events—manifested as high-level structural signals such as spatial histologic patterns or pathway co-activations—are typically sparse, patient-specific, and unannotated, making them inherently difficult to uncover. To address this, we propose **SlotSPE**, a slot-based framework for structural prognostic event modeling. Specifically, inspired by the principle of factorial coding, we compress each patient's multimodal inputs into compact, modality-specific sets of mutually distinctive slots using slot attention. By leveraging these slot representations as encodings for prognostic events, our framework enables both efficient and effective modeling of complex intra- and inter-modal interactions, while also facilitating seamless incorporation of biological priors that enhance prognostic relevance. Extensive experiments on ten cancer benchmarks show that SlotSPE outperforms existing methods in 8 out of 10 cohorts, achieving an overall improvement of 2.9%. It remains robust under missing genomic data and delivers markedly improved interpretability through structured event decomposition.

1 INTRODUCTION

Cancer survival analysis aims to estimate patient mortality risk or survival time, serving as a key tool for personalized prognosis and precision oncology, as it helps quantify outcomes, identify prognostic factors, and guide individualized treatment decisions (Cox & Reid, 2004; Jenkins, 2005; Wiegrebe et al., 2024). Whole-slide images (WSIs) and genomics are two widely-used tools in cancer survival analysis, as they capture complementary prognostic signals: WSIs reveal how cancer manifests morphologically in tissue, while genomics uncovers its underlying molecular drivers. Their integration offers the opportunity for more accurate, robust, and biologically grounded survival prediction, which has in turn sparked growing interest in multimodal learning approaches for survival analysis (Chen et al., 2021b; Xu & Chen, 2023; Jaume et al., 2024; Zhou et al., 2025).

Effective and efficient multimodal learning for survival analysis faces a fundamental challenge: although patient outcomes are often driven by a small number of critical events (Vogelstein et al., 2013; Hosseini et al., 2024), WSI and genomic signals are rather complex and high-dimensional. High-resolution WSIs can exceed $10^5 \times 10^5$ pixels (Song et al., 2023) and genomics involve thousands of genes (Győrffy, 2021). These events manifest as meaningful structural associations within WSI and genomic signals—e.g., in histopathology, immune activity and tumor aggressiveness (Chan et al., 2024; Liang et al., 2025) are reflected in the density and proximity of the pixel organizations of tumor-infiltrating lymphocytes and cancer nests, while in genomics, the tumor-promoting effects emerge from the co-activation of PI3K/AKT and MAPK pathways (Glaviano et al., 2023). However, learning to uncover such event-related patterns structures from the input signals is difficult for machines: these high-level associations are typically sparse, patient-specific, and unannotated. Compounding the challenge, the enormous scale of input units—millions of WSI pixels and thousands of genomic features—creates massive self-interactions within each modality, further amplified by the need for cross-modal alignment, rendering direct computation prohibitively expensive.

Existing attempts, e.g., (Ilse et al., 2018; Shao et al., 2021; Chen et al., 2021b; Xu & Chen, 2023; Jaume et al., 2024), fall short of capturing high-level prognostic associations. The most relevant studies are prototype-based methods (Zhang et al., 2024; Song et al., 2024), which employ fixed prototype representations and therefore lack adaptability to individual patients. For example, PIBD (Zhang et al., 2024) approximates a prototypical distribution for each risk group via information bottleneck (Tishby et al., 2000; Alemi et al., 2016). Yet, these prototypes are static-once trained, the same set is applied to all patients. MMP (Song et al., 2024) takes a different route by adopting Gaussian mixture model (GMM) (Bishop & Nasrabadi, 2006) to define prototypes for each WSI. However, the image-wise prototypes are unlearnable (i.e., fixed) during training, making end-to-end optimization suboptimal.

In this paper, we introduce SlotSPE, a slot-based framework for structural prognostic event modeling. Inspired by factorial coding (Schmidhuber, 1992; Higgins et al., 2017; Greff et al., 2020), our method compresses high-dimensional multimodal inputs into a compact set of semantic slots using a slot attention module (Locatello et al., 2020), where each slot corresponds to a latent prognostic event. Unlike fixed prototypes, these slots are dynamically instantiated for each patient, allowing the model to capture individualized prognostic patterns while preserving a concise representation. To reduce slot redundancy and encourage slots to capture distinct prognostic events, we introduce selective slot activation through a Mixture-of-Experts–style decoder (Hampshire & Waibel, 1989; Jacobs et al., 1991; Jordan & Jacobs, 1994; Shazeer et al., 2017), which activates only the most predictive slots for each patient. Beyond compression, we further model cross-modal interactions by incorporating a biological prior: histopathological phenotypes often reflect underlying molecular events, revealing molecular-morphology mappings (Wang et al., 2021; Liu et al., 2025b). Guided by this insight, we design a cross-modal reconstruction task where omics-derived slots are encouraged to predict gene expression from histology images. This alignment not only enforces biologically meaningful interactions across modalities but also improves robustness in missing genomic data.

The key contributions are: (1) A structural prognostic event modeling framework for multimodal survival prediction that effectively and efficiently extracts patient-specific events; (2) A cross-modal reconstruction that encodes biological priors to enhance alignment and robustness; (3) Extensive evaluation on ten cancer cohorts demonstrating the superiority and robustness; and (4) Enhanced interpretability through the structured decoupling ability of the proposed representation.

2 RELATED WORKS

Single Modality. The advent of deep learning marks a paradigm shift in cancer prognosis, where artificial neural networks (NN) can uncover prognostic patterns from radiological (Aerts et al., 2014; Lou et al., 2019; Xu et al., 2019; Yao et al., 2021), pathological (Ciresan et al., 2012; Cireşan et al., 2013), and molecular data (Chaudhary et al., 2018; Ching et al., 2018) that were previously inaccessible to conventional methods. In histopathology, whole-slide images (WSIs) under the multiple instance learning (MIL) (Dietterich et al., 1997; Maron & Lozano-Pérez, 1997) paradigm—where a slide is represented as a bag of patches with only slide- or patient-level survival labels—have become standard (Hou et al., 2016; Ilse et al., 2018; Shao et al., 2021; Lu et al., 2021). Recent MIL approaches address the challenge of abundant patches by using clustering (e.g. k-means) (Yao et al., 2020; Vu et al., 2023; Yang et al., 2024; Zhou et al., 2024), graph models (Li et al., 2018; Chen et al., 2021a), and hierarchical WSI representations (Shao et al., 2023). Meanwhile, genomic methods moved from simple MLP or shallow nets on high-dimensional gene profiles (Haykin, 1998; Klambauer et al., 2017) to pathway-based representations that group genes into biologically meaningful sets for better interpretability (Zhao et al., 2021; Hou et al., 2023; Wang et al., 2024). Building on these advances, our method preserves the MIL standard for histopathology, adopts pathway-level genomic features.

Multiple Modality. Multimodal approaches, especially the integration of histopathology and genomics, have attracted increasing attention because combining morphological and molecular information often yields better prognostic performance than either modality alone (Chen et al., 2020; 2021b; Yuan et al., 2025). Typical strategies include late fusion (e.g., concatenation or Kronecker product) (Chen et al., 2020; Ding et al., 2023; Volinsky-Fremond et al., 2024; Zhao et al., 2025; Luo et al., 2025), and early fusion via co-attention mechanisms operating at the patch-gene level (Chen et al., 2021b; Xu & Chen, 2023; Zhou & Chen, 2023; Jaume et al., 2024; Yuan et al., 2025; Wang et al., 2025; Zhou et al., 2025). Since inputs are complex and high-dimensional, recent works have

adopted adaptive token selection and aggregation (Zhang et al., 2025) or compact representations by compressing WSI and pathway data via prototypes (Liu et al., 2025a; Song et al., 2024; Zhang et al., 2024). The former, however, lacks modeling of interpretable structural associations; the latter fails to address sparsity and patient-specific event variation, relying instead on fixed prototypes.

Structural Event Modeling. A key challenge in multimodal survival analysis lies in handling the high-dimensional histology and omics inputs, which not only incurs substantial computational costs but also complicates modeling of intricate intra-/inter-modal interactions. An important insight is that patient outcomes are often driven by a small number of critical events (Vogelstein et al., 2013). This motivates approaches that aim to uncover the underlying event-related structures within complex inputs. For examples, PIBD (Zhang et al., 2024) and MMP (Song et al., 2024) attempt to derive structured representations from complex multimodal data. However, relying on fixed prototypes, these methods cannot dynamically capture the sparse and patient-specific structural events that often determine outcomes. From the perspective of machine learning, this challenge can be framed as a problem of factorial coding, a well-established paradigm concerned with decomposing and compressing high-dimensional data into disentangled latent factors (Schmidhuber, 1992; Tipping & Bishop, 1999; Higgins et al., 2017; Kim & Mnih, 2018). Building on this connection, our framework advances beyond static prototypes by introducing learnable slots through slot attention (Locatello et al., 2020), modeling prognostic events that are selectively activated per patient and aligned via biologically guided cross-modal reconstruction. This design enables a structured decomposition of survival-related events, binding multimodal prognostic signals to interpretable representations.

3 METHOD

Survival prediction is the task of estimating a patient risk of an event (e.g., death) given the pathological observations (e.g. WSIs and genomic) of the patient. In this paper, we tackle the problem of learning such hazard predictor from the multi-modal data. Formally, let the hazard prediction function be denoted as h_t, representing the probability of death occurring at time t (i.e., survival up to time t). The model is trained using the negative log-likelihood (NLL) loss (Zadeh & Schmid, 2020), denoted as $\mathcal{L}_{\mathrm{surv}}$ (see Appendix B.1 for details). Each patient sample is represented as $\left(\mathbf{X}_h^{(i)}, \mathbf{X}_g^{(i)}, t^{(i)}, c^{(i)}\right)$, where i indexes the patient, $\mathbf{X}_h^{(i)}$ denotes the pathology data, $\mathbf{X}_g^{(i)}$ denotes the genomic profile, $t^{(i)} \in \{1, \ldots, N_t\}$, and $c^{(i)} \in \{0, 1\}$ is the censoring indicator ($c^{(i)} = 0$ for an observed event, $c^{(i)} = 1$ for right-censoring). Operationally, following standard practices in computational pathology (Chen et al., 2021b; Jaume et al., 2024), WSI and genomic inputs are pre-processed by dedicated image and gene encoders and represented as bags of instances: $\mathbf{X}_h^{(i)} = \left\{\mathbf{x}_{h,j}^{(i)} \in \mathbb{R}^d\right\}_{j=1}^{M_h}$ and $\mathbf{X}_g^{(i)} = \left\{\mathbf{x}_{g,j}^{(i)} \in \mathbb{R}^d\right\}_{j=1}^{M_g}$, where M_h and M_g denote the numbers of WSI patches and biological pathways, respectively, and d is the feature dimension. When modeled with state-of-the-art transformers like prior works (Xu & Chen, 2023; Zhou & Chen, 2023; Jaume et al., 2024), the resulting intra- and inter-modal interactions scale quadratically as $O((M_h + M_g)^2)$, creating a major challenge for both efficacy and efficiency.

In this section, we present **SlotSPE**, a novel **Slot**-based **St**urctual **P**rognostic **E**vent modeling framework for multimodal cancer survival analysis (Figure. 1).

3.1 SLOT-BASED INFORMATION COMPRESSION

Inspired by factorial coding (Schmidhuber, 1992; Higgins et al., 2017), which aims to disentangle data into independent latent factors, we perform structural decomposition for two bags of histology and genomic instances, i.e., $\mathbf{X}_h^{(i)}$ and $\mathbf{X}_g^{(i)}$, using a slot attention module (Locatello et al., 2020). This allows us to compress the large input sets into much smaller slot-based representations, significantly reducing computational overhead. These slots are then decoded using a MoE-style decoder that selectively activates patient-specific slots, promoting both sparsity and competition among them.

Slot Attention. Slot attention (Locatello et al., 2020) learns to encode distinct structural patterns present in the input into a finite set of latent vectors, i.e., the "*slots*". Since WSI and genomic data inherently contain massive redundancy (Zhang et al., 2024; Song et al., 2024), it is desirable to aggregate them into a smaller set of compact representations while preserving their semantic
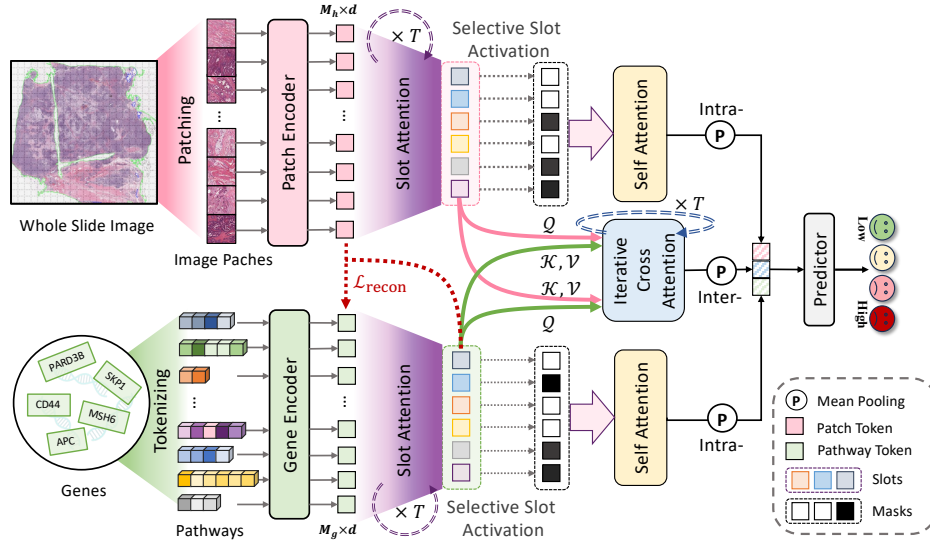
Figure 1: **Framework of SlotSPE**. Histology and gene features are extracted into bag structures, then compressed into slots via slot attention. Selective slot activation enforces sparsity and mutual competition, while a biologically guided cross-modal reconstruction aligns modalities. Finally, slot interactions are modeled using self- and cross-attention to predict survival.

content. This makes slot attention particularly well-suited for our task, where the resulting compact representations can be interpreted as latent prognostic events modeled by slots.

We re-organize the input features for both modalities as matrices, $\mathbf{X} \in \mathbb{R}^{M \times d}$. With a set of S slots, organized as $\mathbf{S}^{(0)} \in \mathbb{R}^{S \times d_{\text{slot}}}$, initialized from some learnable distributions, the slot attention compresses $\mathbf{X}$ into S slots through T iterations of cross-attention-based routing. At each iteration $\tau \in \{0, \ldots, T-1\}$, the slots $\mathbf{S}^{(\tau)}$ attend to the input $\mathbf{X}$ to produce updated representations, which are then refined using a multilayer perceptron (MLP). Slot updates across iterations are applied recurrently using a recurrent neural network (RNNs) (Hochreiter & Schmidhuber, 1997; Cho et al., 2014) to preserve temporal consistency and iterative refinement. Mathematically

$$\boldsymbol{\alpha}_j^{(\tau)} = \underset{\mathbf{S}^{(\tau)}}{\text{softmax}}\left(\frac{\mathcal{Q}(\mathbf{S}^{(\tau)})\,\mathcal{K}(\mathbf{x}_j)}{\sqrt{d_{\text{slot}}}}\right) \in \mathbb{R}^S, \quad \mathbf{u}_k^{(\tau)} = \sum_{j=1}^{M} \alpha_{kj}^{(\tau)}\,\mathcal{V}(\mathbf{x}_j),$$

$$\mathbf{s}_k^{(\tau+1)} = \text{RNN}\big(\mathbf{s}_k^{(\tau)}, \mathbf{u}_k^{(\tau)}\big), \quad \mathbf{s}_k^{(\tau+1)} \leftarrow \mathbf{s}_k^{(\tau+1)} + \text{MLP}\big(\mathbf{s}_k^{(\tau+1)}\big),$$

(1)

where $\mathcal{Q}(\cdot)$, $\mathcal{K}(\cdot)$, and $\mathcal{V}(\cdot)$ are linear projections into query, key, and value spaces. Note that the softmax is applied over S (slots) for each j, enforcing competition among slots and enabling a structured decomposition. Through iterative updates, the learnable bottleneck maps $\mathbb{R}^{M \times d} \rightarrow \mathbb{R}^{S \times d_{\text{slot}}}$ with $S \ll M$, yielding $\mathbf{S}^{(T)}$ as the final set of slots (simply as $\mathbf{S}$). The complexity of slot attentions for both histology and genomic instances will be $O(S_g * M_g + S_h * M_h)$.

Selective Slot Activation. In our task, to avoid slots redundantly attend to similar structures or collapse onto spurious and prognostic irrelevant patterns, we propose *selective slot activation* (Figure 2(a)): a Mixture-of-Experts (MoE)–style decoder with a learnable gating mechanism that sparsely activates the most predictive slots per patient, encouraging specialization and improving discriminative performance across the cohort.

