Cancer Survival Prediction by Cyclic Generation and Multi-grained Alignment

Yongqi Bu[1,2], **Qinggang Niu**[1,2], **Zhen Li**[3], **Yanyu Xu**[2], **Jun Wang**[2], **Guoxian Yu**[1,2,*]

[1]School of Software, Shandong University, Jinan 250101, Shandong, China
[2]Joint SDU-NTU Centre for Artificial Intelligence Research, Shandong University, Jinan 250101, Shandong, China
[3]Qilu Hospital of Shandong University, Jinan 250012, Shandong, China
{byq, qgniu}@mail.sdu.edu.cn, {qilulizhen, xu_yanyu, kingjun, gxyu}@sdu.edu.cn

Abstract

Cancer survival analysis with multimodal data is crucial for precise treatments and patient benefits. However, the following challenges prohibit integrating histopathology and genomics: (i) multimodal data is not always complete, especially for the more costly genomics data; (ii) intricate interactions between different modalities are difficult to capture and understand. To response, we propose an end-to-end framework (CIMA) that coordinates **C**yc**l**i**c** modality generation and **M**ulti-grained multimodal **A**lignment. Specifically, CIMA designs a cyclic modality reconstruction module to reciprocally impute missing modalities and infer the interactions between them. Next, it introduces the multi-grained alignment module over the imputed data and interactions to mine fine-grained alignments between histopathology (slide patches) and genomics (biological pathways). CIMA then constructs the adaptive fusion module to leverage multimodal data and alignments for survival prediction. Extensive experiments on cancer benchmark datasets demonstrate that CIMA outperforms existing methods and exhibits good interpretability, providing valuable insights into intricate relationships between pathological phenotypes and biological pathways. Our code is released in the supplementary materials.

Extended version —
https://www.sdu-idea.cn/codes.php?name=CIMA

Introduction

Cancer survival analysis is a critical issue in medical time-to-event prediction and has wide applications in clinical management decisions, such as treatment and monitoring (Bray et al. 2024). Compared to relying solely on histopathology images, integrating morphological information from histopathology with molecular mechanism information from genomic profiles enables a more comprehensive and accurate estimation of patient mortality risk for prognosis. However, this process is labor-intensive, time-consuming, and dependent on the subjective factors of clinicians (Van der Laak, Litjens, and Ciompi 2021; Xu et al. 2023). Given that, it is imperative to develop accurate and robust computational methods for cancer survival analysis.

Recent works employ multimodal learning for integration of different perspectives that characterize patients, providing more detailed evidence-based support for survival prediction (Boehm et al. 2022). For example, several studies (Chen et al. 2021; Zhou et al. 2023) have improved the prognosis of a majority of cancers. However, the inevitable **incompleteness** of multimodal data and the **insufficient exploration of relationships** between modalities hinder the further advancement of these methods and undermine their potential for clinical application.

To mitigate the negative impact posed by data incompleteness, various strategies have been explored (Miao et al. 2022; Wang et al. 2022; Wu et al. 2025), such as statistical imputation methods (e.g., zero and mean padding) and autoencoder-based ones (van Loon et al. 2024). Nevertheless, directly applying these approaches to multimodal cancer survival prediction remains problematic for several reasons (Hou et al. 2023), such as the significant heterogeneity between multimodal medical data and the complex interactions between modalities (Qiu et al. 2023; Zhou et al. 2024; Bu et al. 2024). On the one hand, the information content across medical modalities differs greatly, which can lead to severe mode collapse during imputation. For instance, WSIs often contain gigapixel-scale resolution, whereas genomic profiles consist of ten thousand-dimensional sequences, underscoring the necessity of effectively aligning these imbalanced modalities. On the other hand, there are complex interactions between modalities. Inadequate consideration and modeling of these relationships during imputation can introduce bias and noise, increasing the prediction errors. In light of these challenges, there is *a pressing need to develop methods for robust multimodal survival prediction in scenarios where data is frequently incomplete*, particularly in realistic clinical settings where genomic information is frequently missing. This is the first motivation for the present work.

Attention mechanisms have recently emerged as promising solutions for modeling various intricate interactions between modalities (Xu and Chen 2023; Ding et al. 2024; Zhou et al. 2023; Qiu et al. 2023). While these attention mechanisms can simulate dense interactions between modalities, few efforts (Jaume et al. 2024; Chen et al. 2024) concentrate on fine-grained cross-modal relationships. Specifically, morphological manifestations in regions of WSIs should correspond to certain abnormalities in cellular functions, which

*Corresponding author

have the potential to provide critical insights for a deeper understanding of the biological mechanisms of cancer. For instance, the Wnt signal pathway shapes the histologic variation in diffuse gastric cancer (Togasaki et al. 2021). Additionally, the aforementioned methods tend to update attention scores based solely on the losses derived from downstream tasks, lacking specialized guidance that could encourage fine-grained alignment between modalities. To sum up, the *explicit modeling of interactions between patches in WSIs and cellular functions to make fine-grained alignment* constitutes our second motivation.

