Together, Then Apart: Balancing Alignment and Distinctiveness for Multimodal Survival Analysis

Wenjing Liu[1,2]*, Qin Ren[1]*, Wen Zhang[1,3], Yuewei Lin[4], and Chenyu You[1]✉

[1] Stony Brook University, Stony Brook, NY, USA
[2] Stanford University, Stanford, CA, USA
[3] Johns Hopkins University, Baltimore, MD, USA
[4] Brookhaven National Laboratory, Upton, NY, USA
chenyu.you@stonybrook.edu
https://y-research-sbu.github.io/TTA

Abstract. Multimodal survival analysis aims to improve cancer prognosis using heterogeneous biomedical data, such as histopathology images and genomic profiles. A common strategy is to align representations across modalities so that shared signals can be captured. However, strong cross-modal alignment can also remove modality-specific evidence that is critical for survival prediction. In this paper, we revisit multimodal survival learning from a simple observation: effective models should first discover shared patterns across modalities, and then preserve modality-specific signals. This motivates a representation learning principle that we refer to as *Together Then Apart*. Based on this idea, we propose **TTA**, a framework that balances cross-modal alignment and representation distinctiveness. TTA first performs prototype-based alignment to capture shared survival-related structures between modalities. It then encourages modality-specific distinctiveness through an anchor-guided contrastive objective. To further account for modality imbalance and noisy correspondences, we model cross-modal interactions using unbalanced optimal transport. We evaluate the proposed approach on multiple TCGA cancer cohorts with paired histopathology and genomic data. TTA consistently improves survival prediction over recent multimodal survival models. Moreover, the learned prototype structures reveal interpretable cross-modal patterns associated with clinical outcomes. Code is available at https://github.com/Y-Research-SBU/TTA.

Keywords: Multimodal Learning · Survival Analysis · Computational Pathology · Cross-Modal Representation Learning

1 Introduction

Multimodal survival analysis seeks to improve cancer prognosis by combining heterogeneous biomedical data sources. Among them, pathology whole-slide im-

* Equal contribution.
✉ Corresponding author.

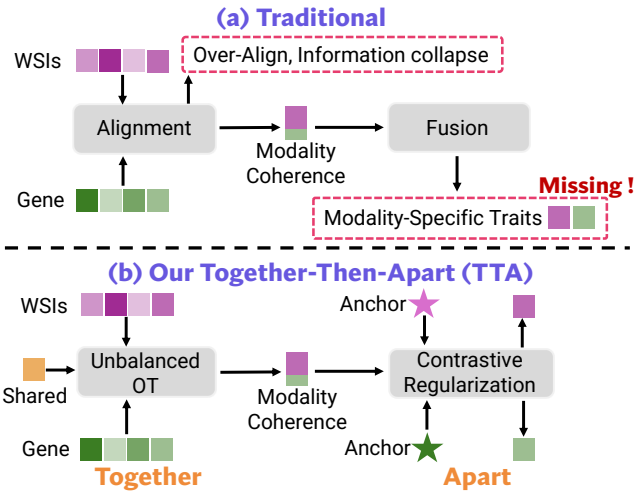

Fig. 1: Traditional multimodal learning vs. our TTA framework. Conventional approaches align heterogeneous modalities in a shared representation space, which can suppress modality-specific signals. TTA follows a *Together–Then–Apart* strategy: modalities are first aligned to capture shared patterns (*Together*), and then separated to preserve modality-specific information (*Apart*).

ages (WSIs) and genomic profiles provide two complementary views of tumor biology. WSIs capture rich morphological patterns and histopathological biomarkers and are typically modeled using Multiple Instance Learning (MIL) frameworks that aggregate patch-level features into slide-level representations [19, 20, 34, 44]. Genomic measurements, often derived from transcriptomic or pathway-level analyses [15,24,32], reveal molecular programs and mutation-driven signals that are not observable in tissue morphology [29]. Combining these modalities has therefore become a central direction in multimodal survival modeling [8–10, 14, 21, 28, 41]. Most recent methods adopt attention-based fusion mechanisms to learn joint representations that capture cross-modal correlations [8, 9, 21, 37].

A common assumption behind these models is that aligning heterogeneous modalities in a shared latent space will improve prediction. However, recent evidence suggests that this assumption does not always hold [31,42,49]. When alignment is enforced too aggressively, multimodal representations can lose modality-specific structure. This phenomenon, which we refer to as *over-alignment collapse*, occurs when heterogeneous signals are prematurely forced into a shared representation space. In practice, this may dilute morphology-rich spatial patterns in WSIs and gene-level variation in genomics, reducing representational diversity and limiting the benefit of multimodal integration. Empirical observations further reveal a surprising outcome: multimodal models can sometimes underperform strong WSI-only baselines [19,25,33,34], which naturally preserve detailed morphological information.

Several recent works attempt to mitigate this issue. PIBD [49] disentangles shared and modality-specific representations, while MRePath [31] dynamically reweights modality contributions. Although these approaches partially alleviate redundancy, they do not fully resolve the tension between cross-modal alignment and modality-specific information preservation. In particular, most existing methods treat these objectives independently, lacking a unified principle that jointly encourages semantic agreement across modalities while maintaining the distinctive signals each modality provides.

In this work, we revisit multimodal survival modeling from this perspective. Our key observation is simple: effective multimodal representations should *first discover what modalities share, and only then preserve what makes them different*. We term this learning principle **Together-Then-Apart**. Following this idea, we propose **TTA**, a framework that explicitly balances cross-modal alignment and modality-specific distinctiveness. In the *Together* stage, TTA aligns modalities by projecting both WSI and genomic embeddings onto a shared set of prototypes that serve as semantic anchors. Unlike prior approaches that maintain separate prototype spaces for each modality [37, 49], our shared-prototype design captures cross-modal structure while avoiding premature homogenization. To further accommodate intra-sample heterogeneity, we introduce an **unbalanced optimal transport (UOT)** formulation that learns entropy-regularized instance-to-prototype assignments under relaxed marginal constraints. This allows the model to selectively emphasize informative regions and produce stable MIL representations. In the *Apart* stage, TTA preserves modality-specific information. We introduce modality-specific anchors that retain distinctive semantics for WSIs and genomics. A contrastive regularization encourages features to remain close to their modality anchors while maintaining separation across modalities, preventing representational collapse and complementing the alignment learned in the *Together* stage. Through this alternating process, TTA produces multimodal representations that capture both shared prognostic signals and modality-specific evidence. Our contributions are as follows:

- We identify *over-alignment collapse* as a key challenge in multimodal survival analysis, where aggressive cross-modal alignment suppresses modality-specific signals and limits the benefit of multimodal integration.
- We propose **Together-Then-Apart (TTA)**, a framework that balances cross-modal alignment and modality-specific distinctiveness. TTA aligns modalities through shared prototypes while preserving distinctive information using modality-specific anchors and contrastive regularization.
- Extensive experiments on five TCGA cohorts demonstrate that TTA consistently improves survival prediction over recent multimodal survival models and reveals interpretable cross-modal structures related to clinical outcomes.