Conceptually, each slot is treated as an "expert" potentially encoding a critical prognostic event. From the slot set $\mathbf{S}$, we derive per-slot predictions by projecting them into class logits $\boldsymbol{\ell}_k \in \mathbb{R}^{N_t}$ via an MLP predictor, obtaining S candidate predictions. A lightweight gating function ϕ predict retention score r_k for each slot $\mathbf{s}_k$:

$$r_k = \phi(\mathbf{s}_k), \qquad \phi : \mathbb{R}^D \rightarrow \mathbb{R}. \tag{2}$$

Then the top-K slots are selected using the Gumbel-Top-K trick (Gumbel, 1954; Maddison et al., 2014; Kool et al., 2019) with Straight-Through Estimation (ST) (Jang et al., 2016), producing a
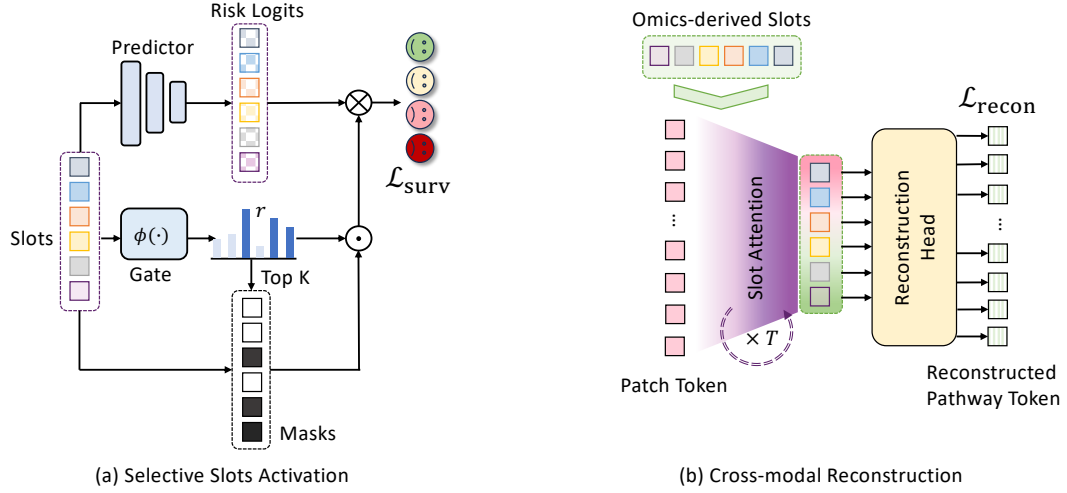
Figure 2: Detailed component structure of SlotSPE. (a) Selective slot activation: a Mixture-of-Experts(MoE)–style decoder sparsely activates only the most predictive slots. (b) Cross-modal reconstruction: guided by biological priors, omics-derived slots are aligned with histology by predicting pathway-level gene expression from WSIs.

differentiable K-hot mask $\hat{\mathbf{G}} \in \{0,1\}^S$ (see Appendix B.2 for details). The selected slots are reweighted by masking and renormalizing the softmax weights:

$$w_k = \frac{\tilde{w}_k \cdot \hat{G}_k}{\sum_{s=1}^{S} \tilde{w}_s \cdot \hat{G}_s}, \qquad \tilde{\mathbf{w}} = \mathrm{softmax}(\mathbf{r}) \in \mathbb{R}^S. \tag{3}$$

Finally, patient-level predictions are obtained as a gated mixture of slot logits:

$$y = \sum_{k=1}^{S} w_k \, \boldsymbol{\ell}_k \ \in \ \mathbb{R}^{N_t}. \tag{4}$$

Thus, each patient is represented by a compact, patient-specific subset of slots, yielding a sparse mixture of prognostic events. The entire decoder is trained end-to-end with the survival loss $\mathcal{L}_{\mathrm{surv}}$, which supervises the gating mechanism to prioritize outcome-relevant slots.

Slots Regularization. While selective slot activation encourages sparsity, it can also lead to the emergence of null slots—slots that fail to attend to any input features—thereby weakening the competition enforced by the gating mechanism. In information compression, as the entropy of the information bottleneck is upper-bounded by the input features $\mathbf{X}$, non-activated slots—though less predictive than activated ones—are leveraged to capture residual information for input reconstruction, thereby mitigating potential information loss and ensuring completeness. To mitigate this potential loss of information during slot-based compression, we introduce a reconstruction objective. Specifically, we decode the slots using a reconstruction head, and minimize the reconstruction loss $\mathcal{L}_{\mathrm{recon}}(\hat{\mathbf{X}}, \mathbf{X})$, where $\hat{\mathbf{X}}$ denotes the reconstructed features (see Appendix B.3 for details).

3.2 MULTIMODAL SLOTS INTERACTION WITH BILOGICAL GUIDANCE

Cross-modal Reconstruction using Biological Prior. To ensure effective cross-modal interaction, we introduce cross-modal reconstruction (Figure 2(b)) to uncover molecular-morphology mappings reflecting the fact that molecular events give rise to histopathological phenotypes (Wang et al., 2021; Liu et al., 2025b).

Given the omics-derived initialization slots $\mathbf{S}_g$, i.e. initialization learned from the training omics data, we perform slot attention (Eq. 1) with histopathology patch features $\mathbf{X}_h$ (for a better interpretability on original WSIs, we choose $\mathbf{X}_h$ instead of using $\mathbf{S}_h$) to obtain cross-modal slot representations $\tilde{\mathbf{S}}_{g \to h}$, which capture gene-related events as expressed in histopathological morphology. Thus the cross-modal interaction of this step is $O(S_g * M_h)$ with $S_g \ll M_h$. These slots are then

decoded by a reconstruction head $\mathcal{R}$ to recover the original genomic features, with position embeddings $\mathbf{Q}_g \in \mathbb{R}^{M_g \times d}$:

$$\mathbf{Q}_g = \text{Embed}([0, \ldots, M_g - 1]). \tag{5}$$

Since the input pathway features are arranged in a fixed index order, corresponding to a specific pathway index, this provides a positional prior that anchors reconstruction to the correct genomic coordinate. The reconstructed features are obtained as:

$$\tilde{\mathbf{X}}_g = \mathcal{R}(\mathbf{Q}_g, \tilde{\mathbf{S}}_{g \rightarrow h}), \tag{6}$$

with reconstruction loss $\mathcal{L}_{\text{recon}}(\tilde{\mathbf{X}}_g, \mathbf{X}_g)$. Here, histology serves as the morphological manifestation of underlying molecular events, guiding omics-derived slots toward biologically meaningful alignment. Notably, because genomic profiling is substantially more costly than histopathology and often unavailable at test time (Payne et al., 2018), this cross-modal reconstruction enhances robustness: the reconstruction head can impute genomic features directly from histology images, enabling inference even when omics data are missing.

Slots Interactions. To capture multimodal dependencies, we model slot interactions at both intra- and inter-modal levels—handled in parallel. Intra-modal interactions are captured using masked self-attention applied only to the top-K most self-predictive slots within each modality. For inter-modal interactions, we use all slots without masking to fully capture inter-modal synergies, including those arising from slots that are originally less self-predictive within their own modality. Instead of relying on one-shot co-attention (Chen et al., 2021b; Xu & Chen, 2023; Jaume et al., 2024), we design an *iterative cross-attention* scheme that follows the recurrent update in slot attention module. At each iteration, slots from one modality attend to those of the other: histology queries attend to omics keys/values, and vice versa (see Appendix B.4 for details). Finally, the output of three branches are concatenated and passed into a risk classification head for risk prediction. The computational complexity of slots interactions is $O((S_g + S_h)^2)$. Overall, instead of the original $O((M_g + M_h)^2)$, our model requires $O(S_g * M_g + S_h * M_h + S_g * M_h + (S_g + S_h)^2)$. Since $S_g, S_h \ll M_g, M_h$ in practice and the $(S_g + S_h)^2$ term is negligible, this expression can be upper-bounded by $O((S_g + S_h)(M_g + M_h))$.

3.3 Training and Inference

Overall Loss. The total training objective of SlotSPE is

$$\mathcal{L} = \mathcal{L}_{\text{surv}} + \lambda \mathcal{L}_{\text{recon}}. \tag{7}$$

where λ balances the survival and reconstruction components. $\mathcal{L}_{\text{surv}}$ includes both the survival prediction loss and the survival prediction losses w.r.t. each single modalities, while $\mathcal{L}_{\text{recon}}$ combines the slots regularization and cross-modal reconstruction objectives (refer Appendix B.5 for more details).

Inference. At test time, neither slots regularization reconstruction nor cross-modal reconstruction is required when both modalities are available. Cross-modal reconstruction is only invoked when genomic data are missing, serving as an imputation mechanism.

4 Experiments

4.1 Datasets and Implementations

We perform comprehensive evaluations across ten publicly available cancer cohorts from TCGA[1]. Case counts for each dataset are listed in Table 1. Following (Jaume et al., 2024), we predict disease-specific survival (DSS) with histological data including all diagnostic WSIs and transcriptomic profiles with DSS labels are obtained from cBioPortal[2]. Pathway gene sets are curated from Hallmarks (Subramanian et al., 2005; Liberzon et al., 2015) and Reactome (Gillespie et al., 2022), with genes absent in cBioPortal removed. We use 5-fold cross-validation and report the concordance index (C-index) (Harrell Jr et al., 1996) with standard deviation (std) and visualize the Kaplan–Meier (KM) Kaplan & Meier (1958) survival curves for stratified risk groups, perform the log-rank test Mantel et al. (1966) and restricted mean survival time (RMST) (Irwin, 1949; Karrison, 1986) to assess differences between them. Further information can be found in Appendix C.

---

[1] https://portal.gdc.cancer.gov/
[2] https://www.cbioportal.org/

Table 1: Comparison of mean C-index values across ten cancer datasets. g. and h. refer to genomic modality and histological modality, respectively. The best results and the second-best results are highlighted in **bold** and in underline. (See Table 4 for the completed table with standard deviation.)

Model	Modality	BRCA (N=1046)	COADREAD (N=573)	KIRC (N=488)	UCEC (N=488)	LUAD (N=467)	LUSC (N=460)	HNSC (N=438)	SKCM (N=409)	BLCA (N=381)	STAD (N=366)	Overall
MLP	g.	0.644	0.661	0.750	0.731	0.634	0.584	0.584	0.652	0.677	0.615	0.653
SNN	g.	0.645	0.664	0.754	0.753	0.652	0.548	0.597	0.673	0.694	0.585	0.656
SNNTrans	g.	0.657	0.710	0.747	0.724	0.648	0.555	0.605	0.684	0.699	0.593	0.662
SlotSPE	g.	0.698	0.733	0.774	0.752	0.646	0.616	0.611	0.678	0.701	0.596	0.681
ABMIL	h.	0.656	0.715	0.782	0.779	0.645	0.593	**0.664**	0.668	0.663	0.642	0.681
TransMIL	h.	0.647	0.723	0.767	0.759	0.655	0.582	0.644	0.632	0.647	0.656	0.671
CLAM-SB	h.	0.696	0.742	0.784	0.757	0.639	0.584	<u>0.649</u>	0.643	0.645	0.637	0.678
CLAM-MB	h.	0.715	0.721	0.782	0.788	0.632	0.608	0.636	0.640	0.657	0.641	0.682
SlotSPE	h.	<u>0.738</u>	0.729	0.790	0.779	0.656	0.598	0.635	0.659	0.642	**0.676**	0.690
Porpoise	g.+h.	0.651	0.699	0.788	0.788	0.647	0.597	0.599	0.682	0.690	0.619	0.676
MCAT	g.+h.	0.709	0.737	0.795	<u>0.802</u>	0.659	0.583	0.622	0.667	0.688	0.639	0.690
MOTCat	g.+h.	0.717	0.742	<u>0.799</u>	0.773	<u>0.663</u>	0.573	0.627	**0.688**	0.700	0.642	<u>0.692</u>
CMTA	g.+h.	0.731	0.720	0.794	0.764	0.662	0.599	0.619	0.681	<u>0.702</u>	0.639	0.691
SurvPath	g.+h.	0.728	0.727	0.787	0.751	0.661	0.577	0.611	0.666	0.666	0.647	0.682
PIBD	g.+h.	0.662	0.716	0.772	0.783	0.634	0.575	0.627	0.649	0.693	0.644	0.676
MMP	g.+h.	0.714	0.686	0.770	0.755	0.634	0.580	0.580	0.682	0.686	0.602	0.669
LD-CVAE	g.+h.	0.705	<u>0.753</u>	0.792	0.788	0.651	<u>0.620</u>	0.635	0.682	0.635	0.653	<u>0.692</u>
SlotSPE	g.+h.	**0.779**	**0.773**	**0.815**	**0.813**	**0.683**	**0.634**	0.642	**0.688**	**0.708**	<u>0.671</u>	**0.721**

4.2 COMPARISONS WITH SOTAS

We benchmark our approach against three categories of state-of-the-art (SOTA) methods: (1) **Transcriptomics**. We consider a feed-forward neural network (MLP) (Haykin, 1998), a pathway-specific baseline (SNNs) (Klambauer et al., 2017), and SNNTrans, which integrates SNNs with transformers (Klambauer et al., 2017; Shao et al., 2021). (2) **Histology**. We compare with leading MIL-based models, including ABMIL (Ilse et al., 2018), TransMIL (Shao et al., 2021), and CLAM (Lu et al., 2021). (3) **Multimodal**. We evaluate against eight multimodal SOTA methods including Porpoise (Chen et al., 2022), MCAT (Chen et al., 2021b), MOTCat (Xu & Chen, 2023), CMTA (Zhou & Chen, 2023), SurvPath (Jaume et al., 2024), PIBD (Zhang et al., 2024), MMP Song et al. (2024), and LD-CVAE (Zhou et al., 2025). Among them, the first five methods adopt instance-level interactions for histology and omics integration, PIBD and MMP employ prototype-based modeling, while LD-CVAE is specifically designed to handle missing genomic data.

From Table 1, we observe that most multimodal approaches (except Porpoise, PIBD, and MMP) outperform unimodal ones, confirming the complementary value of integrating histology and genomics. Although prototype-based methods (PIBD and MMP) use prototypes to represent complex inputs, their limited adaptability in capturing patient-specific events leads to unsatisfactory results on certain datasets, reducing overall effectiveness. In contrast, **SlotSPE** consistently achieves superior results, ranking first or second in nearly all cohorts and outperforming the second-best method by 2.9% in overall C-index across ten cancer types. Notably, this advantage also holds in *single-modality* settings: gene-only variant surpasses the next best model by 1.9%, while WSI-only variant improves by 0.8% in overall C-index. These results highlight the generalizability and flexibility of SlotSPE on cancer types and modalities. By dynamically capturing structured prognostic events and enabling cross-modal event-level alignment, SlotSPE establishes itself as a strong and reliable framework for survival prediction.

4.3 ROBUSTNESS

Since collecting genomic data is substantially more costly than acquiring histology images (Payne et al., 2018), clinical deployment often faces performance degradation when genomic features are unavailable. To evaluate the robustness of our approach, we test SlotSPE under the missing-genomic-modality setting (Table 2). SlotSPE maintains strong performance, and in several cohorts (e.g., HNSC, STAD) its missing-modality variant matches or even surpasses the full-modality version, demonstrating resilience to incomplete omics data. This robustness stems from the cross-modal reconstruction task, which enables the model to infer pathway representations from histology through omics-derived slots. Compared with LD-CVAE (Zhou et al., 2025)—a method explicitly designed for missing genomic data—SlotSPE shows a somewhat larger drop relative to its full-modality variant, yet still achieves superior or comparable performance across most datasets. Notably, its overall performance even exceeds that of full-modality LD-CVAE. These findings highlight

Table 2: C-index across ten cancer datasets under missing-genomics settings. The column *Missing* indicates whether the genomic is missing (✓ = missing). (See Table 6 for the complete table.)

Model	Missing	BRCA	COADREAD	KIRC	UCEC	LUAD	LUSC	HNSC	SKCM	BLCA	STAD	Overall
LD-CVAE		0.705	0.753	0.792	0.788	0.651	0.620	0.635	0.682	0.635	0.653	0.692
LD-CVAE	✓	0.715	0.736	0.798	0.775	0.638	0.596	0.649	0.676	0.644	0.647	0.688
SlotSPE		0.779	0.773	0.815	0.813	0.683	0.634	0.642	0.688	0.708	0.671	0.721
SlotSPE	✓	0.734	0.741	0.789	0.802	0.663	0.634	0.649	0.678	0.678	0.677	0.704
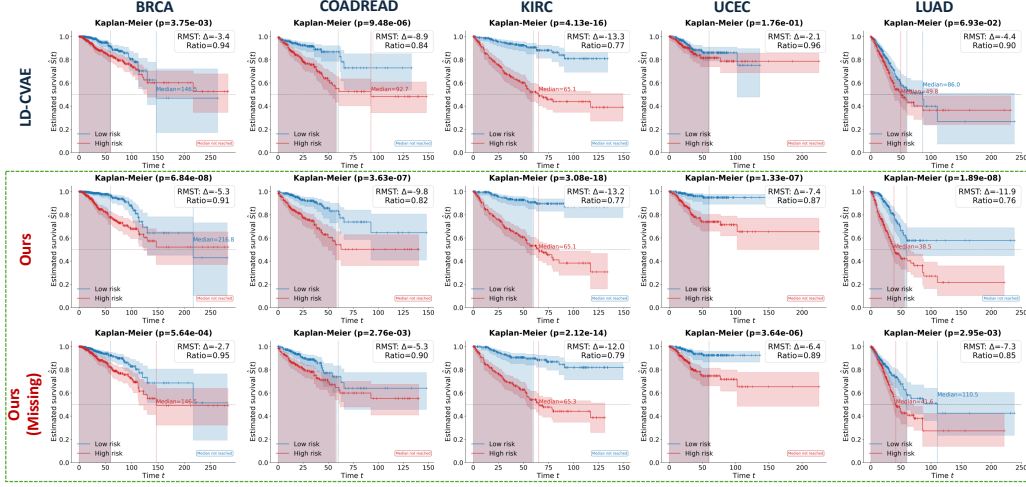
Figure 3: Kaplan–Meier curves of predicted high-risk and low-risk groups. A p-value < 0.05 at the top indicates statistically significant separation between groups. The restricted mean survival time (RMST) up to 60 months is also reported, with values shown as Δ (High–Low), and ratio (High/Low). (Zoom in to view details.)

that SlotSPE not only integrates multimodal features effectively but also delivers robust predictions in the absence of genomic data.