2 Related Works

2.1 Single-Modality Survival Analysis

Recent studies have explored survival prediction using a single data modality, most commonly histopathology whole-slide images (WSIs) or genomic profiles. Due to the gigapixel resolution of WSIs, most approaches formulate the task within a Multiple Instance Learning (MIL) framework, where a slide is represented as a bag of image patches. A large body of work focuses on how to aggregate patch-level features into slide-level representations. Early methods employ attention-based aggregation to assign importance weights to instances before pooling [19,27]. Subsequent works extend this paradigm by modeling contextual dependencies among patches using sequence-based architectures [34, 43]. Other approaches exploit the hierarchical organization of WSI patches through multi-scale or hierarchical frameworks [4,35]. Graph-based methods have also been proposed to model spatial relationships between tissue regions and capture context-aware interactions [7]. Another line of research focuses on reducing redundancy among patch tokens through prototype-based abstraction [11, 36, 39, 44] or instance filtering mechanisms that select informative regions for downstream prediction [23,25,33,47,48]. In parallel, genomic survival modeling typically relies on feedforward neural networks or self-normalizing neural architectures to process transcriptomic or pathway-level features [17,22]. Although these single-modality approaches provide strong predictive baselines and robust modality-specific representations, they capture only one aspect of tumor biology. Integrating complementary modalities therefore becomes a natural direction for improving survival prediction.

2.2 Multimodal Survival Analysis

Recent research efforts increasingly focus on integrating heterogeneous modalities [16,40,45,46], particularly histopathology and genomics, for prognostic modeling. Many multimodal methods adopt attention-based mechanisms to capture interactions between modalities and learn joint representations for survival prediction [1,9,21]. Beyond attention-based approaches, optimal transport has also been explored as a mechanism for aligning pathology and genomic representations in a shared space [37, 41]. To improve scalability, MMP [37] introduces morphological prototypes to compress redundant WSI tokens. PIBD [49] adopts an information-theoretic framework to disentangle shared and modality-specific representations and reduce redundancy. MRePath [31] further introduces dynamic reweighting to rebalance modality contributions. Other studies address robustness in the presence of missing modalities, including LD-CVAE [50] and DisPro [42]. Despite these advances, most existing approaches focus primarily on modeling cross-modal correlations through alignment. How to capture shared structures across modalities while preserving modality-specific signals remains an open challenge in multimodal survival analysis.

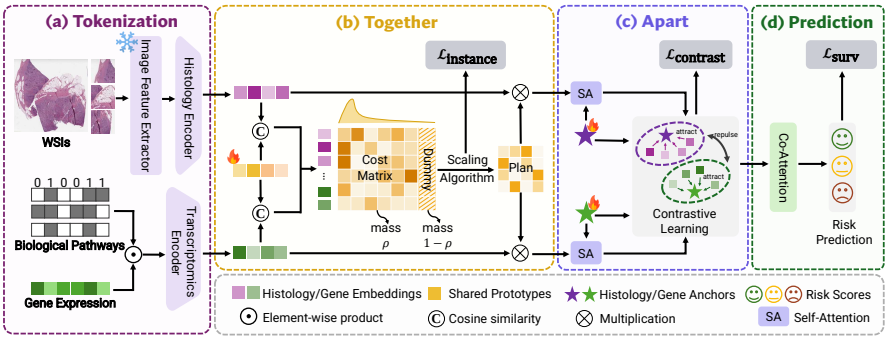

Fig. 2: Overview of TTA. (1) *Tokenization:* WSIs and gene-expression profiles are converted into modality-specific token sets. (2) *Together:* All tokens are assigned to a shared prototype bank via a semi-relaxed unbalanced optimal transport (UOT) module with a curriculum on the mass parameter ρ. (3) *Apart:* Modality representations are refined with modality-specific anchors and a contrastive regularizer to retain modality identity. (4) *Prediction:* A co-attention module exchanges information between modalities and outputs the survival risk.

2.3 Survival Analysis with Optimal Transport

Optimal Transport (OT) has recently emerged as an effective tool for modeling structural relationships and heterogeneity in survival data. In WSI analysis, unbalanced OT has been used for imbalanced clustering and instance-level patch selection, allowing models to focus on salient regions that drive prognosis [33]. A similar idea can be applied to genomic features, where transporting mass toward informative pathway tokens yields sparse and biologically meaningful representations. Beyond within-modality modeling, OT has also been explored for cross-modal alignment between pathology and genomic representations [37, 41]. Existing OT-based multimodal approaches typically compute transport plans either between modality-specific embeddings [41] or between modality-specific prototype assignments [37]. In contrast, our method computes a single transport plan after concatenating pathology and genomic tokens and transporting them to a shared prototype bank. This formulation enables cross-modal evidence to interact under a unified assignment geometry while maintaining flexibility through an unbalanced OT formulation.

3 Method

We consider survival prediction from two complementary patient views: pathology WSIs and transcriptomics. A common practice is to tightly align the two modalities, but in survival analysis this can be brittle: WSIs contain many ambiguous patches, and transcriptomic signals are sparse and patient-dependent.

If the model is forced to agree everywhere, modality-specific cues can be washed out, and the combined model may even fall behind strong single-modality baselines. TTA follows a simple principle: *align where the modalities agree, and keep them separated where they should differ*. As shown in Fig. 2, TTA consists of two stages. TOGETHER (Sec. 3.2) aligns both modalities through shared prototypes using an unbalanced optimal transport assignment, designed to tolerate uncertainty and long-tailed heterogeneity. APART (Sec. 3.3) then re-introduces modality identity using modality-specific anchors and a lightweight contrastive objective. Finally, a co-attention prediction head aggregates complementary evidence for survival risk estimation (Sec. 3.4).

3.1 Problem Formulation

We represent each patient with two token sets, one per modality, so that alignment and refinement can be performed at the token level [37].

WSI tokenization. Each WSI is divided into patches and encoded by a pretrained image encoder such as ResNet50 or UNI [6]. A linear projection maps patch embeddings to D dimensions, yielding pathology tokens $\boldsymbol{X}_n^p \in \mathbb{R}^{N_n^p \times D}$ for patient n, where N_n^p is the number of patches and $\boldsymbol{x}_{n,i}^p$ denotes the i-th patch token.

Gene-expression tokenization. We formulate transcriptomics into pathway-level tokens to obtain compact and interpretable genomic factors. Following MMP [37], we use a fixed pathway collection indexed by $c \in \{1, \ldots, C_g\}$. Each pathway is specified by a binary selector $\boldsymbol{a}_c \in \{0,1\}^G$ over G genes. Given the gene-expression vector $\boldsymbol{x}_n^g \in \mathbb{R}^G$, we extract the pathway-specific subprofile and densify it:

$$\boldsymbol{z}_{n,c}^{\text{agg}} = R(\boldsymbol{x}_n^g \odot \boldsymbol{a}_c) \in \mathbb{R}^{N_c}, \tag{1}$$

where $\odot$ denotes element-wise multiplication, $R(\cdot)$ removes zeros, and N_c is the number of genes selected by pathway c. A lightweight per-pathway head then maps each summary to a D-dimensional token:

$$\boldsymbol{x}_{n,c}^g = E_g\big(\boldsymbol{z}_{n,c}^{\text{agg}}\big) \in \mathbb{R}^D. \tag{2}$$

This produces genomic tokens $\boldsymbol{X}_n^g = [\boldsymbol{x}_{n,1}^g; \ldots; \boldsymbol{x}_{n,C_g}^g] \in \mathbb{R}^{N_n^g \times D}$ with $N_n^g = C_g$. In our experiments, pathways (e.g., Hallmark [24]) are fixed a priori.

3.2 TOGETHER: UOT-guided Prototype Alignment

The TOGETHER stage aligns histology and genomics using a shared set of prototypes. Instead of directly matching tokens across modalities, we map each token to the same prototype bank, which serves as a common coordinate system. Crucially, assignments are computed with unbalanced optimal transport, allowing uncertain tokens to remain diffuse early on and preventing noisy tokens from dominating alignment.

Shared prototypes. Let $P \in \mathbb{R}^{K \times D'}$ be K learnable prototypes in a shared D'-dimensional space. For modality $m \in \{p, g\}$, tokens are projected by W_m and ℓ_2-normalized to obtain $\tilde{X}_n^m \in \mathbb{R}^{N_n^m \times D'}$. Prototypes are also normalized to $\tilde{P} \in \mathbb{R}^{K \times D'}$. We compute cosine similarities with temperature τ:

$$L_n^m = \frac{\tilde{X}_n^m (\tilde{P})^\top}{\tau} \in \mathbb{R}^{N_n^m \times K}. \tag{3}$$

Unbalanced optimal transport for robust assignments. A naive choice is to use $\mathrm{softmax}(L_n^m)$ as token-to-prototype weights. In survival analysis, this can be unstable: WSI bags are large and long-tailed, and pathway tokens exhibit strong patient-specific sparsity. Moreover, both modalities contain noisy or weakly informative tokens, especially early in training. Motivated by OT-Surv [33], we compute assignments via unbalanced optimal transport (UOT), which naturally supports *semi-relaxed* matching: each token can spread its mass across prototypes rather than committing prematurely. We define a transport cost using negative log-probabilities:

$$C_n^m = -\log \mathrm{softmax}(L_n^m) \in \mathbb{R}^{N_n^m \times K}, \quad m \in \{p, g\}. \tag{4}$$

To allow cross-modal evidence sharing under a single assignment geometry, we concatenate the two modalities along the token axis:

$$C_n = \begin{bmatrix} C_n^p \\ C_n^g \end{bmatrix} \in \mathbb{R}^{N_{\mathrm{tot}} \times K}, \quad N_{\mathrm{tot}} = N_n^p + N_n^g. \tag{5}$$

We seek a transport plan $Q \in \mathbb{R}_{\geq 0}^{N_{\mathrm{tot}} \times K}$ that minimizes a general OT objective:

$$\min_{Q \geq 0} \ \langle Q, C_n \rangle_F \ + \ F_1\big(Q\mathbf{1}, u\big) \ + \ F_2\big(Q^\top \mathbf{1}, v\big), \tag{6}$$

where $\langle \cdot, \cdot \rangle_F$ denotes the Frobenius inner product, $\mathbf{1}$ is an all-one vector with context-dependent dimension, u and v are source and target marginals, and F_1, F_2 encode the corresponding marginal constraints. We fix the source as uniform $u = a = \frac{1}{N_{\mathrm{tot}}} \mathbf{1}_{N_{\mathrm{tot}}}$, ensuring each token contributes equally before assignment. The prototype-side marginal is relaxed to handle heterogeneity.

Prototype-side relaxation. We penalize deviation from a uniform prior $b = \frac{1}{K} \mathbf{1}_K$ via KL divergence:

$$\min_{Q \geq 0} \quad \langle Q, C_n \rangle \ + \ \gamma \, \mathrm{KL}\big(Q^\top \mathbf{1} \,\big\|\, b\big) \quad \text{s.t.} \quad Q\,\mathbf{1} = a, \tag{7}$$

where $\gamma > 0$ controls the relaxation strength. This prevents the plan from collapsing onto a few dominant prototypes and keeps prototype usage balanced, while still allowing patient-dependent deviations. We provide additional algorithm details and theoretical derivations in Appendix A.

Curriculum on transported mass. Even with a balanced prior, noisy tokens can still introduce spurious matches early in training. We therefore control the *total mass* transported to real prototypes using $\rho \in [0, 1]$: only a ρ-fraction of

mass is assigned to the K prototypes, and the remaining $1-\rho$ is absorbed by a sink. Small ρ yields conservative, uncertainty-tolerant assignments; larger ρ gradually enforces more decisive matching. Let $b_\rho = \frac{\rho}{K}\mathbf{1}_K$. We obtain:

$$\min_{Q\geq 0} \quad \langle Q, C_n\rangle \;+\; \gamma\,\mathrm{KL}\big(Q^\top \mathbf{1} \,\big\|\, b_\rho\big) \quad \text{s.t.} \quad Q\,\mathbf{1} \leq a,\; \mathbf{1}^\top Q \mathbf{1} = \rho. \tag{8}$$

We schedule ρ with a smooth ramp-up:

$$\rho(t) = \rho_{\text{base}} + (\rho_{\text{upper}} - \rho_{\text{base}})\,\exp\!\left(-5\Big(1-\tfrac{t}{T}\Big)^2\right), \tag{9}$$

where t is the current training iteration clipped to $[0, T]$, T is the ramp-up horizon, and ρ_{base} and ρ_{upper} denote the initial and maximum transported masses, respectively. Thus, assignments remain cautious during the unstable early phase and become sharper only after representations have matured.