4.4 RISK STRATIFICATION

We further evaluate our model using standard survival analysis techniques. Figure 3 shows Kaplan–Meier (KM) curves, where patients are stratified into high- and low-risk groups based on predicted risk scores, with validation-set median as cut-off. Group separation is assessed using the log-rank test, and the corresponding p-values are reported. Across cohorts, our method achieves significantly greater separation between risk groups than the second-best method (LD-CVAE). For instance, in UCEC, LD-CVAE yields a non-significant p-value (> 0.05), while our model demonstrates significant separation ($p = 1.33 \times 10^{-7}$). For a more intuitive quantitative view, our method attains a larger absolute RMST difference Δ (High–Low) and a smaller RMST ratio (High/Low), together indicating robust prognostic stratification—even under missing-modality condition.

4.5 ABLATION STUDY

Table 3 summarizes the ablation results for each component of SlotSPE. We construct a strong baseline by retaining the vanilla slot attention and multimodal fusion modules while removing *selective slot attention* and *reconstructions*, and replacing the *iterative cross-attention* in multimodal fusion with standard dual branch cross-attention. As shown in the first and last rows, SlotSPE yields a substantial improvement over this baseline. Moreover, removing any of the four key components of SlotSPE results in a consistent performance drop across most datasets (Rows 2–6), indicating that each module contributes critically to the framework. Specifically, excluding selective slot attention reduces the model's ability to focus on discriminative prognostic patterns; removing cross-modal reconstruction with biological priors weakens the alignment between histology and genomics; and dropping iterative cross-attention impairs the iterative refinement needed to capture complex cross-modal dependencies. Slots regularization reconstruction further helps prevent null slots. Overall,

Table 3: Ablation of model components reported as mean C-index across ten cancer datasets. (See Table 8 for the complete table.)

| Variants	BRCA	COADREAD	KIRC	UCEC	LUAD	LUSC	HNSC	SKCM	BLCA	STAD	Overall
Baseline (Vanilla Slots)	0.711	0.734	0.784	0.779	0.661	0.592	0.610	0.668	0.686	0.644	0.687
w/o Selective Slot Attention	0.728	0.755	0.787	0.800	0.659	0.620	0.615	0.685	0.688	0.650	0.699
w/o Slots Regularization	0.734	0.763	0.794	0.811	0.662	0.611	0.621	0.679	0.693	0.667	0.704
w/o Cross-modal Reconstruction	0.710	0.758	0.787	0.794	0.663	0.619	0.611	0.671	0.698	0.645	0.696
w/o Iterative Cross-attention	0.727	0.761	0.803	0.794	0.653	0.602	0.624	0.676	0.702	0.648	0.699
SlotSPE	**0.779**	**0.773**	**0.815**	**0.813**	**0.683**	**0.634**	**0.642**	**0.688**	**0.708**	**0.671**	**0.721**
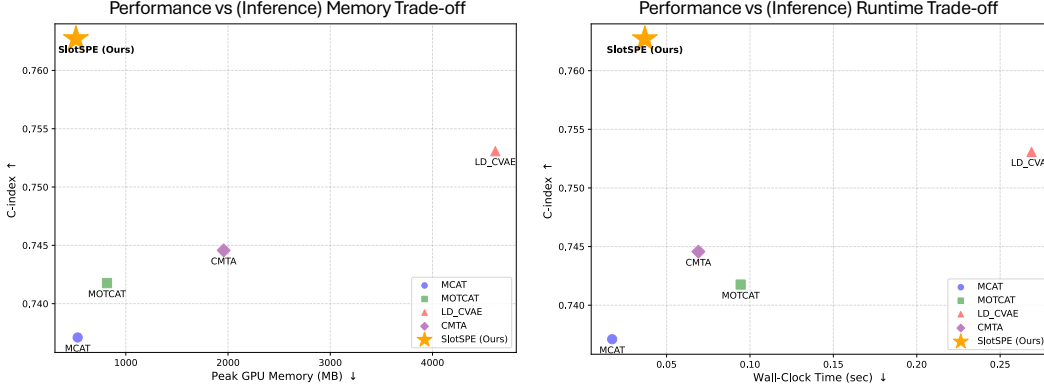
Figure 4: Performance vs (Inference) Memory/Runtime Trade-off

the full SlotSPE achieves the best performance, with an average C-index of 0.721, demonstrating that all four modules act synergistically to enable effective multimodal survival prediction.

4.6 EFFICIENCY

In this section, we assess the inference efficiency of our model by comparing it against the four strongest baselines listed in Table 1. Because all methods share the same WSI and omics encoders, we exclude these components and evaluate only the model-specific modules to ensure a fair and meaningful comparison. We measure peak inference memory and wall-clock time on the COAD-READ dataset, using all WSI patches during inference and configuring SlotSPE with 16 slots and 3 iterations. As shown in Figure 4, SlotSPE achieves the highest C-index while maintaining low runtime and memory usage. MCAT and MOTCAT achieve shorter runtime primarily because they omit self-attention computations within WSI patches; however, this design choice leads to a considerable drop in predictive performance. CMTA improves upon these baselines by employing Nyström attention (Xiong et al., 2021) as an approximation to full self-attention, but its capacity remains insufficient to capture the sparse, patient-specific structural prognostic signals. LD-CVAE attains the second-best performance, yet its reliance on a variational autoencoder introduces substantial memory and runtime overhead. Overall, SlotSPE offers the most favorable balance between efficiency and predictive performance, with minimal memory usage and the second-fastest runtime among all compared methods. Additional analysis is provided in the Appendix G.

4.7 INTERPRETABILITY

Cross-modality Alignment. Visualization of slot assignments (Figure 5A) shows that histology- and genomics-derived slots often align to structurally similar tissue regions in WSIs, reflecting event-level cross-modal consistency. Subtle differences remain, which can be attributed to modality-specific information: morphology-driven slots emphasize structural patterns, whereas genomics-driven slots highlight regions linked to molecular features. **Intra- and Inter-modal Interactions.** For intra-modal analysis, we visualize attention maps of histology-derived slots (first three plots in Figure 5B) and report the top-3 most relevant and irrelevant pathways per omics-derived slot (Figure 5C). For inter-modal analysis, we link pathway-level attention scores with the corresponding genomic-slot attention maps on WSIs (last three plots in Figure 5B). For instance, in Slot 2, highly
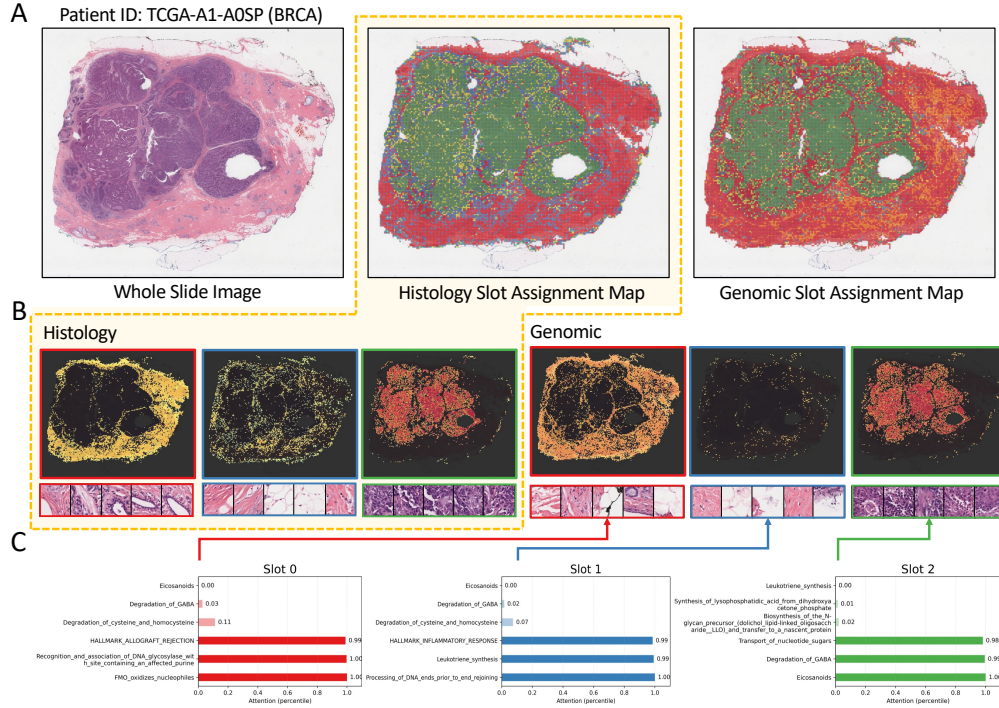
Figure 5: Interpretability of slots. (A) Original WSI, assignment map of histology-derived slots, and assignment map of omics-derived slots. (B) Attention maps for each slot with the top-5 most relevant patches highlighted. (C) Top-3 relevant and irrelevant pathways identified per slot.

attended pathways correspond to tumor-enriched regions, consistent with prior studies on breast cancer prognosis (Wang & DuBois, 2010; Neman et al., 2014; Chen et al., 2019; Ta et al., 2021; Xie et al., 2023).

In general, these interpretability analyzes demonstrate the ability of SlotSPE to establish event-level alignment across modalities and reveal pathway–morphology correspondences, offering the potential for biologically plausible explanations and new hypotheses for further validation. Additional examples are provided in the Appendix I.

5 CONCLUSION

In this work, we introduced **SlotSPE**, a slot-based framework for structural prognostic event modeling in multimodal cancer survival analysis. By compressing high-dimensional histology and omics data into interpretable slots, selectively activating the most predictive ones, and aligning them through biologically guided cross-modal reconstruction, SlotSPE effectively captures sparse, patient-specific prognostic events and effectively reduces the computational complexity of multimodal interactions. Our extensive evaluation across ten TCGA cohorts demonstrates that SlotSPE not only achieves state-of-the-art predictive performance but also maintains robustness under missing-modality conditions. Moreover, the interpretability analyses highlight its ability to uncover event-level morphology–molecular correspondences. These results underscore the potential of SlotSPE as a powerful and interpretable framework for multimodal precision oncology.

REFERENCES

Hugo JWL Aerts, Emmanuel Rios Velazquez, Ralph TH Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications*, 5(1):4006, 2014.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Ronald Chan, Chaiwat Aphivatanasiri, Ivan K Poon, Julia Y Tsang, Yunbi Ni, Maribel Lacambra, Joshua Li, Conrad Lee, and Gary M Tse. Spatial distribution and densities of cd103+ and foxp3+ tumor infiltrating lymphocytes by digital analysis for outcome prediction in breast cancer. *The Oncologist*, 29(3):e299–e308, 2024.

Kumardeep Chaudhary, Olivier B Poirion, Liangqun Lu, and Lana X Garmire. Deep learning–based multi-omics integration robustly predicts survival in liver cancer. *Clinical cancer research*, 24(6): 1248–1259, 2018.

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, Muhammad Shaban, Chengkuan Chen, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 339–349. Springer, 2021a.

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, 2021b.

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, 2022.

Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature medicine*, 30(3):850–862, 2024.

Xingyu Chen, Qianhua Cao, Ruocen Liao, Xuebiao Wu, Shining Xun, Jian Huang, and Chenfang Dong. Loss of abat-mediated gabaergic system promotes basal-like breast cancer progression by activating ca2+-nfat1 axis. *Theranostics*, 9(1):34, 2019.

Travers Ching, Xun Zhu, and Lana X Garmire. Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS computational biology*, 14(4): e1006076, 2018.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pp. 103–111, 2014.

Dan Ciresan, Alessandro Giusti, Luca Gambardella, and Jürgen Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. *Advances in neural information processing systems*, 25, 2012.

Dan C Cireşan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Mitosis detection in breast cancer histology images with deep neural networks. In *International conference on medical image computing and computer-assisted intervention*, pp. 411–418. Springer, 2013.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

David R Cox and Nancy Reid. A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91(3):729–737, 2004.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Carsten Denkert, Gunter von Minckwitz, Silvia Darb-Esfahani, Bianca Lederer, Barbara I Heppner, Karsten E Weber, Jan Budczies, Jens Huober, Frederick Klauschen, Jenny Furlanetto, et al. Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: a pooled analysis of 3771 patients treated with neoadjuvant therapy. *The lancet oncology*, 19(1):40–50, 2018.

Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997.

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.

Tong Ding, Sophia J Wagner, Andrew H Song, Richard J Chen, Ming Y Lu, Andrew Zhang, Anurag J Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, et al. A multimodal whole-slide foundation model for pathology. *Nature Medicine*, pp. 1–13, 2025.

Stephen B Edge and Carolyn C Compton. The american joint committee on cancer: the 7th edition of the ajcc cancer staging manual and the future of tnm. *Annals of surgical oncology*, 17(6):1471–1474, 2010.

Thomas A Gerds and Martin Schumacher. Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040, 2006.

Marc Gillespie, Bijay Jassal, Ralf Stephan, Marija Milacic, Karen Rothfels, Andrea Senff-Ribeiro, Johannes Griss, Cristoffer Sevilla, Lisa Matthews, Chuqiao Gong, et al. The reactome pathway knowledgebase 2022. *Nucleic acids research*, 50(D1):D687–D692, 2022.

Antonino Glaviano, Aaron SC Foo, Hiu Y Lam, Kenneth CH Yap, William Jacot, Robert H Jones, Huiyan Eng, Madhumathy G Nair, Pooyan Makvandi, Birgit Geoerger, et al. Pi3k/akt/mtor signaling transduction pathway and targeted therapies in cancer. *Molecular cancer*, 22(1):138, 2023.

Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18(17-18):2529–2545, 1999.

Klaus Greff, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks. *arXiv preprint arXiv:2012.05208*, 2020.

Emil Julius Gumbel. *Statistical theory of extreme values and some practical applications: a series of lectures*, volume 33. US Government Printing Office, 1954.

Balázs Győrffy. Survival analysis across the entire transcriptome identifies biomarkers with the highest prognostic power in breast cancer. *Computational and structural biotechnology journal*, 19:4101–4109, 2021.

John Hampshire and Alex Waibel. Connectionist architectures for multi-speaker phoneme recognition. *Advances in neural information processing systems*, 2, 1989.

Frank E Harrell Jr, Kerry L Lee, and Daniel B Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387, 1996.

Simon Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1998.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Patrick J Heagerty and Yingye Zheng. Survival model predictive accuracy and roc curves. *Biometrics*, 61(1):92–105, 2005.

Patrick J Heagerty, Thomas Lumley, and Margaret S Pepe. Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2):337–344, 2000.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8): 1735–1780, 1997.

Mahdi S Hosseini, Babak Ehteshami Bejnordi, Vincent Quoc-Huy Trinh, Lyndon Chan, Danial Hasan, Xingwen Li, Stephen Yang, Taehyo Kim, Haochen Zhang, Theodore Wu, et al. Computational pathology: a survey review and the way forward. *Journal of Pathology Informatics*, 15: 100357, 2024.

Le Hou, Dimitris Samaras, Tahsin M Kurc, Yi Gao, James E Davis, and Joel H Saltz. Patch-based convolutional neural network for whole slide tissue image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2424–2433, 2016.

Zhichao Hou, Jiacheng Leng, Jiating Yu, Zheng Xia, and Ling-Yun Wu. Pathexpsurv: pathway expansion for explainable survival analysis and disease gene discovery. *BMC bioinformatics*, 24 (1):434, 2023.

Jing Huang, Chenyang Yuan, Jiahui Jiang, Jianfeng Chen, Sunil S Badve, Yesim Gokmen-Polar, Rossana L Segura, Xinmiao Yan, Alexander Lazar, Jianjun Gao, et al. Bridging cell morphological behaviors and molecular dynamics in multi-modal spatial omics with morphlink. *Nature Communications*, 16(1):5878, 2025.

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

JO Irwin. The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice. *Epidemiology & Infection*, 47(2):188–189, 1949.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Guillaume Jaume, Anurag Vaidya, Richard J Chen, Drew FK Williamson, Paul Pu Liang, and Faisal Mahmood. Modeling dense multimodal interactions between biological pathways and histology for survival prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11579–11590, 2024.

Stephen P Jenkins. Survival analysis. *Unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester, UK*, 42:54–56, 2005.

Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994.

Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

Theodore Karrison. Restricted mean life with adjustment for covariates. *Controlled Clinical Trials*, 7(3):247, 1986.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International conference on machine learning*, pp. 2649–2658. PMLR, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. *Advances in neural information processing systems*, 30, 2017.

Wouter Kool, Herke Van Hoof, and Max Welling. Stochastic beams and where to find them: The gumbel-top-k trick for sampling sequences without replacement. In *International conference on machine learning*, pp. 3499–3508. PMLR, 2019.

Ruoyu Li, Jiawen Yao, Xinliang Zhu, Yeqing Li, and Junzhou Huang. Graph cnn for survival analysis on whole slide pathological images. In *International conference on medical image computing and computer-assisted intervention*, pp. 174–182. Springer, 2018.

Hongling Liang, Jianqing Huang, Hongsheng Li, Weixing He, Xiang Ao, Zhi Xie, Yu Chen, Zhiyi Lv, Leyao Zhang, Yanhua Zhong, et al. Spatial proximity of cd8+ t cells to tumor cells predicts neoadjuvant therapy efficacy in breast cancer. *NPJ Breast Cancer*, 11(1):13, 2025.

Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell systems*, 1(6): 417–425, 2015.