To solve the relaxed problem efficiently, we append a zero-cost sink column and define $\tilde{C}_n = [C_n \mid \mathbf{0}] \in \mathbb{R}^{N_{\text{tot}}\times(K+1)}$. The corresponding target prior becomes:

$$\tilde{b}(\rho) = \begin{bmatrix} \frac{\rho}{K}\mathbf{1}_K \\ 1-\rho \end{bmatrix} \in \mathbb{R}^{K+1}. \tag{10}$$

We then solve:

$$\min_{\tilde{Q}\geq 0} \quad \langle \tilde{Q}, \tilde{C}_n\rangle \;+\; \gamma\,\mathrm{KL}\Big(\tilde{Q}^\top \mathbf{1} \,\Big\|\, \tilde{b}(\rho)\Big) \quad \text{s.t.} \quad \tilde{Q}\,\mathbf{1} = a, \tag{11}$$

After removing the sink column, we multiply the remaining plan by N_{tot} and denote the resulting row-scaled UOT assignment by $Q^\star \in \mathbb{R}^{N_{\text{tot}}\times K}$. Because the plan is computed over concatenated tokens, strong evidence from one modality can support the other, while the sink prevents forced agreement when the signals conflict. Let $Q^{\star,p}$ and $Q^{\star,g}$ denote the modality-specific blocks of $Q^\star$. We blend similarity-based weights and transport-based assignments, then row-normalize the mixture over prototypes:

$$W_n^m = \mathrm{Normalize}\big((1-\beta)\,\mathrm{softmax}(L_n^m) + \beta\, Q^{\star,m}\big), \quad \beta \in [0,1], \tag{12}$$

where $\mathrm{Normalize}(A)_{ik} = A_{ik}/(\sum_j A_{ij} + \epsilon)$ denotes row-wise normalization with a small $\epsilon > 0$ for numerical stability. Prototype tokens are aggregated as:

$$H_n^m = (W_n^m)^\top X_n^m \in \mathbb{R}^{K\times D}, \quad m \in \{p, g\}. \tag{13}$$

Instance-level supervision. We further use UOT assignments as soft pseudo-labels to regularize token-to-prototype predictions. Let $\boldsymbol{\pi}_i^m = \mathrm{Normalize}(Q^{\star,m})_{i,:}$ be the row-normalized UOT pseudo-label and $\boldsymbol{p}_i^m = \mathrm{softmax}(\boldsymbol{\ell}_i^m)$ be the predicted prototype distribution, where $\boldsymbol{\ell}_i^m$ is the i-th row of L_n^m. We define:

$$\mathcal{L}_{\text{CE}}^m = -\frac{1}{N_n^m}\sum_{i=1}^{N_n^m} \langle \boldsymbol{\pi}_i^m, \log \boldsymbol{p}_i^m\rangle, \quad m \in \{p, g\}, \tag{14}$$

and combine the two modalities:

$$\mathcal{L}_{\text{instance}} = \lambda_{\text{wsi}}\,\mathcal{L}_{\text{CE}}^p + \lambda_{\text{gen}}\,\mathcal{L}_{\text{CE}}^g. \tag{15}$$

3.3 Apart: Contrastive Diversification

After Together, both modalities live in the same prototype coordinate system. This is useful for sharing information, but it can also blur modality identity. The Apart stage counteracts this effect by explicitly keeping histology and genomics distinguishable.

Anchor refinement. We introduce a learnable anchor for each modality, a^p and $a^g \in \mathbb{R}^{D'}$, which serve as modality identifiers rather than patient-specific centers. Prototype tokens are projected by a modality-specific map $U_m \in \mathbb{R}^{D\times D'}$ and normalized to obtain $\tilde{H}_n^m \in \mathbb{R}^{K\times D'}$. We append the anchor as an extra token and refine the set with a lightweight self-attention module $\mathcal{R}_\theta$:

$$Y_n^m = \mathcal{R}_\theta([\tilde{H}_n^m; a^m]) \in \mathbb{R}^{(K+1)\times D'}, \quad m \in \{p, g\}, \tag{16}$$

then keep the first K tokens:

$$\hat{H}_n^m = \mathrm{head}_K(Y_n^m) \in \mathbb{R}^{K\times D'}. \tag{17}$$

Anchor-based contrastive regularization. We encourage each modality to stay close to its own anchor while remaining separated from the other modality's anchor. Let $\phi(\cdot)$ be a projection head followed by unit normalization, and let $\bar{h}_n^m = \frac{1}{K}\sum_{k=1}^K \hat{h}_{n,k}^m$ be the mean refined token. We compute:

$$s_+^m = \frac{\langle \phi(\bar{h}_n^m), \phi(a^m)\rangle}{\tau_r}, \quad s_-^m = \frac{\langle \phi(\bar{h}_n^m), \phi(a^{\bar{m}})\rangle}{\tau_r}, \quad m \in \{p, g\}, \tag{18}$$

where $\bar{m}$ denotes the other modality and $\tau_r > 0$ is the contrastive temperature. We then define the InfoNCE-style loss [3, 30]:

$$\mathcal{L}_{\text{contrast}} = -\log \frac{\exp(s_+^p)}{\exp(s_+^p) + \exp(s_-^p)} - \log \frac{\exp(s_+^g)}{\exp(s_+^g) + \exp(s_-^g)}. \tag{19}$$

3.4 Prediction Head

We next combine the two refined modality streams for survival prediction. We use a transformer-style co-attention module [38] over prototype tokens $\hat{H}_n^p$ and $\hat{H}_n^g$. Co-attention allows pathology to query genomics and vice versa, capturing cross-modal dependencies at the prototype level. We then mean-pool within each modality to obtain h_n^p and h_n^g, and map them to a joint representation $f_n = F(h_n^p, h_n^g) \in \mathbb{R}^{d_f}$. A linear risk head predicts log-risk $r_n = \langle w, f_n\rangle$, with $w \in \mathbb{R}^{d_f}$. We train with a standard survival objective $\mathcal{L}_{\text{surv}}$ (e.g., Cox partial likelihood [12]). The full training loss is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{surv}} + \lambda_{\text{contrast}}\mathcal{L}_{\text{contrast}} + \lambda_{\text{inst}}\mathcal{L}_{\text{instance}}. \tag{20}$$

Discussion. The two auxiliary terms play different roles. $\mathcal{L}_{\text{instance}}$ stabilizes prototype assignments in Together, while $\mathcal{L}_{\text{contrast}}$ preserves modality identity in Apart. In practice, we find this separation improves both robustness and interpretability without complicating optimization (see ablation studies in Section 4.3 and Appendix A.1).

4 Experiments

In this section, we first describe the implementation details (Section 4.1). We then conduct comprehensive comparisons with seventeen state-of-the-art methods across five benchmarks (Section 4.2). We further perform ablation studies to analyze the contributions of the main components of TTA (Section 4.3).

4.1 Implementation Details

Datasets. We evaluate our method on five cancer cohorts from The Cancer Genome Atlas (TCGA): Bladder urothelial carcinoma (BLCA, $n = 359$), Breast invasive carcinoma (BRCA, $n = 868$), Stomach adenocarcinoma (STAD, $n = 318$), Colon and Rectum adenocarcinoma (CRC, $n = 296$), and Kidney renal clear cell carcinoma (KIRC, $n = 340$). These cohorts span distinct tissue systems, including gastrointestinal (CRC, STAD), urological (BLCA, KIRC), and breast (BRCA) cancers, and exhibit substantially different histology, genomic profiles, and survival patterns. This diversity provides a challenging testbed for evaluating cross-tissue generalization.