Hong Liu, Haosen Yang, Federica Eduati, Josien PW Pluim, and Mitko Veta. Adaptive prototype learning for multimodal cancer survival analysis. *arXiv preprint arXiv:2503.04643*, 2025a.

Junzhuo Liu, Markus Eckstein, Zhixiang Wang, Friedrich Feuerhake, and Dorit Merhof. Spatial transcriptomics expression prediction from histopathology based on cross-modal mask reconstruction and contrastive learning. *bioRxiv*, pp. 2025–06, 2025b.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33:11525–11538, 2020.

Bin Lou, Semihcan Doken, Tingliang Zhuang, Danielle Wingerter, Mishka Gidwani, Nilesh Mistry, Lance Ladic, Ali Kamen, and Mohamed E Abazeed. An image-based deep learning framework for individualising radiotherapy dose: a retrospective analysis of outcome prediction. *The Lancet Digital Health*, 1(3):e136–e147, 2019.

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.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature medicine*, 30(3):863–874, 2024.

Hui Luo, Jiashuang Huang, Hengrong Ju, Tianyi Zhou, and Weiping Ding. Multimodal multi-instance evidence fusion neural networks for cancer survival prediction. *Scientific Reports*, 15(1): 10470, 2025.

Chris J Maddison, Daniel Tarlow, and Tom Minka. A* sampling. *Advances in neural information processing systems*, 27, 2014.

Nathan Mantel et al. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*, 50(3):163–170, 1966.

Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, 10, 1997.

Josh Neman, John Termini, Sharon Wilczynski, Nagarajan Vaidehi, Cecilia Choy, Claudia M Kowolik, Hubert Li, Amanda C Hambrecht, Eugene Roberts, and Rahul Jandial. Human breast cancer metastases to the brain display gabaergic properties in the neural niche. *Proceedings of the National Academy of Sciences*, 111(3):984–989, 2014.

Emma Nolan, Peter Savas, Antonia N Policheni, Phillip K Darcy, François Vaillant, Christopher P Mintoff, Sathana Dushyanthen, Mariam Mansour, Jia-Min B Pang, Stephen B Fox, et al. Combined immune checkpoint blockade as a therapeutic strategy for brca1-mutated breast cancer. *Science translational medicine*, 9(393):eaal4922, 2017.

Natalya N Pavlova and Craig B Thompson. The emerging hallmarks of cancer metabolism. *Cell metabolism*, 23(1):27–47, 2016.

Katherine Payne, Sean P Gavan, Stuart J Wright, and Alexander J Thompson. Cost-effectiveness analyses of genetic and genomic diagnostic tests. *Nature Reviews Genetics*, 19(4):235–246, 2018.

Florian Röhrig and Almut Schulze. The multifaceted roles of fatty acid synthesis in cancer. *Nature reviews cancer*, 16(11):732–749, 2016.

Jürgen Schmidhuber. Learning factorial codes by predictability minimization. *Neural computation*, 4(6):863–879, 1992.

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.

Zhuchen Shao, Yang Chen, Hao Bian, Jian Zhang, Guojun Liu, and Yongbing Zhang. Hvtsurv: Hierarchical vision transformer for patient-level survival prediction from whole slide image. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 2209–2217, 2023.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

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.

Andrew H Song, Richard J Chen, Guillaume Jaume, Anurag J Vaidya, Alexander S Baras, and Faisal Mahmood. Multimodal prototyping for cancer survival prediction. *arXiv preprint arXiv:2407.00224*, 2024.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *ICML Workshop on Deep Learning*, 2015.

Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

Hoang Dang Khoa Ta, Do Thi Minh Xuan, Wan-Chun Tang, Gangga Anuraga, Yi-Chun Ni, Syu-Ruei Pan, Yung-Fu Wu, Fenny Fitriani, Elvira Mustikawati Putri Hermanto, Muhammad Athoillah, et al. Novel insights into the prognosis and immunological value of the slc35a (solute carrier 35a) family genes in human breast cancer. *Biomedicines*, 9(12):1804, 2021.

Robert J Tibshirani and Bradley Efron. An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57(1):1–436, 1993.

Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 61(3):611–622, 1999.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

G Tomasello, F Petrelli, M Ghidini, E Pezzica, R Passalacqua, F Steccanella, L Turati, G Sgroi, and S Barni. Tumor regression grade and survival after neoadjuvant treatment in gastro-esophageal cancer: a meta-analysis of 17 published studies. *European Journal of Surgical Oncology*, 43(9): 1607–1616, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Bert Vogelstein, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis A Diaz Jr, and Kenneth W Kinzler. Cancer genome landscapes. *science*, 339(6127):1546–1558, 2013.

Sarah Volinsky-Fremond, Nanda Horeweg, Sonali Andani, Jurriaan Barkey Wolf, Maxime W Lafarge, Cor D de Kroon, Gitte Ørtoft, Estrid Høgdall, Jouke Dijkstra, Jan J Jobsen, et al. Prediction of recurrence risk in endometrial cancer with multimodal deep learning. *Nature medicine*, 30(7): 1962–1973, 2024.

Quoc Dang Vu, Kashif Rajpoot, Shan E Ahmed Raza, and Nasir Rajpoot. Handcrafted histological transformer (h2t): Unsupervised representation of whole slide images. *Medical image analysis*, 85:102743, 2023.

Dingzhi Wang and Raymond N DuBois. Eicosanoids and cancer. *Nature Reviews Cancer*, 10(3): 181–193, 2010.

Shuo Wang, Yuanning Liu, Hao Zhang, and Zhen Liu. Survconvmixer: robust and interpretable cancer survival prediction based on convmixer using pathway-level gene expression images. *BMC bioinformatics*, 25(1):133, 2024.

Yinxi Wang, Kimmo Kartasalo, Philippe Weitz, Balazs Acs, Masi Valkonen, Christer Larsson, Pekka Ruusuvuori, Johan Hartman, and Mattias Rantalainen. Predicting molecular phenotypes from histopathology images: a transcriptome-wide expression–morphology analysis in breast cancer. *Cancer research*, 81(19):5115–5126, 2021.

Zhikang Wang, Yumeng Zhang, Yingxue Xu, Seiya Imoto, Hao Chen, and Jiangning Song. Histo-genomic knowledge association for cancer prognosis from histopathology whole slide images. *IEEE Transactions on Medical Imaging*, 2025.

Simon Wiegrebe, Philipp Kopper, Raphael Sonabend, Bernd Bischl, and Andreas Bender. Deep learning for survival analysis: a review. *Artificial Intelligence Review*, 57(3):65, 2024.

Mingchen Xie, Fuxu Wang, Bing Chen, Zeyu Wu, Ci Chen, and Jian Xu. Systematic pan-cancer analysis identifies slc35c1 as an immunological and prognostic biomarker. *Scientific Reports*, 13 (1):5331, 2023.

Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A nyström-based algorithm for approximating self-attention. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 14138–14148, 2021.

Yingxue Xu and Hao Chen. Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 21241–21251, 2023.

Yiwen Xu, Ahmed Hosny, Roman Zeleznik, Chintan Parmar, Thibaud Coroller, Idalid Franco, Raymond H Mak, and Hugo JWL Aerts. Deep learning predicts lung cancer treatment response from serial medical imaging. *Clinical cancer research*, 25(11):3266–3275, 2019.

Zekang Yang, Hong Liu, and Xiangdong Wang. Scmil: Sparse context-aware multiple instance learning for predicting cancer survival probability distribution in whole slide images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 448–458. 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.

Jiawen Yao, Yu Shi, Kai Cao, Le Lu, Jianping Lu, Qike Song, Gang Jin, Jing Xiao, Yang Hou, and Ling Zhang. Deepprognosis: Preoperative prediction of pancreatic cancer survival and surgical margin via comprehensive understanding of dynamic contrast-enhanced ct imaging and tumor-vascular contact parsing. *Medical image analysis*, 73:102150, 2021.

Xue Qin Yu, Mei Ling Yap, Elvin S Cheng, Preston J Ngo, Pavla Vaneckova, Deme Karikios, Karen Canfell, and Marianne F Weber. Evaluating prognostic factors for sex differences in lung cancer survival: findings from a large australian cohort. *Journal of Thoracic Oncology*, 17(5):688–699, 2022.

Wei Yuan, Yijiang Chen, Biyue Zhu, Sen Yang, Jiayu Zhang, Ning Mao, Jinxi Xiang, Yuchen Li, Yuanfeng Ji, Xiangde Luo, et al. Pancancer outcome prediction via a unified weakly supervised deep learning model. *Signal transduction and targeted therapy*, 10(1):285, 2025.

Shekoufeh Gorgi Zadeh and Matthias Schmid. Bias in cross-entropy-based training of deep survival networks. *IEEE transactions on pattern analysis and machine intelligence*, 43(9):3126–3137, 2020.

Shuaiyu Zhang, Xun Lin, Rongxiang Zhang, Yu Bai, Yong Xu, Tao Tan, Xunbin Zheng, and Zitong Yu. Adamhf: Adaptive multimodal hierarchical fusion for survival prediction. *arXiv preprint arXiv:2503.21124*, 2025.

Yilan Zhang, Yingxue Xu, Jianqi Chen, Fengying Xie, and Hao Chen. Prototypical information bottlenecking and disentangling for multimodal cancer survival prediction. *arXiv preprint arXiv:2401.01646*, 2024.

Lianhe Zhao, Qiongye Dong, Chunlong Luo, Yang Wu, Dechao Bu, Xiaoning Qi, Yufan Luo, and Yi Zhao. Deepomix: a scalable and interpretable multi-omics deep learning framework and application in cancer survival analysis. *Computational and structural biotechnology journal*, 19: 2719–2725, 2021.

Tingting Zhao, Yongyong Ren, Hui Lu, and Yan Kong. Decision level scheme for fusing multiomics and histology slide images using deep neural network for tumor prognosis prediction. *Scientific Reports*, 15(1):25479, 2025.

Fengtao Zhou and Hao Chen. Cross-modal translation and alignment for survival analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21485–21494, 2023.

Huajun Zhou, Fengtao Zhou, and Hao Chen. Cohort-individual cooperative learning for multimodal cancer survival analysis. *IEEE Transactions on Medical Imaging*, 2024.

Junjie Zhou, Jiao Tang, Yingli Zuo, Peng Wan, Daoqiang Zhang, and Wei Shao. Robust multimodal survival prediction with conditional latent differentiation variational autoencoder. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10384–10393, 2025.

APPENDIX TABLE OF CONTENTS

A   THE USE OF LARGE LANGUAGE MODELS (LLMS)

In this work, we use Large Language Models (LLMs) to correct grammar mistakes and polish the writing.

B   METHOD DETAILS

B.1   PROBLEM FORMULATION

Given the i-th patient with multimodal inputs $\mathbf{I}^{(i)} = \left(\mathbf{X}_h^{(i)}, \mathbf{X}_g^{(i)}, t^{(i)}, c^{(i)} \right)$ where $\mathbf{X}_h^{(i)}$ denotes the pathology data, $\mathbf{X}_g^{(i)}$ denotes the genomic profile, $t^{(i)} \in \{1, \ldots, N_t\}$ is the discretized survival time interval (risk band), and $c^{(i)} \in \{0, 1\}$ is the censoring indicator ($c^{(i)} = 0$ for an observed event, $c^{(i)} = 1$ for right-censoring). Following the multiple instance learning (MIL) paradigm (Chen et al., 2021b; Jaume et al., 2024), features extracted from the WSI patch encoder and gene encoder are organized as bags of instances: $\mathbf{X}_h^{(i)} = \left\{ \mathbf{x}_{h,j}^{(i)} \in \mathbb{R}^d \right\}_{j=1}^{M_h}$ and $\mathbf{X}_g^{(i)} = \left\{ \mathbf{x}_{g,j}^{(i)} \in \mathbb{R}^d \right\}_{j=1}^{M_g}$, where M_h and M_g denote the numbers of WSI patches and biological pathways, respectively, and d is the feature dimension. The objective is to learn $\mathcal{F}_\theta : \mathcal{X}_h \times \mathcal{X}_g \to \mathbb{R}^D$ mapping $(\mathbf{X}_h^{(i)}, \mathbf{X}_g^{(i)}) \mapsto \mathbf{z}^{(i)}$ for survival prediction. We denote by $h_t^{(i)}$ the predicted hazard probability at a discrete time interval t, i.e., the probability of an event occurring at t given survival up to t. The survival function is $\mathcal{S}_t^{(i)}$. With N_D patients in the training set, the hazard, survival, and the negative log-likelihood (NLL) loss (Zadeh & Schmid, 2020) are:

$$h_t^{(i)} = P(T = t \mid T \geq t, \mathbf{z}^{(i)}), \quad \mathcal{S}_t^{(i)} = \prod_{k=1}^{t} \left(1 - h_k^{(i)} \right),$$

$$\mathcal{L}_{\text{surv}}(\{\mathbf{z}^{(i)}, t^{(i)}, c^{(i)}\}_{i=1}^{N_D}) = -\sum_{i=1}^{N_D} \Big[c^{(i)} \log \mathcal{S}_{t^{(i)}}^{(i)} + (1 - c^{(i)}) \log \mathcal{S}_{t^{(i)}-1}^{(i)} + (1 - c^{(i)}) \log h_{t^{(i)}}^{(i)} \Big]. \tag{8}$$

B.2   DETAILS OF SELECTIVE SLOT ACTIVATION

To achieve sparse yet differentiable slot selection, we adopt the Gumbel-Top-K (Gumbel, 1954; Maddison et al., 2014; Kool et al., 2019) trick combined with a Straight-Through (ST) estimator (Jang et al., 2016).

Given slot scores $\mathbf{r} \in \mathbb{R}^S$, we add i.i.d. Gumbel noise $g_k \sim \text{Gumbel}(0, 1)$:

$$\tilde{r}_k = r_k + g_k. \tag{9}$$

This perturbation transforms deterministic scores into a stochastic sampling process where larger r_k values are more likely to be selected. To obtain a differentiable relaxation, we then apply softmax to the perturbed scores:

$$\boldsymbol{\alpha} = \text{softmax}(\tilde{\mathbf{r}}), \tag{10}$$

In practice, we use an ST estimator that combines a hard mask for the forward pass with a soft relaxation for the backward pass:

$$\hat{\mathbf{G}} \leftarrow \text{onehot\_topk}(\boldsymbol{\alpha}) + \big(\boldsymbol{\alpha} - \text{stopgrad}(\boldsymbol{\alpha}) \big). \tag{11}$$

where $\text{onehot\_topk}(\cdot)$ produces a discrete K-hot mask in the forward pass, while gradients flow through the soft assignment $\boldsymbol{\alpha}$.

Finally, the selected slots are reweighted by masking and renormalizing the softmax weights in Eq. 3, which ensures that only the top-K slots contribute, while their relative importance is adaptively normalized. The resulting $\hat{\mathbf{G}} \in [0, 1]^S$ acts as a differentiable K-hot mask, enforcing sparse slot selection while preserving end-to-end trainability.

B.3 DETAILS OF RECONSTRUCTIONS

Slots Regularization as Reconstruction. Within each modality, slots $\mathbf{S}$ are required to reconstruct the original modality-specific features to address the emergence of null slots. We use a Transformer decoder (Vaswani et al., 2017) as the reconstruction head. For genomics, slot embeddings $\mathbf{S}_g$ are decoded to approximate the original pathway embeddings $\mathbf{X}_g$, guided by the positional embeddings $\mathbf{Q}_g$ (Eq. 5), with reconstruction optimized via MSE loss:

$$\hat{\mathbf{X}}_g = \mathcal{R}(\mathbf{Q}_g, \mathbf{S}_g), \qquad \mathcal{L}_{\text{recon}}^g = \|\hat{\mathbf{X}}_g - \mathbf{X}_g\|_2^2. \tag{12}$$

For histopathology, reconstructing entire WSIs is infea1sible due to their scale and random patch sampling. Instead, we employ a fixed MLP to map the original patch features $\mathbf{X}_h$ to query embeddings $\mathbf{Q}_h$. The WSI slots $\mathbf{S}_h$ are then decoded back to patch-level embeddings, with cosine similarity enforcing feature-level consistency:

$$\hat{\mathbf{X}}_h = \mathcal{R}(\mathbf{Q}_h, \mathbf{S}_h), \qquad \mathcal{L}_{\text{recon}}^h = 1 - \frac{1}{M_h} \sum_{j=1}^{M_h} \cos(\hat{\mathbf{x}}_{h,j}, \mathbf{x}_{h,j}). \tag{13}$$

Cross-modal Reconstruction. To capture molecular-morphological mappings, we introduce a cross-modal reconstruction task in which omics-derived slots are guided to predict pathway embeddings from histopathology. Specifically, WSI patch features are re-encoded through omics-derived slots $\mathbf{S}_g$ and decoded to approximate $\mathbf{X}_g$ with an MSE loss:

$$\tilde{\mathbf{X}}_g = \mathcal{R}(\mathbf{Q}_g, \tilde{\mathbf{S}}_{g \to h}), \qquad \mathcal{L}_{\text{recon}}^{g \to h} = \|\tilde{\mathbf{X}}_g - \mathbf{X}_g\|_2^2. \tag{14}$$

B.4 DETAILS OF MULTIMODAL INTERACTIONS

In contrast to one-shot co-attention methods (Chen et al., 2021b; Xu & Chen, 2023; Jaume et al., 2024), we employ an *iterative cross-attention* scheme that reuses a single fusion layer for L iterations of bidirectional updates. At each iteration l, omics- and histology-derived slots alternately attend to one another, absorbing complementary information while preserving their own context via recurrent updates, akin to slot attention. This iterative design supports multi-step fusion, progressively aligning molecular and morphological evidence that may require hierarchical inference (e.g., molecular dysregulation propagating through cellular morphology to tissue-level structure (Huang et al., 2025)). Concretely, at iteration l, the update of histology slots attending to omics slots is defined as

$$\mathbf{A}_{h \leftarrow g}^{(l)} = \text{softmax}\left(\frac{\mathcal{Q}(\hat{\mathbf{S}}_h^{(l)})\mathcal{K}(\hat{\mathbf{S}}_g^{(l)})^\top}{\sqrt{d}} \right) \in \mathbb{R}^{S_h \times S_g}, \quad \mathbf{U}_h^{(l)} = \mathbf{A}_{h \leftarrow g}^{(l)} \mathcal{V}(\hat{\mathbf{S}}_g^{(l)}),$$
$$\hat{\mathbf{S}}_h^{(l+1)} = \text{RNN}(\hat{\mathbf{S}}_h^{(l)}, \mathbf{U}_h^{(l)}), \quad \hat{\mathbf{S}}_h^{(l+1)} \leftarrow \hat{\mathbf{S}}_h^{(l+1)} + \text{MLP}(\hat{\mathbf{S}}_h^{(l+1)}), \tag{15}$$

with a symmetric update for omics slots attending to histology slots. Repeating for $l = \{0, \ldots, L - 1\}$ yields progressively aligned multimodal slot representations.

In parallel with cross-attention, each modality undergoes intra-modal self-attention, yielding refined representations $\bar{\mathbf{S}}_h$ and $\bar{\mathbf{S}}_g$. To form the final multimodal representation, we compute pooled summaries (mean over slots) and concatenate three branches outputs:

$$\mathbf{z} = \left[\text{Pool}([\hat{\mathbf{S}}_h \| \hat{\mathbf{S}}_g]) \, \| \, \text{Pool}(\bar{\mathbf{S}}_h) \, \| \, \text{Pool}(\bar{\mathbf{S}}_g) \right] \in \mathbb{R}^{3d}, \tag{16}$$

where $\|$ denotes concatenation along the slot dimension and $\text{Pool}(\cdot)$ indicates mean pooling over slots. Finally, a prediction head maps $\mathbf{z}$ to discrete risk classes, with training jointly optimizing the survival loss and the auxiliary reconstruction losses described above.

B.5 DETAILS OF OVERALL LOSS

The total training objective of SlotSPE is

$$\mathcal{L} = \mathcal{L}_{\text{surv}} + \lambda \mathcal{L}_{\text{recon}}. \tag{17}$$

where λ balances the survival and reconstruction components. The survival loss $\mathcal{L}_{\text{surv}}$ consists of the survival prediction loss together with each single-modal survival losses. The single-modal survival

predictions, i.e., y_h (histology) and y_g (genomics) in Eq. 18, are obtained by applying their modal-specific MoE decoders to the corresponding slot representations, $\mathbf{S}_h$ and $\mathbf{S}_g$. These single-modal survival losses incorporate task supervision for the gated slot logits y_h and y_g:

$$\mathcal{L}_{\text{surv}} = \mathcal{L}_{\text{surv}}(\mathbf{z}, t, c) + \mathcal{L}_{\text{surv}}^h(y_h, t, c) + \mathcal{L}_{\text{surv}}^g(y_g, t, c) \tag{18}$$

while $\mathcal{L}_{\text{recon}}$ combines the slots regularization and cross-modal reconstruction objectives:

$$\mathcal{L}_{\text{recon}} = \mathcal{L}_{\text{recon}}^g + \mathcal{L}_{\text{recon}}^h + \mathcal{L}_{\text{recon}}^{g \to h} \tag{19}$$

C   DATASETS AND IMPLEMENTATIONS

Dataset. We perform comprehensive evaluations across ten publicly available cancer cohorts from TCGA[3] , including Breast Invasive Carcinoma (BRCA), Bladder Urothelial Carcinoma (BLCA), Colon and Rectum Adenocarcinoma (COADREAD), Stomach Adenocarcinoma (STAD), Head and Neck Squamous Cell Carcinoma (HNSC), as well as Lung Adenocarcinoma (LUAD), Lung Squamous Cell Carcinoma (LUSC), Kidney Renal Clear Cell Carcinoma (KIRC), and Skin Cutaneous Melanoma (SKCM). Case counts for each dataset are listed in Table 1. Following (Jaume et al., 2024), we predict disease-specific survival (DSS), a more precise indicator of patient status than overall survival. Histological data include all diagnostic WSIs, while transcriptomic profiles with DSS labels are obtained from cBioPortal[4]. Pathway gene sets are curated from Hallmarks (Subramanian et al., 2005; Liberzon et al., 2015) and Reactome (Gillespie et al., 2022), with genes absent in cBioPortal removed, yielding 330 pathways.

Evaluation. We adopt 5-fold cross-validation and report the concordance index (C-index) (Harrell Jr et al., 1996) together with the standard deviation (std) across folds. To further assess clinical relevance, we visualize Kaplan–Meier (KM) survival curves (Kaplan & Meier, 1958) for stratified risk groups and apply the log-rank test (Mantel et al., 1966) to evaluate global differences in survival distributions. In addition, we compute the restricted mean survival time (RMST) (Irwin, 1949; Karrison, 1986). RMST, defined as the area under the estimated survival curve up to a clinically meaningful truncation time (60 months in our experiments), provides an interpretable summary of the average survival time within the specified horizon and does not rely on the proportional hazards assumption. We report: (a) the RMST values for both high- and low-risk groups, (b) the RMST difference (Δ = High–Low) with a 95% bootstrap confidence interval (CI) and a two-sided p-value testing whether the difference is significantly different from zero, and (c) the RMST ratio (High/Low) with a log-transformed 95% bootstrap CI as a scale-free measure of group separation. To ensure robust inference, both CI and p-values are estimated from 1000 bootstrap replicates (Tibshirani & Efron, 1993). These statistics are annotated on the KM plots, providing an interpretable and robust evaluation of prognostic stratification.

Taken together, the three metrics are complementary: the C-index measures overall ranking concordance of predicted risks, the log-rank test evaluates survival curve separation under proportional hazards, and RMST provides a clinically grounded and assumption-light measure of absolute and relative survival differences.

Implementation. Tissue regions are segmented from WSIs, and non-overlapping 256×256 patches are extracted at $20\times$ magnification. We adopt UNI (Chen et al., 2024) as the patch encoder and SNNs as the pathway encoder. The method is implemented in Python using PyTorch and executed on only one NVIDIA A100 GPU. For survival prediction, disease-specific survival (DSS) time is discretized into four intervals. Bag features are embedded via an MLP with a 512-dimensional hidden layer and a 256-dimensional output, thus d and d_{slot} are all set to 256. The number of iterations for slot attention is set to $T = 10$, and for iterative cross-attention to $L = 3$. The top-K in selective slot activation decoder is set to 25% slot numbers since patients are divided into 4 risk groups. We report the best performance by ablating the slot numbers in Appendix F.3.1. Pathway tokens are reconstructed using MSE loss, while histopathology tokens are optimized with cosine similarity loss using a hyperparameter $\lambda = 0.1$ (searched from $[0.01, 0.1, 1.0]$). To support batch training, 4096 patches are randomly sampled per WSI, whereas all patches are used during inference. Optimization is carried out with Adam (Kingma & Ba, 2014) at a learning rate of 5×10^{-4}. All models are trained for 30 epochs with a batch size of 32.

---

[3] https://portal.gdc.cancer.gov/
[4] https://www.cbioportal.org/

Table 4: Comparison of C-index (mean ± std) across ten cancer datasets using **biological pathways**. Best and second-best results are in **bold** and <u>underline</u>.

Model	Modality	BRCA (N=1046)	COADREAD (N=573)	KIRC (N=488)	UCEC (N=488)	LUAD (N=467)	Overall
MLP	g.	0.644± 0.097	0.661± 0.053	0.750± 0.035	0.731± 0.071	0.634± 0.069	-
SNN	g.	0.645±0.086	0.664±0.021	0.754± 0.035	0.753± 0.071	0.652± 0.063	-
SNNTrans	g.	0.657±0.060	0.710±0.053	0.747±0.035	0.724±0.061	0.648±0.059	-
SlotSPE	g.	0.698±0.046	0.733±0.045	0.774±0.022	0.752±0.065	0.646±0.064	-
ABMIL	h.	0.656±0.037	0.715±0.029	0.782±0.026	0.779±0.057	0.645±0.061	-
TransMIL	h.	0.647±0.032	0.723±0.027	0.767±0.035	0.759±0.059	0.655±0.052	-
CLAM-SB	h.	0.696±0.049	0.742±0.046	0.784±0.033	0.757±0.072	0.639±0.059	-
CLAM-MB	h.	0.715±0.034	0.721±0.041	0.782±0.014	0.788±0.074	0.632±0.070	-
SlotSPE	h.	0.738±0.019	0.729±0.038	0.790±0.023	0.779±0.057	0.656±0.038	-
Porpoise	g.+h.	0.651±0.046	0.699±0.050	0.788±0.028	0.788±0.092	0.647±0.091	-
MCAT	g.+h.	0.709±0.052	0.737±0.033	0.795±0.018	0.802±0.060	0.659±0.054	-
MOTCat	g.+h.	0.717±0.064	0.742±0.042	0.799±0.013	0.773±0.060	0.663±0.052	-
CMTA	g.+h.	0.731±0.066	0.720±0.008	0.794±0.018	0.764±0.069	0.662±0.053	-
SurvPath	g.+h.	0.728±0.040	0.727±0.024	0.787±0.025	0.751±0.084	0.661±0.050	-
PIBD	g.+h.	0.662±0.054	0.716±0.054	0.772±0.007	0.783±0.053	0.634±0.065	-
MMP	g.+h.	0.714 ± 0.050	0.686 ± 0.016	0.770 ±0.011	0.755 ±0.073	0.634 ± 0.048	-
LD-CVAE	g.+h.	0.705±0.050	0.753±0.029	0.792±0.025	0.788±0.068	0.651±0.055	-
SlotSPE (8 slots/modality)	g.+h.	<u>0.748±0.052</u>	0.759±0.027	0.802±0.009	<u>0.807±0.067</u>	0.653±0.056	-
SlotSPE ($T=3$)	g.+h.	0.732±0.056	<u>0.763±0.031</u>	**0.817±0.020**	0.788±0.067	<u>0.672±0.056</u>	-
SlotSPE	g.+h.	**0.779±0.061**	**0.773±0.042**	<u>0.815±0.017</u>	**0.813±0.059**	**0.683±0.058**	-

Model	Modality	LUSC (N=460)	HNSC (N=438)	SKCM (N=409)	BLCA (N=381)	STAD (N=366)	Overall
MLP	g.	0.584±0.072	0.584±0.061	0.652±0.038	0.677±0.053	0.615±0.063	0.653
SNN	g.	0.548±0.041	0.597±0.039	0.673±0.021	0.694±0.040	0.585±0.048	0.656
SNNTrans	g.	0.555±0.042	0.605±0.065	0.684±0.029	0.699±0.031	0.593±0.038	0.662
SlotSPE	g.	0.616±0.058	0.611±0.038	0.678±0.025	0.701±0.017	0.596±0.036	0.681
ABMIL	h.	0.593±0.055	**0.664±0.043**	0.668±0.042	0.663±0.072	0.642±0.044	0.681
TransMIL	h.	0.582±0.039	0.644±0.020	0.632±0.050	0.647±0.032	0.656±0.038	0.671
CLAM-SB	h.	0.584±0.065	<u>0.649±0.042</u>	0.643±0.049	0.645±0.052	0.637±0.022	0.678
CLAM-MB	h.	0.608±0.056	0.636±0.038	0.640±0.032	0.657±0.029	0.641±0.032	0.682
SlotSPE	h.	0.598±0.054	0.635±0.038	0.659±0.063	0.642±0.035	**0.676±0.033**	0.690
Porpoise	g.+h.	0.597±0.050	0.599±0.048	0.682±0.021	0.690±0.025	0.619±0.022	0.676
MCAT	g.+h.	0.583±0.051	0.622±0.049	0.667±0.049	0.688±0.024	0.639±0.032	0.690
MOTCat	g.+h.	0.573±0.026	0.627±0.049	**0.688±0.054**	0.700±0.010	0.642±0.019	0.692
CMTA	g.+h.	0.599±0.071	0.619±0.051	0.681±0.036	0.702±0.033	0.639±0.038	0.691
SurvPath	g.+h.	0.577±0.024	0.611±0.024	0.666±0.054	0.666±0.034	0.647±0.027	0.682
PIBD	g.+h.	0.575±0.098	0.627±0.069	0.649±0.048	0.693±0.032	0.644±0.038	0.676
MMP	g.+h.	0.580 ± 0.043	0.580 ± 0.049	0.682 ± 0.029	0.686 ± 0.004	0.602 ± 0.020	0.669
LD-CVAE	g.+h.	<u>0.620±0.045</u>	0.635±0.051	0.682±0.037	0.635±0.037	<u>0.653±0.036</u>	0.692
SlotSPE (8 slots/modality)	g.+h.	<u>0.620±0.058</u>	0.642±0.035	**0.688±0.051**	<u>0.705±0.037</u>	0.637±0.018	<u>0.706</u>
SlotSPE ($T=3$)	g.+h.	0.610±0.064	0.624±0.052	0.676±0.056	0.703±0.023	0.657±0.048	0.704
SlotSPE	g.+h.	**0.634±0.079**	0.642±0.035	**0.688±0.051**	**0.708±0.025**	<u>0.671±0.028</u>	**0.721**

D COMPARISONS WITH SOTAS

The complete results with standard deviations (std) for Table 1, using the biological pathway format for gene data, are provided in Table 4. In addition, we also report results using 8 slots per modality and $T = 3$ iterations in slot attention, with further ablations on the number of slots provided in Appendix F.3.1 and on the number of iterations in Appendix F.3.2. The results demonstrate that even with a very small number of slots for both histopathology and genomics, SlotSPE maintains strong performance, indicating that the slots effectively capture structured prognostic events through highly compressed representations.

To further assess the generalization ability of our approach, we compare SlotSPE with other multi-modal methods under a higher-level grouping of genes into six functional categories (Subramanian

Table 5: Comparison of C-index (mean ± std) across ten cancer datasets using **six gene functional groups**. Best and second-best results are in **bold** and <u>underline</u>.

Model	Modality	BRCA (N=1046)	COADREAD (N=573)	KIRC (N=488)	UCEC (N=488)	LUAD (N=467)	Overall
Porpoise	g.+h.	0.681±0.075	0.742±0.038	0.778±0.023	0.798±0.086	0.655±0.066	-
MCAT	g.+h.	0.706±0.038	0.736±0.024	0.797±0.018	0.800±0.065	0.641±0.062	-
MOTCat	g.+h.	0.683±0.031	<u>0.759±0.039</u>	0.791±0.010	<u>0.808±0.041</u>	**0.683±0.057**	-
CMTA	g.+h.	0.670±0.031	0.751±0.031	0.786±0.029	0.779±0.059	0.644±0.031	-
SurvPath	g.+h.	<u>0.707±0.049</u>	0.751±0.029	0.768±0.010	0.764±0.072	0.663±0.050	-
PIBD	g.+h.	0.683±0.016	0.733±0.030	0.791±0.019	0.774±0.054	0.649±0.077	-
MMP	g.+h.	0.684 ±0.081	0.699 ± 0.032	0.767±0.017	0.772±0.056	0.643±0.041	-
LD-CVAE	g.+h.	0.695±0.043	0.736±0.029	<u>0.797±0.025</u>	0.792±0.052	<u>0.677±0.050</u>	-
SlotSPE	g.+h.	**0.735±0.051**	**0.780±0.044**	**0.804±0.020**	**0.824±0.062**	0.669±0.063	-

Model	Modality	LUSC (N=460)	HNSC (N=438)	SKCM (N=409)	BLCA (N=381)	STAD (N=366)	Overall
Porpoise	g.+h.	0.585±0.065	0.617±0.058	0.674±0.035	0.683±0.040	0.620±0.036	0.683
MCAT	g.+h.	0.571±0.056	**0.658±0.025**	<u>0.679±0.041</u>	0.682±0.029	0.630±0.011	0.690
MOTCat	g.+h.	0.578±0.053	0.634±0.036	**0.688±0.049**	0.689±0.045	0.657±0.007	0.692
CMTA	g.+h.	0.582±0.048	<u>0.651±0.056</u>	0.667±0.031	0.682±0.039	0.627±0.045	0.684
SurvPath	g.+h.	0.564±0.033	0.636±0.036	0.658±0.064	**0.707±0.049**	0.637±0.021	0.686
PIBD	g.+h.	0.595±0.100	0.626±0.029	0.649±0.037	0.676±0.058	0.648±0.044	0.682
MMP	g.+h.	0.584±0.048	0.588±0.054	0.667±0.034	0.680±0.051	0.597±0.029	0.668
LD-CVAE	g.+h.	<u>0.613±0.051</u>	0.633±0.035	0.645±0.040	0.684±0.043	<u>0.663±0.041</u>	<u>0.693</u>
SlotSPE	g.+h.	**0.616±0.049**	0.645±0.030	<u>0.679±0.035</u>	<u>0.701±0.026</u>	**0.667±0.017**	**0.712**

et al., 2005; Liberzon et al., 2015; Chen et al., 2021b; Xu & Chen, 2023; Zhou & Chen, 2023): tumor supression, oncogenesis, protein kinases, cellular differentiation, transcription, and cytokines and growth. In this setting, the number of gene slots is fixed to six, while the number of histopathology slots is varied and the best performance is reported. Results in Table 5 show that SlotSPE achieves the best overall performance, ranking first or second in 8 out of 10 cohorts, and outperforming the second-best method (LD-CVAE) by 1.9% on average. This indicates that even when genes are grouped into higher-level functional categories, our SlotSPE is able to maintain strong predictive performance across different granularities of gene input.