We follow the data splits in [37] and use disease-specific survival (DSS) [26] as the prediction target. Performance is evaluated using the concordance index (C-Index) with 5-fold site-stratified cross-validation [18], which is widely adopted for survival analysis on TCGA cohorts with limited sample sizes [9,21,37,41]. To avoid overfitting, a single configuration is used for all folds and datasets without per-fold tuning, and we report mean±std C-Index.

Settings. Key configurations are summarized here. Additional implementation details are provided in Appendix B, and further experiments are reported in Appendix C.

WSI patches are encoded using UNI [5], a DINOv2-based ViT-Large pretrained on 1×10^8 pathology patches. To verify robustness across feature extractors, we also evaluate ResNet50 [13] (Table 9 in Appendix C). For bags exceeding a fixed size, we apply uniform subsampling. Shorter bags are zero-padded. The UOT transport plan then learns importance weights during aggregation. Following MMP [37], gene expression profiles are organized into 50 Hallmark pathways covering 4,241 genes, which correspond to well-defined biological processes and provide interpretable genomic groupings [24]. In the TOGETHER stage, we use K=32 shared prototypes with KL regularization weight γ=0.1. The curriculum mass is scheduled from ρ_{base}=0.1 to ρ_{upper}=1.0. The auxiliary loss weights are $\lambda_{\text{contrast}}$=$\lambda_{\text{inst}}$=0.5, and the modality weights are λ_{wsi}=λ_{gen}=1. For the survival objective $\mathcal{L}_{\text{surv}}$, we use the Cox partial likelihood [12]. Both Cox and discrete-time NLL are standard objectives in survival modeling, and we verify their interchangeability in Appendix C. Note that Cox partial likelihood requires batch sizes larger than 1, as risk set computation involves pairwise comparisons between samples.

Table 1: Comparison with state-of-the-art methods in terms of C-index (mean ± std) across five TCGA cancer cohorts. "Modal." denotes Modality, "h." and "g." denote models that rely solely on histopathology WSIs and genomic data, respectively, while "g.+h." indicates multimodal models using both modalities. The overall score is computed as the average C-index across datasets. The best and second-best results are highlighted in **bold** and underline, respectively.

Model	Modal.	BRCA	BLCA	STAD	CRC	KIRC	Overall
SNN [22]	g.	0.619±0.063	0.618±0.027	0.557±0.072	0.576±0.102	0.689±0.098	0.611
SNNTrans [34]	g.	0.630±0.062	0.622±0.038	0.559±0.070	0.570±0.098	0.698±0.126	0.616
ABMIL [19]	h.	0.566±0.097	0.553±0.052	0.566±0.033	0.655±0.118	0.675±0.124	0.603
CLAM [27]	h.	0.627±0.188	0.618±0.104	0.538±0.077	0.629±0.132	0.670±0.124	0.616
HIPT [4]	h.	0.566±0.092	0.606±0.126	0.529±0.074	0.600±0.066	0.685±0.099	0.597
AttnMISL [44]	h.	0.585±0.073	0.533±0.065	0.541±0.052	**0.723±0.111**	0.656±0.112	0.607
TransMIL [34]	h.	0.599±0.056	0.595±0.102	0.500±0.060	0.550±0.143	0.676±0.132	0.584
OTSurv [33]	h.	0.625±0.071	0.637±0.065	0.556±0.057	0.663±0.102	0.739±0.149	0.644
PANTHER [36]	h.	0.643±0.128	0.593±0.083	0.503±0.082	0.639±0.133	0.700±0.124	0.615
MCAT [9]	g.+ h.	0.650±0.096	0.623±0.055	0.528±0.112	0.579±0.134	0.699±0.121	0.615
CMTA [1]	g.+ h.	0.687±0.077	0.622±0.056	0.539±0.076	0.568±0.102	0.711±0.121	0.625
MOTCat [41]	g.+ h.	0.717±0.058	0.631±0.059	0.576±0.075	0.598±0.128	0.708±0.109	0.646
SurvPath [21]	g.+ h.	0.696±0.061	0.619±0.051	0.555±0.133	0.588±0.134	0.742±0.101	0.640
PIBD [49]	g.+ h.	0.683±0.056	0.661±0.034	0.552±0.054	0.582±0.077	0.732±0.139	0.642
MRePath [31]	g.+ h.	0.703±0.036	0.651±0.060	0.579±0.062	0.639±0.101	0.743±0.112	0.663
MMP [37]	g.+ h.	0.724±0.067	0.640±0.050	0.594±0.066	0.634±0.118	0.743±0.133	0.667
LD-CVAE [50]	g.+ h.	0.709±0.047	**0.676±0.035**	0.589±0.083	0.602±0.120	0.751±0.139	0.665
TTA(Ours)	g.+ h.	**0.726±0.039**	0.662±0.079	**0.613±0.079**	0.685±0.131	**0.778±0.117**	**0.693**

4.2 Main Results

Comparisons with SOTAs. We compare TTA with a wide range of recent survival analysis methods. *Unimodal baselines*, including SNN [22] and SNNTrans [22, 34] for genomic survival prediction. For histopathology-only survival modeling, we compare against seven representative MIL methods: ABMIL [19], CLAM [27], HIPT [4], AttnMISL [44], TransMIL [34], OTSurv [33], and PANTHER [36]. For multimodal survival prediction, we compare against eight recent methods including MCAT [9], CMTA [1], MOTCat [41], SurvPath [21], PIBD [49], MMP [37], MRePath [31], and LD-CVAE [50]. All comparison methods were re-run with the same splits, gene tokens, and WSI features.

The results are reported in Table 1. As is shown, compared with unimodal methods, most multimodal approaches including ours achieve higher overall C-index, indicating complementary benefits from fusing WSIs and genomics. Among multimodal methods, TTA attains the best overall performance, outperforming the second-best by **+2.6%** in Overall C-index. Across cohorts, TTA ranks first on BRCA, STAD, KIRC and second on BLCA and CRC. Relative to strong recent multimodal methods (e.g., MMP [37], LD-CVAE [50]), per-dataset gains typically range from about **+0.2% to +5.1%**, and up to **+8.3%** on CRC. This supports that our Together-Then-Apart design, which *minimizes semantic discrepancies* during alignment and *maximizes modality distinctiveness* thereafter, yields more discriminative multimodal representations for survival prediction.

Table 2: Ablation study on the TOGETHER and APART stages. C-index (mean ± std) is reported across five cohorts, with the last column showing the average across cohorts. Gray rows denote the default setting.