E ROBUSTNESS AND RISK STRATIFICATION

The complete results with standard deviations for Table 2, using the biological pathway representation for gene data, are reported in Table 6. While most baselines are not explicitly designed to handle missing modalities, they can still be evaluated under missing-genomics settings by supplying a neutral placeholder input in place of genomic features (Zhang et al., 2025). For completeness, we also include two strong methods (MOTCAT and CMTA). The results show that when these approaches rely solely on histology—without modeling or reconstructing genomic information—their prognostic performance drops substantially. Moreover, under missing-genomics conditions, their limited robustness leads to performance even worse than strong single-modal baselines (e.g., ABMIL).

Kaplan–Meier curves for all ten datasets are shown in Figure 6. For clarity, Table 7 summarizes four complementary statistics: (a) log-rank p-values across all cancer types, (b) RMST values for both high- and low-risk groups, (c) the RMST difference (Δ = High–Low) with 95% confidence intervals, and (d) the RMST ratio (High/Low) with 95% confidence intervals. Across most datasets, SlotSPE consistently achieves larger RMST differences (Because Δ is typically negative, a "larger" separation corresponds to a more negative value, i.e. smaller numerically but larger in magnitude) and smaller RMST ratios than competing methods, demonstrating clearer prognostic separation. Even in datasets where performance is comparable (e.g., KIRC, LUSC, HNSC), SlotSPE generally maintains equal or better significance with tighter confidence intervals.
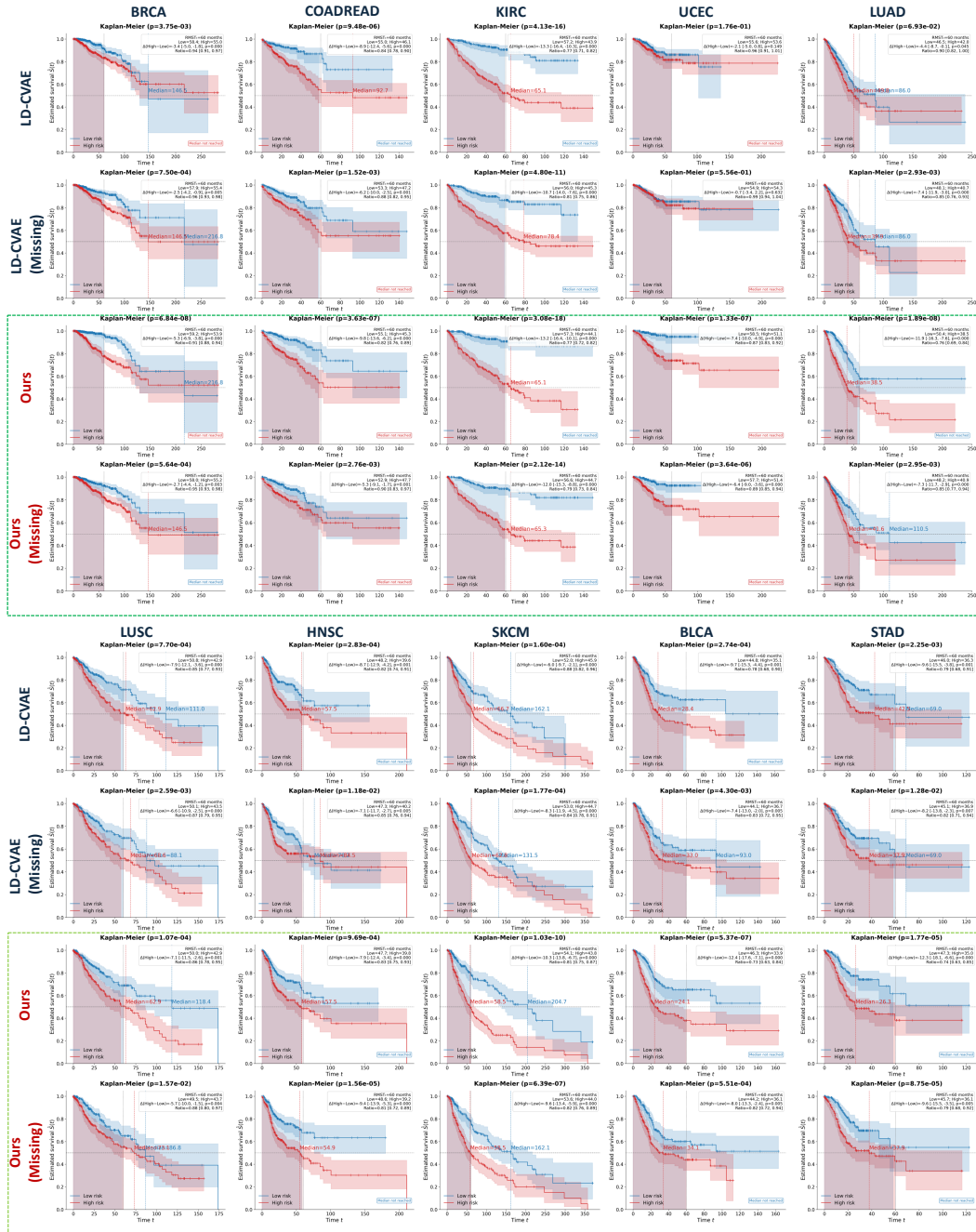
Figure 6: Kaplan–Meier curves of predicted high-risk and low-risk groups. A p-value < 0.05 at the top indicates statistically significant separation between groups. The restricted mean survival time (RMST) up to 60 months is also reported, with values shown as Δ (High–Low), and RMST ratio (High/Low). (Zoom in to view details.)

Table 6: C-index (mean $\pm$ std) across ten cancer datasets under missing-genomics settings. The column *Missing* indicates whether the genomic modality is missing ($\checkmark$ = missing). Best and second-best results are highlighted in **bold** and <u>underline</u>.

Model	Missing	BRCA (N=1046)	COADREAD (N=573)	KIRC (N=488)	UCEC (N=488)	LUAD (N=467)	Overall
MOTCAT	$\checkmark$	0.600±0.077	0.645±0.028	0.738±0.034	0.659±0.144	0.643±0.023	-
CMTA	$\checkmark$	0.604±0.112	0.683±0.065	0.662±0.045	0.676±0.124	0.593±0.051	-
LD-CVAE		0.705±0.050	<u>0.753±0.029</u>	0.792±0.025	0.788±0.068	0.651±0.055	-
LD-CVAE	$\checkmark$	0.715±0.056	0.736±0.042	<u>0.798±0.014</u>	0.775±0.073	0.638±0.016	-
SlotSPE		**0.779±0.061**	**0.773±0.042**	**0.815±0.018**	**0.813±0.059**	**0.683±0.057**	-
SlotSPE	$\checkmark$	<u>0.734±0.056</u>	0.741±0.040	0.789±0.029	<u>0.802±0.047</u>	<u>0.663±0.054</u>	-

Model	Missing	LUSC (N=460)	HNSC (N=438)	SKCM (N=409)	BLCA (N=381)	STAD (N=366)	Overall
MOTCAT	$\checkmark$	0.540±0.024	0.578±0.030	0.642±0.050	0.565±0.098	0.585±0.071	0.619
CMTA	$\checkmark$	0.541±0.045	0.564±0.074	0.541±0.048	0.558±0.054	0.608±0.065	0.603
LD-CVAE		0.620±0.045	0.635±0.051	<u>0.682±0.037</u>	0.635±0.037	0.653±0.036	0.692
LD-CVAE	$\checkmark$	0.596±0.042	**0.649±0.049**	0.676±0.038	0.644±0.048	0.647±0.049	0.688
SlotSPE		**0.634±0.079**	0.642±0.035	**0.688±0.051**	**0.708±0.025**	<u>0.671±0.028</u>	**0.721**
SlotSPE	$\checkmark$	**0.634±0.032**	**0.649±0.031**	0.678±0.068	<u>0.678±0.041</u>	**0.677±0.029**	<u>0.704</u>

Table 7: Summary of log-rank p-values and RMST-based statistics (RMST values, differences, and ratios) for group separation across ten cancer datasets in Figure 6.

Dataset	Method	Log-rank p-value	RMST (Low)	RMST (High)	Δ (High–Low) [95% CI]	Ratio (High/Low) [95% CI]
BRCA	LD-CVAE	3.75×10^{-3}	58.4	55.0	-3.4 [-5.0, -1.8]	0.94 [0.91, 0.97]
	SlotSPE	6.84×10^{-8}	59.2	53.9	-5.3 [-6.9, -3.8]	0.91 [0.88, 0.94]
COADREAD	LD-CVAE	9.48×10^{-6}	55.0	46.1	-8.9 [-12.4, -5.6]	0.84 [0.78, 0.90]
	SlotSPE	3.63×10^{-7}	55.1	45.3	-9.8 [-13.6, -6.2]	0.82 [0.76, 0.89]
KIRC	LD-CVAE	4.13×10^{-16}	57.2	43.9	-13.3 [-16.4, -10.3]	0.77 [0.71, 0.82]
	SlotSPE	3.08×10^{-18}	57.3	44.1	-13.2 [-16.4 -10.1]	0.77 [0.72, 0.82]
UCEC	LD-CVAE	1.76×10^{-1}	55.6	53.6	-2.1 [-5.0, 0.8]	0.96 [0.91, 1.00]
	SlotSPE	1.33×10^{-7}	58.5	51.1	-7.4 [-10.0, -4.9]	0.87 [0.83, 0.92]
LUAD	LD-CVAE	6.93×10^{-2}	46.5	42.0	-4.4 [-8.7, -0.1]	0.90 [0.82, 1.00]
	SlotSPE	1.89×10^{-8}	50.4	38.5	-11.9 [-16.3 -7.6]	0.76 [0.69, 0.84]
LUSC	LD-CVAE	7.70×10^{-4}	50.8	42.9	-7.9 [-12.1, -3.6]	0.85 [0.77, 0.93]
	SlotSPE	1.07×10^{-4}	50.0	42.9	-7.1 [-11.5, -2.6]	0.86 [0.78, 0.95]
HNSC	LD-CVAE	2.83×10^{-4}	48.2	39.6	-8.7 [-12.9, -4.2]	0.82 [0.74, 0.91]
	SlotSPE	9.69×10^{-4}	47.7	39.8	-7.9 [-12.4, -3.4]	0.83 [0.75, 0.93]
SKCM	LD-CVAE	1.60×10^{-4}	52.0	45.9	-6.0 [-9.7, -2.1]	0.88 [0.82, 0.96]
	SlotSPE	1.03×10^{-10}	54.1	43.8	-10.3 [-13.8, -6.7]	0.81 [0.75, 0.87]
BLCA	LD-CVAE	2.74×10^{-4}	44.8	35.1	-9.7 [-15.3, -4.4]	0.78 [0.68, 0.90]
	SlotSPE	5.37×10^{-7}	46.3	33.9	-12.4 [-17.6, -7.1]	0.73 [0.63, 0.84]
STAD	LD-CVAE	2.25×10^{-3}	46.0	36.3	-9.6 [-15.5, -3.8]	0.79 [0.68, 0.91]
	SlotSPE	1.77×10^{-5}	47.3	35.0	-12.3 [-18.1, -6.6]	0.74 [0.63, 0.85]

F ABLATION STUDY

F.1 COMPONENT VALIDATION

The complete results with standard deviations for Table 3, using the biological pathway representation for gene data, are reported in Table 8.

F.2 HISTOLOGY ENCODER ABLATIONS

To assess the robustness of SlotSPE to the choice of visual encoder, we first replace the histology encoder with a ResNet50 (Srivastava et al., 2015; He et al., 2016) pretrained only on ImageNet (Deng et al., 2009), rather than a pathology-specific foundation model. This setting tests whether SlotSPE

Table 8: Ablation of model components reported as C-index (mean ± std) across ten cancer datasets. Best and second-best results are in **bold** and <u>underline</u>.

Variants	BRCA (N=1046)	COADREAD (N=573)	KIRC (N=488)	UCEC (N=488)	LUAD (N=467)	Overall
Baseline	0.711±0.052	0.734±0.021	0.784±0.010	0.779±0.081	0.661±0.059	-
w/o Selective Slot Attention	0.728±0.057	0.755±0.030	0.787±0.011	0.800±0.064	0.659±0.067	-
w/o Slots Regularization	<u>0.734±0.065</u>	<u>0.763±0.029</u>	<u>0.794±0.017</u>	<u>0.811±0.069</u>	0.663±0.053	-
w/o Cross-modal Reconstruction	0.710±0.067	0.758±0.027	0.787±0.008	0.794±0.056	<u>0.663±0.062</u>	-
w/o Iterative Cross-attention	0.727±0.031	0.761±0.033	0.803±0.027	0.794±0.056	0.653±0.062	-
SlotSPE	**0.779±0.061**	**0.773±0.042**	**0.815±0.018**	**0.813±0.059**	**0.683±0.057**	-

Variants	LUSC (N=460)	HNSC (N=438)	SKCM (N=409)	BLCA (N=381)	STAD (N=366)	Overall
Baseline	0.592±0.064	0.610±0.056	0.668±0.043	0.686±0.030	0.644±0.021	0.687
w/o Selective Slot Attention	<u>0.620±0.071</u>	0.615±0.045	<u>0.685±0.052</u>	0.688±0.042	0.650±0.026	0.699
w/o Slots Regularization	0.611±0.060	<u>0.621±0.061</u>	0.679±0.053	0.693±0.038	<u>0.667±0.018</u>	<u>0.704</u>
w/o Cross-model Reconstruction	0.619±0.090	0.611±0.058	0.671±0.026	0.698±0.036	0.645±0.017	0.696
w/o Iterative Cross-attention	0.602±0.071	0.624±0.048	0.676±0.039	<u>0.702±0.024</u>	0.648±0.017	0.699
SlotSPE	**0.634±0.079**	**0.642±0.035**	**0.688±0.051**	**0.708±0.025**	**0.671±0.028**	**0.721**

Table 9: Comparison of C-index (mean ± std) across five datasets using ResNet50 as histology encoder. Best and second-best results are in **bold** and <u>underline</u>.

Methods	Modality	BRCA (N=1046)	COADREAD (N=573)	KIRC (N=488)	UCEC (N=488)	LUAD (N=467)	Overall
TransMIL	g.	0.608±0.029	0.641±0.019	0.661±0.036	0.690±0.081	0.599±0.054	0.640
CLAM_MB	g.	0.573±0.035	0.616±0.081	0.637±0.076	0.655±0.092	0.590±0.066	0.620
MCAT	g.+h.	0.694±0.074	<u>0.712±0.034</u>	<u>0.754±0.023</u>	0.771±0.062	0.633±0.058	<u>0.713</u>
MOTCAT	g.+h.	<u>0.706±0.062</u>	0.674±0.034	0.752±0.026	**0.790±0.053**	0.637±0.054	0.711
SurvPath	g.+h.	0.650±0.038	0.674±0.051	0.743±0.037	0.761±0.052	<u>0.643±0.046</u>	0.694
LD-CVAE	g.+h.	0.656±0.051	0.676±0.025	0.714±0.061	0.733±0.081	0.620±0.037	0.680
SlotSPE	g.+h.	**0.730±0.072**	**0.730±0.035**	**0.769±0.024**	<u>0.775±0.052</u>	**0.645±0.064**	**0.730**

can still achieve strong performance when the visual backbone lacks domain-specific pretraining. For this experiment, we evaluate representative unimodal and multimodal baselines listed in Table 1. As summarized in Table 9, SlotSPE continues to outperform all baselines under this weaker configuration, suggesting that its performance is not tightly coupled to a specialized encoder. We further extend this analysis by examining SlotSPE with recent pathology foundation models, including CONCH (Lu et al., 2024), CONCH v1.5 (used in TITAN (Ding et al., 2025)), and UNI (Chen et al., 2024). The results in Figure 7 show that performance varies across different encoders: while recent pathology foundation models generally improve results compared to ImageNet-pretrained encoder, the best performance is obtained with UNI. These results indicate that SlotSPE is both robust to weaker, non-specialized encoders and compatible with strong pathology-specific foundation models.