TOGETHER	APART	BRCA	BLCA	STAD	CRC	KIRC	Overall
		0.682±0.082	0.660±0.063	0.558±0.071	0.561±0.170	0.756±0.124	0.643
✓		0.675±0.078	0.682±0.067	0.582±0.043	0.587±0.177	0.784±0.088	0.662
	✓	0.683±0.034	0.651±0.077	0.585±0.099	0.625±0.145	0.785±0.098	0.666
✓	✓	**0.726**±0.039	**0.662**±0.079	**0.613**±0.079	**0.685**±0.131	**0.778**±0.117	**0.693**

Table 3: Ablation study on the TOGETHER stage: shared vs. modality-specific prototypes and joint vs. modality-specific unbalanced optimal transport. C-index (mean ± std) is reported across five cohorts, with the last column showing the average across cohorts. Gray rows denote the default setting.

Shared Prototypes	joint UOT	BRCA	BLCA	STAD	CRC	KIRC	Overall
		0.755±0.025	0.671±0.025	0.594±0.063	0.609±0.199	0.780±0.108	0.681
✓		0.713±0.034	0.662±0.081	0.605±0.081	0.672±0.110	0.768±0.130	0.684
✓	✓	**0.726**±0.039	**0.662**±0.079	**0.613**±0.079	**0.685**±0.131	**0.778**±0.117	**0.693**

Analysis on the CRC Cohort. CRC is particularly challenging for multimodal survival modeling. Clinical prognosis in colorectal cancer is strongly driven by histological morphology, meaning that most predictive signals originate from WSIs. In this setting, strong cross-modal alignment can dilute discriminative morphological cues, causing several multimodal methods to underperform WSI-only baselines. TTA alleviates this issue through the APART stage, which explicitly preserves WSI-specific evidence via anchor-guided contrastive regularization. The ablation in Table 2 confirms this behavior: removing APART leads to a $-\mathbf{9.8\%}$ drop on CRC ($0.685 \rightarrow 0.587$). This result highlights the importance of preserving modality-specific representations in settings where predictive signals are unevenly distributed across modalities.

4.3 Ablation Study

We conduct ablation experiments to analyze the contributions of the main components in TTA. Hyperparameter sensitivity is summarized in Fig. 3. We first evaluate the individual and combined effects of the TOGETHER and APART stages. We then analyze key design choices within TOGETHER, including *shared vs. respective prototypes*, *joint vs. respective UOT*, and the *instance-to-prototype assignment strategy* with *different mass schedules*. Finally, we study the two key components of APART: *anchor refinement* and *contrastive regularization*.

Effect of the TOGETHER and APART Stages. We evaluate four variants: (i) neither stage, (ii) TOGETHER only, (iii) APART only, and (iv) both stages. Removing TOGETHER replaces shared prototypes with modality-specific prototypes and disables UOT alignment. Removing APART disables the anchor refiner and its contrastive objective. As shown in Table 2, enabling only TOGETHER im-

Table 4: Ablation study on anchor refinement and contrastive regularization in the APART stage. "Contrast. Regular." denotes Contrastive Regularization. C-index (mean ± std) is reported across five cohorts, with the last column showing the average across cohorts. Gray rows denote the default setting.

Anchor Refinement	Contrast. Regular.	BRCA	BLCA	STAD	CRC	KIRC	Overall
		0.675±0.078	0.682±0.067	0.582±0.043	0.587±0.177	0.784±0.088	0.662
✓		0.694±0.048	0.648±0.066	0.598±0.099	0.636±0.155	0.771±0.122	0.669
	✓	0.688±0.072	0.678±0.053	0.589±0.032	0.627±0.134	0.778±0.083	0.672
✓	✓	**0.726**±0.039	**0.662**±0.079	**0.613**±0.079	**0.685**±0.131	**0.778**±0.117	**0.693**

Table 5: Ablation on assignment and curriculum mass ρ. C-index (mean ± std) is reported across five cohorts, with the last column showing the average across cohorts. "curr." denotes curriculum. Gray rows denote the default setting.

Module	Setting	BRCA	BLCA	STAD	CRC	KIRC	Avg
Assignment	KMeans	0.726±0.057	0.656±0.070	0.582±0.078	0.670±0.096	0.772±0.119	0.681
	General OT	0.716±0.036	0.652±0.085	0.585±0.082	0.692±0.112	0.769±0.124	0.683
	w/o curr. ρ	0.720±0.039	0.662±0.083	0.589±0.081	0.686±0.117	0.775±0.126	0.686
	w/ curr. ρ	**0.726**±0.039	**0.662**±0.079	**0.613**±0.079	**0.685**±0.131	**0.778**±0.117	**0.693**
ρ ramp-up Schedule	**Sigmoid**	**0.726**±0.039	**0.662**±0.079	**0.613**±0.079	**0.685**±0.131	**0.778**±0.117	**0.693**
	Linear	0.735±0.072	0.661±0.060	0.604±0.053	0.667±0.116	0.771±0.113	0.687

proves Overall C-index by +**1.9%**, while enabling only APART yields +**2.3%**. Enabling both stages further increases performance to +**5.0%**. These results indicate that cross-modal alignment and modality-specific representation learning are complementary and jointly necessary.

Shared vs. Respective Prototypes and Joint vs. Respective UOT. We further analyze two design choices within the TOGETHER stage. When shared prototypes are disabled, each modality maintains its own prototype set. When joint UOT is disabled, optimal transport is solved independently for each modality. With both components enabled, prototype tokens from both modalities are concatenated and matched through a joint UOT plan with a dummy sink and curriculum mass. Results in Table 3 show that replacing modality-specific prototypes with shared prototypes improves Overall C-index by +**0.3%**. Further enabling joint UOT improves performance by an additional +**0.9%**, resulting in a total gain of +**1.2%**. This indicates that a shared prototype space combined with a unified transport plan yields more coherent cross-modal alignment.

Anchor Refinement and Contrastive Regularization. We next examine the two components in APART stage. Table 4 shows that anchor refinement alone improves performance by +**0.7%**, while contrastive regularization alone yields +**1.0%**. Combining both components leads to a larger improvement of +**3.1%**. These results suggest that lightweight refinement together with contrastive coupling helps preserve modality-specific structure and prevents over-alignment.