F.3 HYPERPARAMETERS ABLATIONS IN SLOT ATTENTION

F.3.1 SLOT NUMBERS ABLATIONS

Figure 8 illustrates how the number of slots for histology (S_h) and genomics (S_g) affects performance across six TCGA cohorts. In most cases, increasing the number of slots usually improves performance compared to the minimal baseline (red line), suggesting that additional slots enable the model to capture a broader set of prognostic events. However, beyond a moderate slot size (e.g., 16–32 slots), further increases often yield little to no gain, and in some cases even slight declines, indicating that excessive slots may introduce redundancy or prognostic irrelevant patterns.
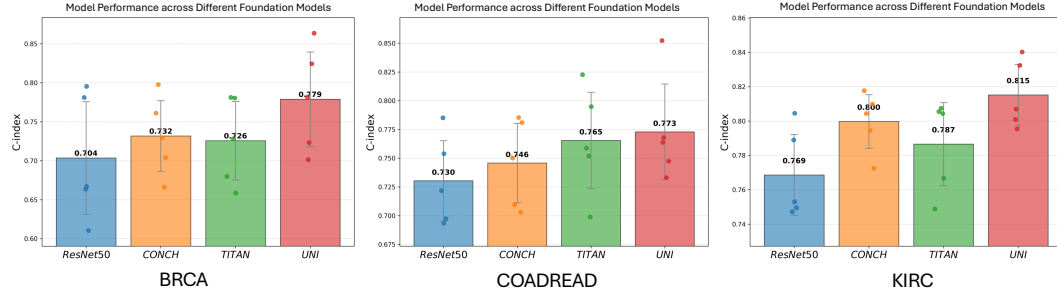
Figure 7: Effect of different foundation models used as histology encoders. Performance (C-index) is shown for three cohorts (BRCA, COADREAD, KIRC) using four backbone encoders: ResNet50, CONCH, TITAN, and UNI. Bars represent the mean across five folds, with error bars indicating standard deviation. Higher values indicate better performance.
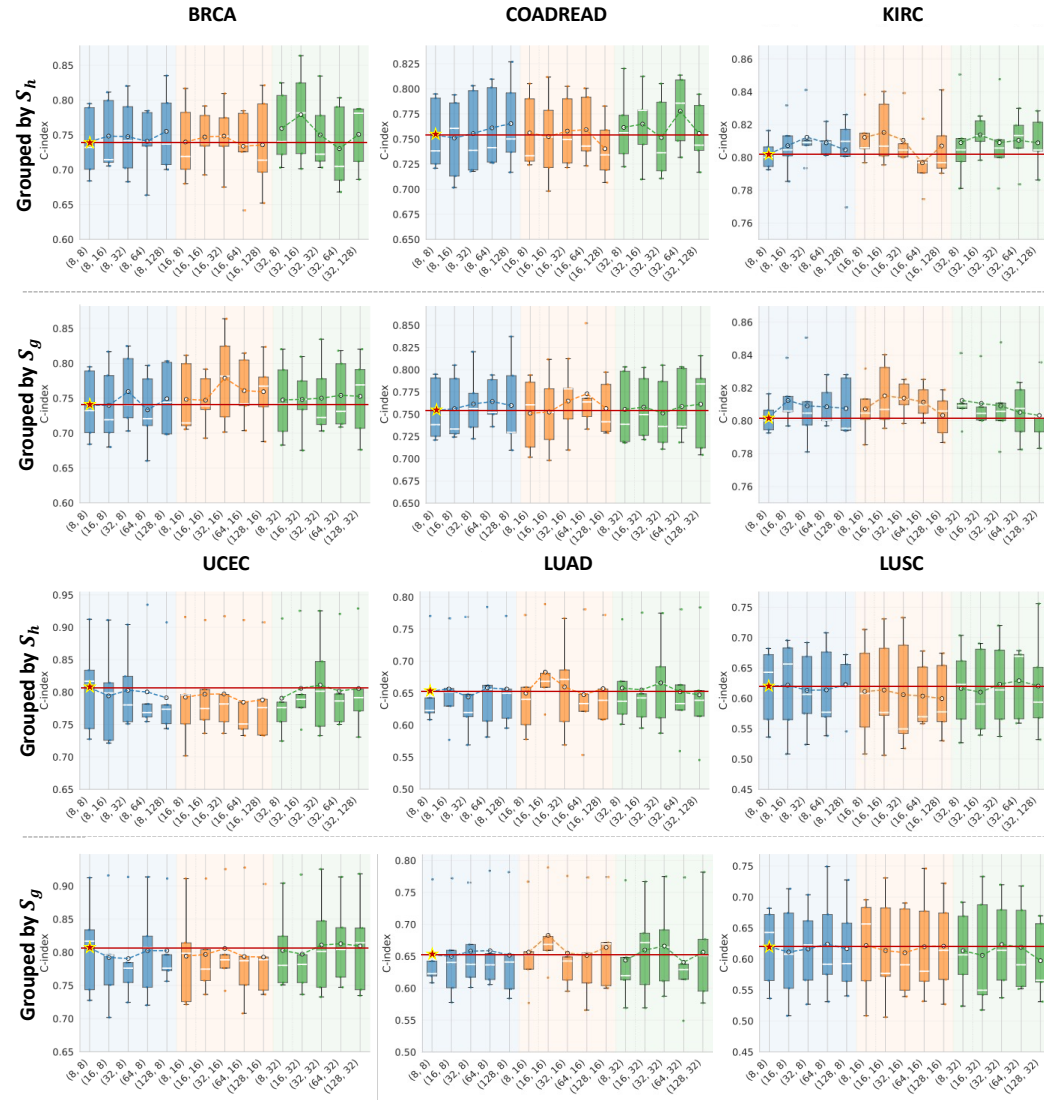

Figure 8: Ablation on slot numbers. Boxplots show C-index across cohorts as the number of histology slots (S_h, top) and genomic slots (S_g, bottom) varies, with the x-axis indicating the pair (S_h, S_g). The red line denotes the baseline with the minimum slot setting. Increasing the number of slots generally improves performance, though gains plateau once saturation is reached.
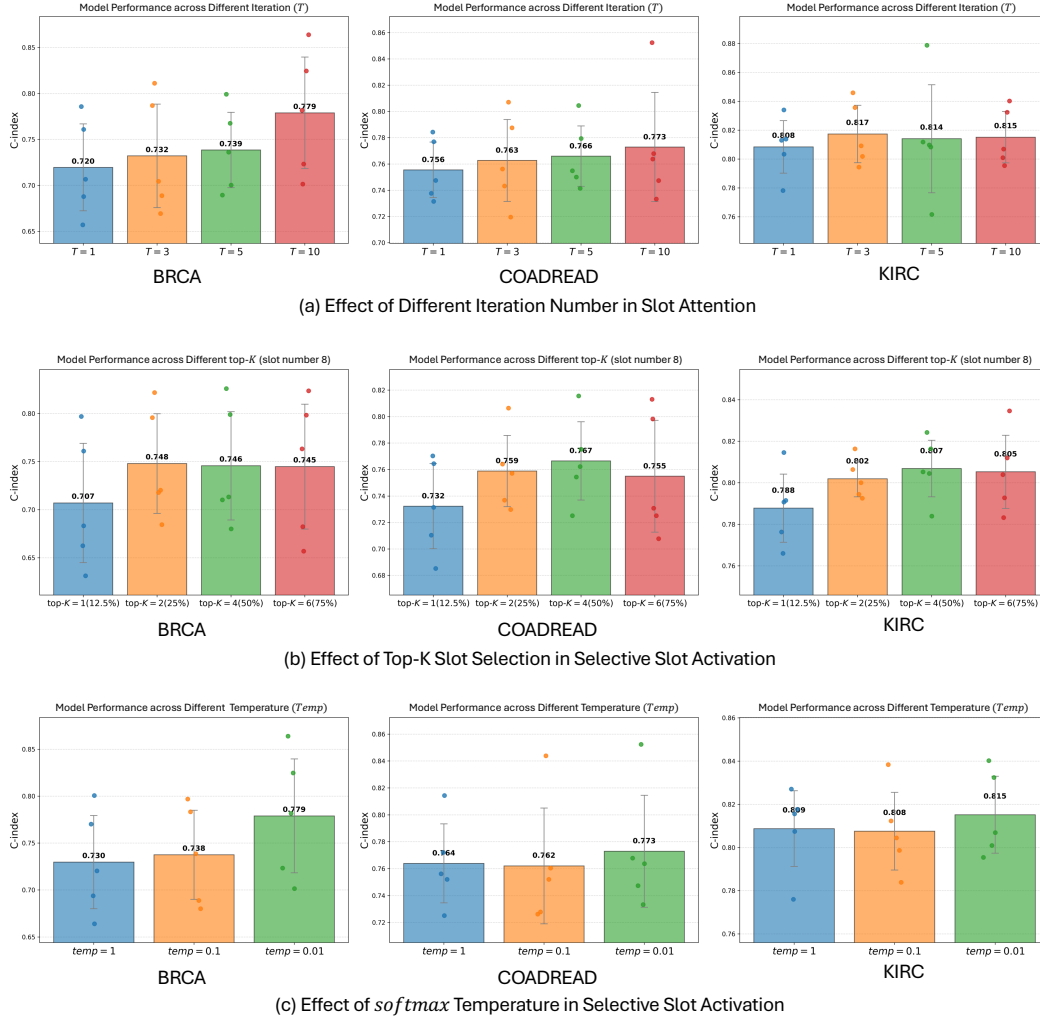
Figure 9: Ablation study of hyperparameters. (a) Effect of different iteration number T in the slot attention. (b) Effect of different top-K(percentage of slots) in selective slot activation (slot number is fixed as 8). (c) Effect of $softmax$ temperature on in selective slot activation. Bars represent mean across five folds, with error bars indicating standard deviation. Higher values indicate better performance.

F.3.2 SLOT ATTENTION ITERATIONS ABLATIONS.

We ablate the number of iterations T in slot attention to examine its impact on performance shown in Figure 9(a). With a small number of iterations ($T = 1$), slots are updated only once, limiting their ability to refine representations and capture complex prognostic patterns, which results in lower C-index values (e.g., 0.720 on BRCA). As T increases, slot updates stabilize, enabling better information bindings and improved outcomes (best at $T = 10$ on BRCA with 0.779). In datasets such as KIRC, performance saturates around $T = 3$–5, suggesting diminishing returns after sufficient updates. Although larger T enhances stability and accuracy, it also increases training time. Overall, our results indicate that moderate values provide a favorable trade-off between performance and efficiency. For consistency, we also report main results under the $T = 3$ setting in Table 4.

F.4 HYPERPARAMETERS ABLATIONS IN SELECTIVE SLOT ACTIVATION

F.4.1 TOP-K SLOT SELECTION ABLATIONS

We analyze the effect of top-K selection by fixing the slot number to 8 for both modalities and evaluate four settings: top-1 (12.5% all slots), top-2 (25% all slots), top-4 (50% all slots), and top-6 (75% all slots), as shown in Figure 9 (b). Using only top-1 results in the weakest performance, which is expected, because patients are divided into four risk groups, activating only a single slot per patient causes several slots to be rarely activated during training. Conversely, when top-K becomes too large (e.g., top-6), some slots that are not discriminative for a given patient are still forced to activate. While the 25% and 50% settings perform similarly across datasets, we adopt the 25% configuration because it aligns naturally with our design: with four risk groups, selecting roughly one-quarter of the slots provides a trade-off between sparsity and coverage.

F.4.2 SOFTMAX TEMPERATURE ABLATIONS

We further evaluate the effect of $softmax$ temperature (denoted as $temp$) in selective slot activation as shown in Figure 9 (c). A higher temperature (e.g., $temp = 1$) produces soft, diffuse gating and results in weaker performance. Lowering the temperature sharpens the gating distribution and leads to more discriminative slot selection, generally improving results. The setting $temp = 0.01$ achieves the best C-index across three datasets, and we therefore adopt it in our experiments.

G   EFFICIENCY ANALYSIS

In this section, we analyze the efficiency of our model by comparing it with the four strongest baselines in Table 1 without missing modality setting on identical hardware (An NVIDIA RTX A6000, 48GB). Since all methods use the same WSI and omics encoders, we omit these components and focus our comparison solely on the remaining model modules for a more meaningful efficiency assessment.

G.1   INFERENCE RUNTIME AND MEMORY

We compare inference (where all patches are used) peak memory and wall-clock time against model performance on the COADREAD dataset. In this comparison, we set our model with slot number as 16 and iteration as 3. As shown in Figure 10(a), SlotSPE achieves the highest C-index while maintaining low runtime and memory usage. MCAT and MOTCAT achieve shorter runtime primarily because they omit self-attention computations within WSI patches; however, this design choice leads to a considerable drop in predictive performance. CMTA improves upon these baselines by employing Nyström attention (Xiong et al., 2021) as an approximation to full self-attention, but its capacity remains insufficient to capture the sparse, patient-specific prognostic signals. LD-CVAE attains the second-best performance, yet its reliance on a variational autoencoder introduces substantial memory and runtime overhead.

G.2   TRAINING RUNTIME AND MEMORY

We also compare training-time runtime and memory usage. Since the official implementation of MCAT, MOTCAT and LD-CVAE don't support batch training, we set the batch size to 1 for all methods and randomly sample 4,096 patches to ensure a fair comparison. For SlotSPE, we retain the same configuration of 16 slots and 3 iterations. As shown in Figure 10(b), SlotSPE exhibits moderately higher memory usage and training time compared with its inference cost, primarily due to the additional reconstruction component that is active during training. Nonetheless, its overall training resource footprint remains well balanced and sits near the middle of all evaluated methods.

G.3   MODULE-LEVEL COST BREAKDOWN

To identify the main computational bottlenecks, Figure 10(c) reports the module-level breakdown of GPU memory consumption and runtime within SlotSPE. The reconstruction component dominates both memory (45.0%) and runtime (55.4%), accounting for the largest share of training cost. Slot attention ranks second, contributing 26.2% of memory usage and 36.9% of runtime. The remaining modules—including the MoE-style decoder, slot interaction, and the final predictor—together account for a relatively small proportion of the overall cost. The heavier cost of the reconstruction branch arises from the inclusion of two slot regularizations for both modalities and one cross-modal reconstruction. These reconstruction components are used only during training and are disabled at inference unless the omics modality is missing, in which case the cross-modal reconstruction head is activated.

(a) Performance vs (Inference) Memory/Runtime Trade-off

(b) Training Memory/Runtime Comparison

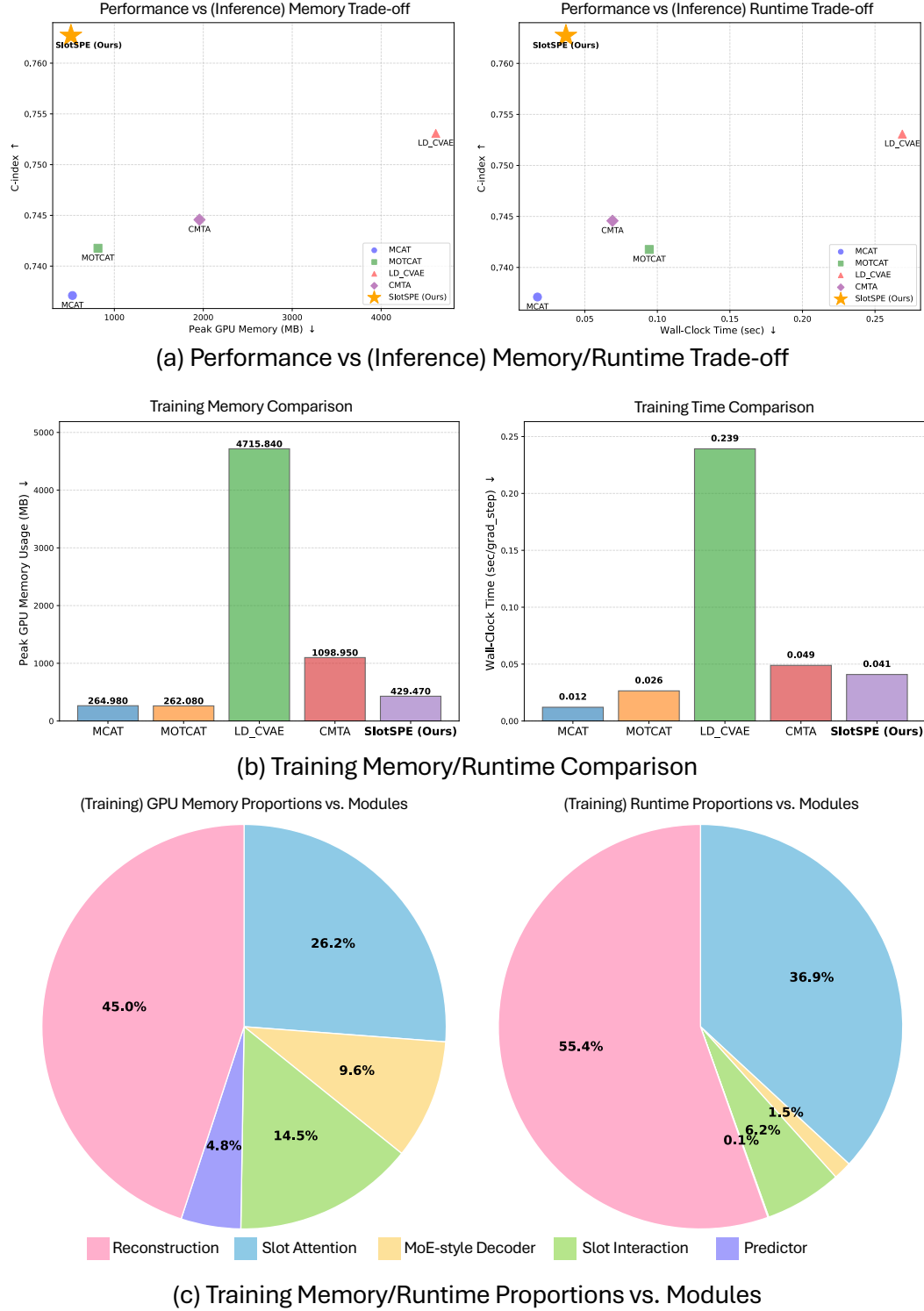(c) Training Memory/Runtime Proportions vs. Modules

Figure 10: Efficiency analysis. (a) Performance vs. inference memory/runtime trade-off across compared models. (b) Training memory/runtime comparison under a unified batch-size setting. (c) Module-level breakdown of training GPU memory/runtime proportions.

Table 10: Multivariable Cox analysis with clinical variables and SlotSPE predictions. Comparison of clinical-only vs. clinical + model configurations using AUC, IBS, C-index, and ΔC-index across five cohorts. The clinical variables of the TCGA cohort were downloaded from cBioPortal.