Instance-to-Prototype Assignment. We ablate the soft instance-to-prototype assignment used in TOGETHER (Table 5). Besides hard KMeans, we compare

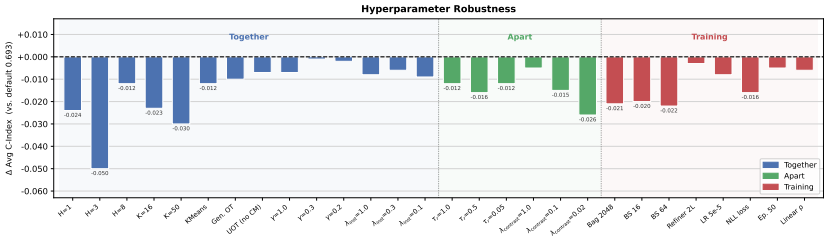

Fig. 3: Hyperparameter robustness. Each bar shows the change in average C-index across five cohorts when perturbing a single factor from the default configuration. The default Avg. (0.693) is indicated by the dashed line. Colors indicate groups of hyperparameters for the TOGETHER stage, the APART stage, and training.

OT-based soft assignments (General OT and UOT): "w/o curr. ρ" denotes without curriculum mass ρ and uses a fixed transported mass, while "w/ curr. ρ" ramps up ρ to gradually tighten matching and route uncertain tokens to a sink early on. UOT with curriculum mass ρ achieves the best Overall C-index, indicating that our model benefits from a heterogeneity-aware soft assignment with curriculum scheduling.

Curriculum Mass Schedule. We further examine the scheduling strategy of transported mass $\rho(t)$. Replacing the sigmoid ramp-up with a linear schedule reduces the Avg C-index from 0.693 to 0.687 (Table 5). This observation suggests that gradually enforcing strict matching stabilizes training during early stages.

Robustness and Hyperparameter Sensitivity. We adopt a *single configuration* across all datasets and folds without per-dataset tuning. Fig. 3 shows one-at-a-time perturbations around the default configuration. Detailed results are provided in Appendix C. Most variants lead to only minor performance changes, indicating stable performance across a range of settings. When enabling the multi-head OT assignment, too few heads can be unstable (e.g., H=3). Meanwhile, even a single head (H=1) achieves a reasonable Avg C-index.

More Ablations and Visualizations. Additional ablation experiments and hyperparameter studies are provided in Appendix C. We also provide additional visualization results in Appendix D.

4.4 Stratification Visualization

To further evaluate risk stratification, we perform Kaplan–Meier survival analysis based on predicted risk groups. Fig. 4 presents Kaplan–Meier curves for all five cancer cohorts, showing clear separation between predicted high- and low-risk groups. The log-rank test [2] yields statistically significant p-values across all cohorts: 2.61×10^{-3} (BRCA), 2.51×10^{-4} (BLCA), 8.93×10^{-3} (STAD), 3.02×10^{-2} (CRC), and 1.86×10^{-10} (KIRC). All p-values are well below the 0.05 significance threshold, confirming effective survival stratification.

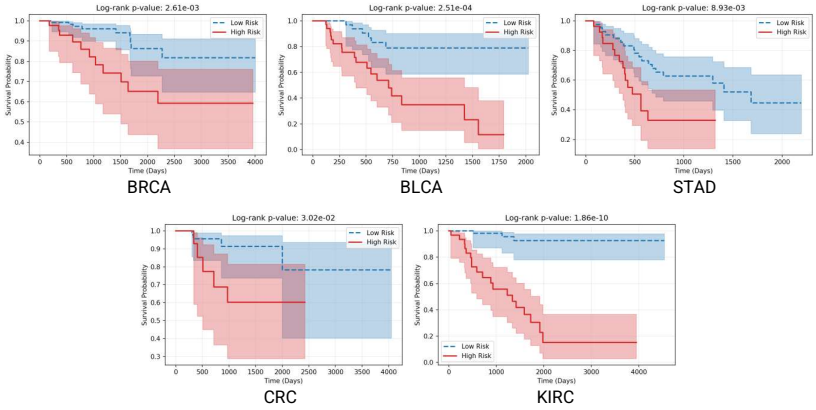

Fig. 4: Kaplan-Meier curves for the predicted high-risk (red) and low-risk (blue) groups. A p-value < 0.05 indicates statistical significance, and shaded regions denote confidence intervals.

5 Conclusion

We introduced Together-Then-Apart (TTA), a framework for multimodal survival analysis that balances cross-modal alignment (Together) with modality-specific distinctiveness (Apart). In the Together stage, TTA aligns modalities through shared prototypes using an unbalanced optimal transport formulation with a curriculum mass schedule, producing stable assignments that account for intra-instance heterogeneity. The Apart stage then preserves modality-specific structure through anchor refinement and contrastive regularization, preventing representational collapse while retaining distinctive signals from each modality. Experiments on five TCGA cohorts show that TTA consistently improves survival prediction over recent multimodal approaches and reveals interpretable cross-modal structures associated with clinical outcomes. More broadly, our results suggest that effective multimodal models should balance shared semantic alignment with modality-specific information, rather than forcing heterogeneous signals into a single representation. We hope this work encourages further study of representation learning strategies that respect both shared and modality-specific structure in multimodal biomedical data.

References

1. et al, Z.: Cross-modal translation and alignment for survival analysis. In: ICCV. pp. 21485–21494 (2023)
2. Bland, J.M., Altman, D.G.: The logrank test. Bmj **328**(7447), 1073 (2004)
3. Chen, L., Gan, Z., Cheng, Y., Li, L., Carin, L., Liu, J.: Graph optimal transport for cross-domain alignment. In: International Conference on Machine Learning. pp. 1542–1553. PMLR (2020)

4. Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16144–16155 (2022)
5. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F.K., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., Williams, M., Oldenburg, L., Weishaupt, L.L., Wang, J.J., Vaidya, A., Le, L.P., Gerber, G., Sahai, S., Williams, W., Mahmood, F.: Towards a general-purpose foundation model for computational pathology. Nature Medicine (2024)
6. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Chen, B., Zhang, A., Shao, D., Song, A.H., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine (2024)
7. Chen, R.J., Lu, M.Y., Shaban, M., Chen, C., Chen, T.Y., Williamson, D.F., Mahmood, F.: Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 339–349. Springer (2021)
8. Chen, R.J., Lu, M.Y., Wang, J., Williamson, D.F., Rodig, S.J., Lindeman, N.I., Mahmood, F.: 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)
9. Chen, R.J., Lu, M.Y., Weng, W.H., Chen, T.Y., Williamson, D.F., Manz, T., Shady, M., Mahmood, F.: Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4025 (2021)
10. Chen, R.J., Lu, M.Y., Williamson, D.F., Chen, T.Y., Lipkova, J., Noor, Z., Shaban, M., Shady, M., Williams, M., Joo, B., et al.: Pan-cancer integrative histology-genomic analysis via multimodal deep learning. Cancer Cell **40**(8), 865–878 (2022)
11. Claudio Quiros, A., Coudray, N., Yeaton, A., Yang, X., Liu, B., Le, H., Chiriboga, L., Karimkhan, A., Narula, N., Moore, D.A., et al.: Mapping the landscape of histomorphological cancer phenotypes using self-supervised learning on unannotated pathology slides. Nature Communications **15**(1), 4596 (2024)
12. Cox, D.R.: Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological) **34**(2), 187–202 (1972)
13. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. IEEE (2009)
14. Ding, K., Zhou, M., Metaxas, D.N., Zhang, S.: Pathology-and-genomics multimodal transformer for survival outcome prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 622–631. Springer (2023)
15. Elmarakeby, H.A., Hwang, J., Arafeh, R., Crowdis, J., Gang, S., Liu, D., AlDubayan, S.H., Salari, K., Kregel, S., Richter, C., et al.: Biologically informed deep neural network for prostate cancer discovery. Nature **598**(7880), 348–352 (2021)
16. Guo, L., Wang, Y., Wen, T., Wang, Y., Feng, A., Chen, B., Jegelka, S., You, C.: Csrv2: Unlocking ultra-sparse embeddings. arXiv preprint arXiv:2602.05735 (2026)
17. Haykin, S.: Neural networks: a comprehensive foundation. Prentice Hall PTR (1998)
18. Howard, F.M., Dolezal, J., Kochanny, S., Schulte, J., Chen, H., Heij, L., Huo, D., Nanda, R., Olopade, O.I., Kather, J.N., et al.: The impact of site-specific

digital histology signatures on deep learning model accuracy and bias. Nature communications **12**(1), 4423 (2021)
19. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
20. Jackson, H.W., Fischer, J.R., Zanotelli, V.R., Ali, H.R., Mechera, R., Soysal, S.D., Moch, H., Muenst, S., Varga, Z., Weber, W.P., et al.: The single-cell pathology landscape of breast cancer. Nature **578**(7796), 615–620 (2020)
21. Jaume, G., Vaidya, A., Chen, R., Williamson, D., Liang, P., Mahmood, F.: Modeling dense multimodal interactions between biological pathways and histology for survival prediction. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
22. Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S.: Self-normalizing neural networks. Advances in neural information processing systems **30** (2017)
23. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14318–14328 (2021)
24. Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., Tamayo, P.: The molecular signatures database hallmark gene set collection. Cell systems **1**(6), 417–425 (2015)
25. Lin, T., Yu, Z., Hu, H., Xu, Y., Chen, C.W.: Interventional bag multi-instance learning on whole-slide pathological images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19830–19839 (2023)
26. Liu, J., Lichtenberg, T., Hoadley, K.A., Poisson, L.M., Lazar, A.J., Cherniack, A.D., Kovatich, A.J., Benz, C.C., Levine, D.A., Lee, A.V., et al.: An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell **173**(2), 400–416 (2018)
27. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021)
28. Mobadersany, P., Yousefi, S., Amgad, M., Gutman, D.A., Barnholtz-Sloan, J.S., Velázquez Vega, J.E., Brat, D.J., Cooper, L.A.: Predicting cancer outcomes from histology and genomics using convolutional networks. Proceedings of the National Academy of Sciences **115**(13), E2970–E2979 (2018)
29. Nunes, L., Li, F., Wu, M., et al.: Prognostic genome and transcriptome signatures in colorectal cancers. Nature **633**(8028), 137–146 (2024)
30. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
31. Qu, M., Yang, G., Di, D., Su, T., Gao, Y., Song, Y., Fan, L.: Multimodal cancer survival analysis via hypergraph learning with cross-modality rebalance. arXiv preprint arXiv:2505.11997 (2025)
32. Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., Wadi, L., Meyer, M., Wong, J., Xu, C., et al.: Pathway enrichment analysis and visualization of omics data using g: Profiler, GSEA, Cytoscape and EnrichmentMap. Nature protocols **14**(2), 482–517 (2019)
33. Ren, Q., Wang, Y., Fang, R., Ling, H., You, C.: Otsurv: A novel multiple instance learning framework for survival prediction with heterogeneity-aware optimal transport. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 439–449. Springer (2025)

34. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems **34**, 2136–2147 (2021)
35. Shao, Z., Chen, Y., Bian, H., Zhang, J., Liu, G., Zhang, Y.: Hvtsurv: Hierarchical vision transformer for patient-level survival prediction from whole slide image. In: Proceedings of the AAAI conference on artificial intelligence (2023)
36. Song, A.H., Chen, R.J., Ding, T., Williamson, D.F., Jaume, G., Mahmood, F.: Morphological prototyping for unsupervised slide representation learning in computational pathology. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
37. Song, A.H., Chen, R.J., Jaume, G., Vaidya, A.J., Baras, A., Mahmood, F.: Multimodal prototyping for cancer survival prediction. In: Forty-first International Conference on Machine Learning (2024)
38. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
39. Vu, Q.D., Rajpoot, K., Raza, S.E.A., Rajpoot, N.: Handcrafted Histological Transformer (H2T): Unsupervised representation of whole slide images. Medical Image Analysis **85**, 102743 (2023)
40. Wen, T., Wang, Y., Zeng, Z., Peng, Z., Su, Y., Liu, X., Chen, B., Liu, H., Jegelka, S., You, C.: Beyond matryoshka: Revisiting sparse coding for adaptive representation. arXiv preprint arXiv:2503.01776 (2025)
41. Xu, Y., Chen, H.: Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21241–21251 (2023)
42. Xu, Y., Zhou, F., Zhao, C., Wang, Y., Yang, C., Chen, H.: Distilled prompt learning for incomplete multimodal survival prediction. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5102–5111 (2025)
43. Yang, S., Wang, Y., Chen, H.: Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology. In: International conference on medical image computing and computer-assisted intervention. pp. 296–306. Springer (2024)
44. Yao, J., Zhu, X., Jonnagaddala, J., Hawkins, N., Huang, J.: Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. Medical Image Analysis **65**, 101789 (2020)
45. You, C., Dai, H., Min, Y., Sekhon, J.S., Joshi, S., Duncan, J.S.: Uncovering memorization effect in the presence of spurious correlations. Nature Communications (2025)
46. You, C., Mint, Y., Dai, W., Sekhon, J.S., Staib, L., Duncan, J.S.: Calibrating multi-modal representations: A pursuit of group robustness without annotations. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2024)
47. Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., Zheng, Y.: Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18802–18812 (2022)
48. Zhang, W., Ren, Q., Liu, W., Ling, H., You, C.: Supervise less, see more: Training-free nuclear instance segmentation with prototype-guided prompting. arXiv preprint arXiv:2511.19953 (2025)

49. Zhang, Y., Xu, Y., Chen, J., Xie, F., Chen, H.: Prototypical information bottlenecking and disentangling for multimodal cancer survival prediction. In: The Twelfth International Conference on Learning Representations (2024)
50. Zhou, J., Tang, J., Zuo, Y., Wan, P., Zhang, D., Shao, W.: Robust multimodal survival prediction with the latent differentiation conditional variational autoencoder. arXiv preprint arXiv:2503.09496 (2025)