Cohort	AUC ↑		IBS ↓		C-index ↑		ΔC-index
	clinical	clinical+model	clinical	clinical+model	clinical	clinical+model	
BRCA	0.731±0.051	0.782±0.066	0.134±0.021	0.133±0.022	0.768±0.058	0.816±0.071	0.053±0.064
COADREAD	0.778±0.053	0.782±0.062	0.138±0.017	0.140±0.030	0.780±0.072	0.819±0.058	0.040±0.022
KIRC	0.858±0.049	0.889±0.039	0.137±0.020	0.110±0.021	0.844±0.013	0.868±0.015	0.023±0.014
LUAD	0.722±0.114	0.726±0.085	0.176±0.008	0.174±0.011	0.683±0.080	0.739±0.061	0.054±0.044
LUSC	0.681±0.084	0.747±0.111	0.185±0.018	0.166±0.030	0.652±0.059	0.699±0.077	0.047±0.057

(a) Calibration Curve (60 months)　　(b) Decision Curve (Positive Net Benefit Region)　　(c) Kaplan–Meier (KM) Curve ($p = 3.08 \times 10^{-18}$)
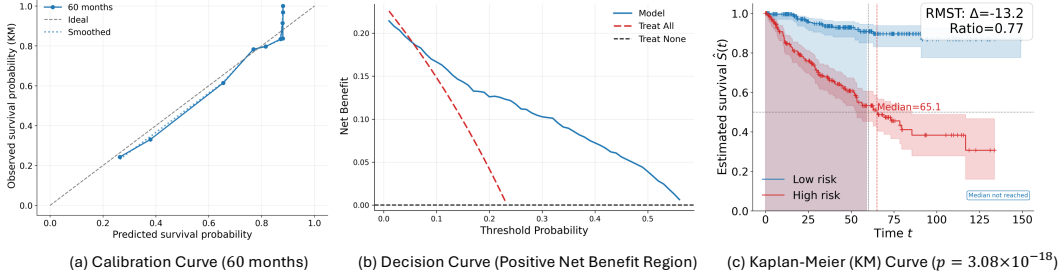
Figure 11: Calibration curve, decision curve analysis, and Kaplan–Meier risk stratification for KIRC.

H CLINICAL UTILITY

H.1 MULTIVARIABLE SURVIVAL ANALYSIS

To assess whether SlotSPE provides prognostic value beyond clinical variables (age, sex, stage, and neoadjuvant treatment), which are empirically shown as crucial prognostic factors (Xu et al., 2019; Tomasello et al., 2017; Edge & Compton, 2010; Yu et al., 2022), we performed multivariable Cox regression (Cox, 1972) using both clinical covariates and our model's predicted risk scores. For each cohort, we merged the risk prediction of SlotSPE with these clinical variables on a per-patient basis. Importantly, the risk scores were obtained from the validation folds, ensuring that the model had no access to these samples during training. We then fitted Cox models across the five validation folds to quantify the incremental prognostic contribution of SlotSPE. In Table 10, we report the time-dependent AUC (Heagerty et al., 2000; Heagerty & Zheng, 2005), integrated brier score (IBS) (Graf et al., 1999; Gerds & Schumacher, 2006), and C-index for two configurations: (i) *clinical-only*, using only the clinical covariates, and (ii) *clinical + model*, where the predicted risk score of SlotSPE is included as an additional covariate. We also report the corresponding improvement ΔC-index. Across all five cohorts, adding the risk score of SlotSPE to the clinical covariates consistently improves the C-index, with ΔC-index ranging from 0.023 (KIRC) to 0.054 (LUAD). Time-dependent AUC likewise increases or remains comparable when including the model, while IBS is generally stable or reduced (e.g., substantial IBS reduction for KIRC and LUSC). These results indicate that Slot-SPE captures complementary prognostic information that is not fully explained by routine clinical variables, showing the clinical utility potential to incorporate our model's prediction as a clinically useful prognostic factor when combined with established clinical factors.

H.2 CALIBRATION, DECISION ANALYSIS, AND RISK STRATIFICATION

To further evaluate the clinical utility of SlotSPE, we examine calibration and decision curve analysis (DCA) on the KIRC cohort. As shown in Figure 11(a), the 60-month calibration curve demonstrates strong agreement between predicted and observed survival probabilities, indicating that SlotSPE provides well-calibrated risk estimates. The decision curve in Figure 11(b) shows that SlotSPE yields positive net benefit across a clinically meaningful range of threshold probabilities. Within this range, SlotSPE consistently yields higher net benefit than both the "treat-all" and "treat-none" strategies, demonstrating potential practical value for supporting clinical decision-making. Finally,

the KM analysis in Figure 11(c) shows a clear and statistically significant separation between high-and low-risk groups, confirming that SlotSPE offers meaningful prognostic stratification.

I INTERPRETABILITY

I.1 PATIENT-LEVEL INTERPRETABILITY

Cross-modality Alignment. We visualize slot assignment maps for both histological and genomic slots (Figure 12A). Slots from both modalities consistently segment similar cellular and tissue regions within the WSI, indicating that SlotSPE captures event-level alignment across modalities. At the same time, differences emerge: morphology-driven slots emphasize structural tissue patterns, whereas genomics-driven slots highlight regions reflecting underlying molecular features.

Intra- and Inter-modal Interactions. For the histological modality (Figure 12B, first row), we display the attention map for its assigned slot, and further show the top-5 patches with the highest attention. An experienced pathologist provided expert annotations for each of these patches. These attention maps provide insight into which tissue regions the model identifies as the most prognostically informative. For the pathway modality, we visualize the assignment maps of pathways for omics-derived slots and also report the top-3 most relevant and irrelevant pathways per slot, revealing how the model differentiates biologically meaningful signals (Figure 12C). Then, to analyze cross-modal interactions, we connect pathway-level attention scores with the corresponding attention map of genomic slot on WSIs (Figure 12B, second row). Another case in UCEC is shown in Figure 13.

Overall, the interpretability analysis highlights the ability of the model to establish event-level alignment across modalities, as well as showing the potential to make novel hypotheses by identifying unexpected pathway–morphology correspondences for further biological validation.
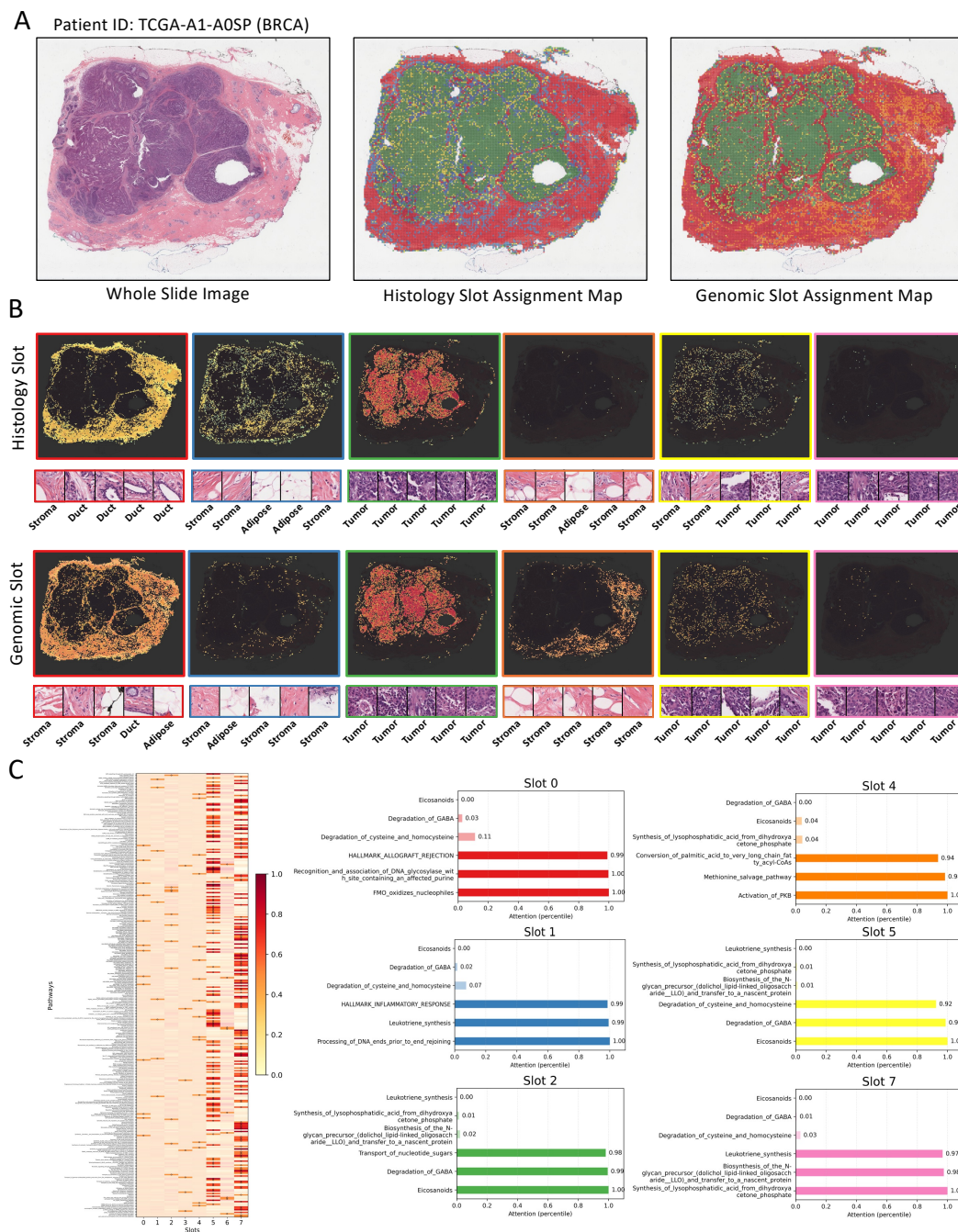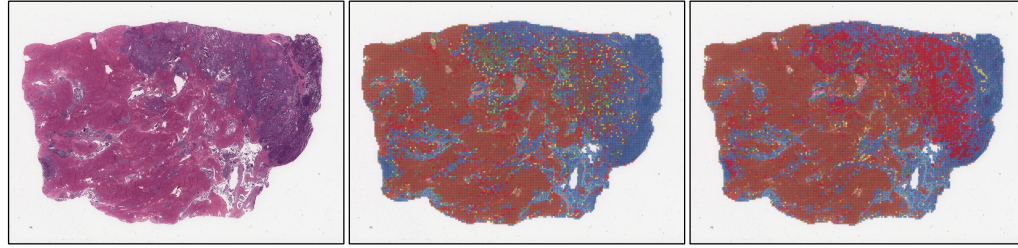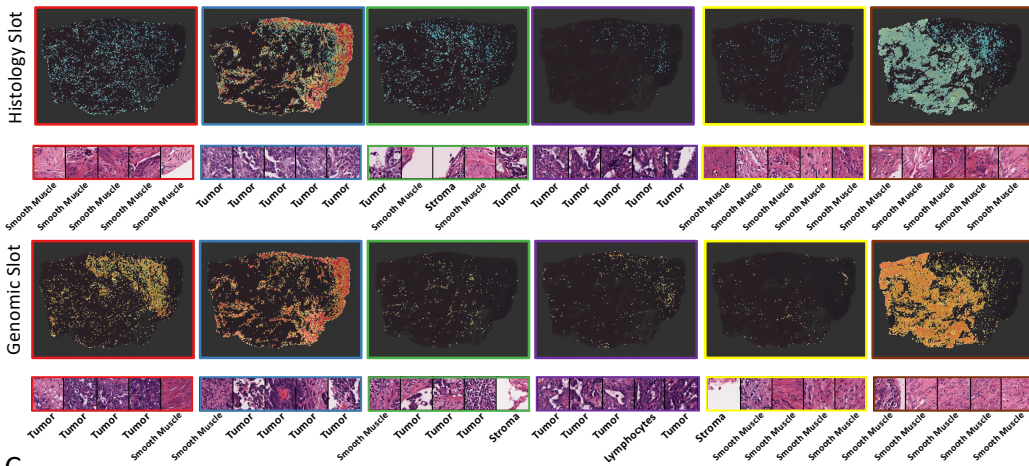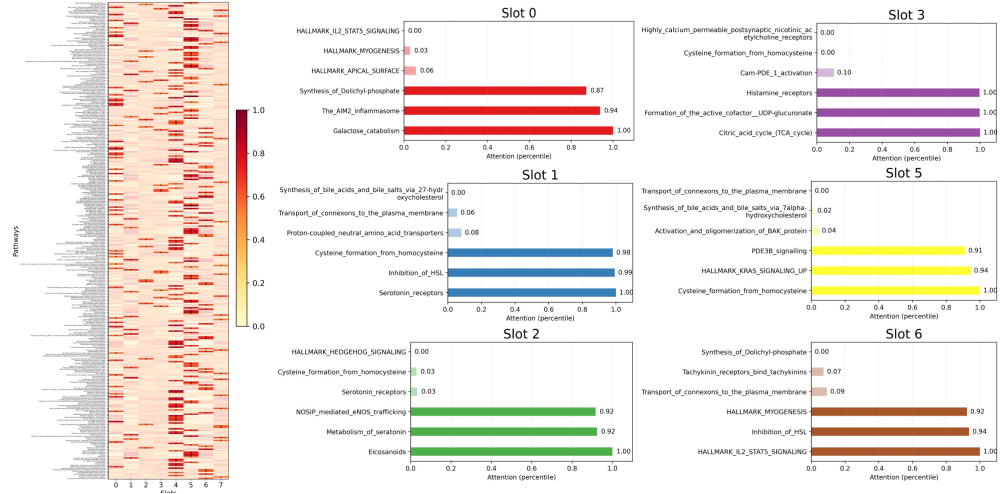
Figure 12: Case study of patient-level interpretability in BRCA. (A) From left to right: original WSI, assignment map of histology-derived slots, and assignment map of omics-derived slots. (B) Attention maps for each slot highlighting the top-5 most relevant patches. (C) Assignment maps of pathways for omics-derived slots with corresponding attentions, showing the top-3 relevant and irrelevant pathways per slot.

Figure 13: Case study of patient-level interpretability in UCEC. (A) From left to right: original WSI, assignment map of histology-derived slots, and assignment map of omics-derived slots. (B) Attention maps for each slot highlighting the top-5 most relevant patches. (C) Assignment maps of pathways for omics-derived slots with corresponding attentions, showing the top-3 relevant and irrelevant pathways per slot.
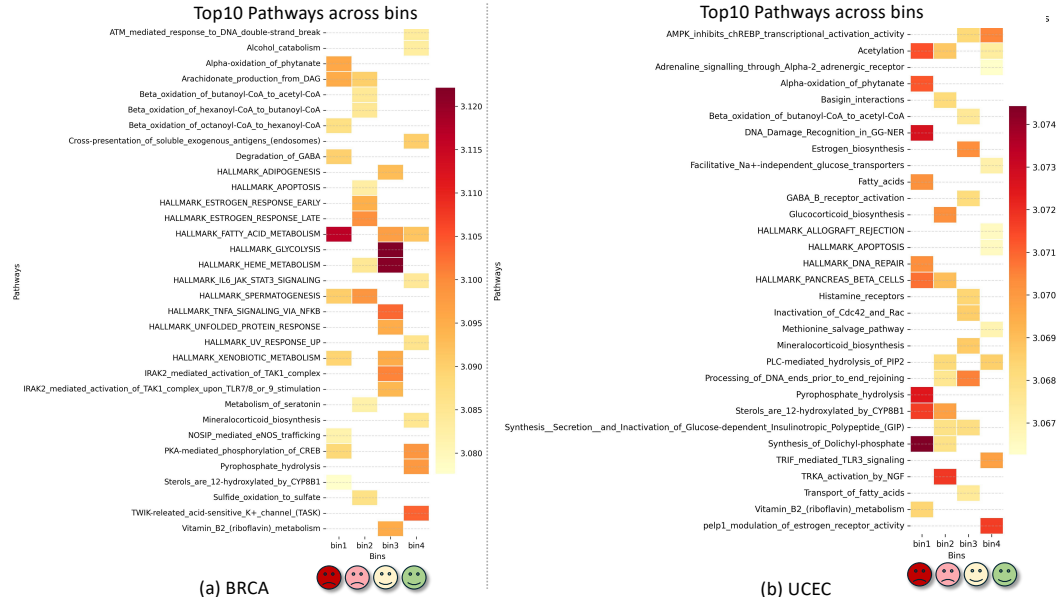
Figure 14: Cohort-level pathway analysis for BRCA and UCEC. Patients are stratified into four risk groups, and top-10 pathways are identified per group based on slot attentions. Distinct patterns emerge across bins, highlighting biologically meaningful differences between high- and low-risk groups.

I.2 COHORT-LEVEL INTERPRETABILITY

To extend interpretability beyond individual patients, we analyze cohort-level pathway attention patterns in the genomic modality. For each patient, we identify the slot with the highest retention score from the MoE decoder and extract its pathway-level attentions. Patients are then stratified into four risk groups (bins) based on their survival time. Within each group, pathway attentions are averaged to obtain group-level profiles, and the top-10 pathways per bin are reported (Figure 14).

In BRCA, for example, high-risk patients (bin1) are enriched in fatty acid and xenobiotic metabolism pathways, reflecting the metabolic reprogramming characteristic of aggressive tumors (Pavlova & Thompson, 2016; Röhrig & Schulze, 2016). In contrast, low-risk patients (bin4) exhibit stronger DNA repair and immune-related pathways, consistent with improved prognosis (Nolan et al., 2017; Denkert et al., 2018). Such cohort-level analyses illustrate the potential of SlotSPE to uncover pathway–risk associations and generate hypotheses for further biological validation.