With the above analysis, we propose an end-to-end framework named CIMA, which coordinates **C**yc**li**c modality generation and **M**ulti-grained multimodal **A**lignment. CIMA employs a flexible strategy to impute incomplete multimodal data and explicitly model the multi-grained interactions between histopathology and genomics. Specifically, CIMA first introduces the Cyclic Modality Reconstruction (CMR) module to handle incomplete multimodal data. CMR leverages the widely available histopathology to recover the genomics that is often absent in clinical practice, and subsequently reconstructs the corresponding histopathology from the imputed genomics. This reciprocal reconstruction mechanism is capable of preventing the introduction of excessive errors during the generation process and enhancing the exploration of cross-modality interactions. Concurrently, CIMA utilizes samples with complete data to establish modality priors, mitigating the supervisory deficit caused by modality missingness. Next, CIMA constructs the Multi-grained Multimodal Alignment (MMA) module over the imputed data. At the coarse-grained level, it aligns the semantic information of different modalities in a shared feature space, where these semantics should both describe the physical condition for the corresponding patient. At the fine-grained level, it captures interactions between WSI (Whole Slide Image) patches and biological pathways. Moreover, CIMA introduces specially designed contrastive supervision to explicitly encourage alignment at both granularities. Finally, guided by the fine-grained alignments, CIMA develops the Adaptive Multimodal Fusion (AMF) module to integrate the meticulously imputed and aligned multimodal data for survival prediction.

The main contributions of our work can be summarized as follows: (a) A unified framework CIMA for cancer patient survival prediction that coordinates the cyclic multimodal reconstruction and multi-grained alignment. (b) A flexible and intuitive missing modality cyclic reconstruction paradigm that restores the missing modalities while avoiding accumulated errors. (c) A skillful multimodal alignment paradigm that explicitly encourages multi-granularity interactions between modalities, providing reliable guidance for multimodal integration and enhancing model interpretability. (d) Extensive experiments on five benchmark datasets prove the superiority and rationality of CIMA.

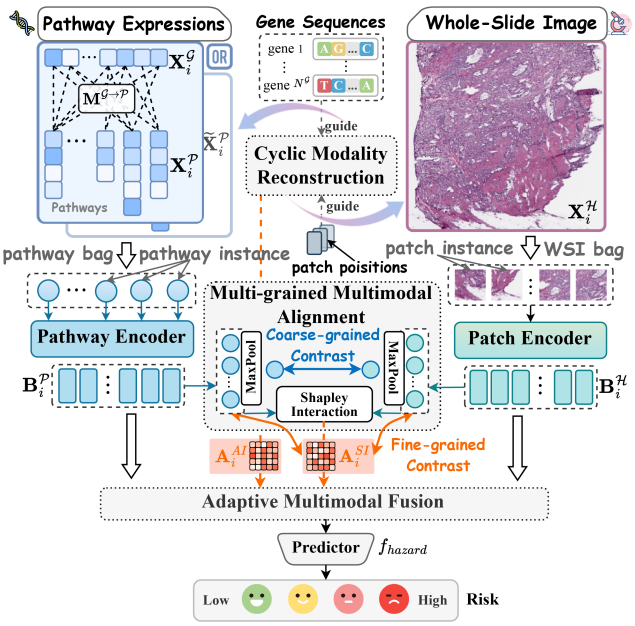

Figure 1: **Conceptual framework of** CIMA. The histopathology and genomics data are organized into bags ($\mathbf{X}_i^{\mathcal{H}}$ and $\mathbf{X}_i^{\mathcal{P}}$). Cyclic Modality Reconstruction (CMR) module leverages available histopathology data to generate pathway expressions, and reversely uses the generated expressions to reconstruct the corresponding pathological features. Next, Multi-grained Multimodal Alignment (MMA) module captures coarse-grained interactions between modalities (histopathology and genomics) and fine-grained interactions between instances (patches and pathways) within the imputed multimodal bags, resulting in a fine-grained interaction map. Finally, Adaptive Multimodal Fusion (AMF) module integrates the multimodal bags under the guidance of the learned interaction maps to predict survival risk scores.

Related Works

Survival Prediction from Unimodal. The Cox proportional hazard regression model (David et al. 1972) and its various extensions have long served as the foundation for cancer survival prediction. The continuous development of digital pathology and high-throughput sequencing technologies has respectively sparked a surge of interest in using WSIs and genomic profiles for unimodal survival prediction (Wiegrebe et al. 2024). These unimodal methods still lack the capacity to dissect underlying pathology and inherently face limitations in robustness and generalization.

Survival Prediction from Multimodal. A recent surge of solutions integrates histopathological and genomic data to improve cancer prognosis analysis (Unger and Kather 2024). The majority of these studies can be categorized into strategies based on tensor operation (Chen et al. 2022; Dwivedi et al. 2022), attention mechanism (Xu and Chen 2023; Zhou et al. 2023), and bilinear pooling (Zhang et al. 2020). However, these methods typically build on the prerequisite of complete multimodal (i.e., each patient possesses data from all modalities simultaneously, and these modalities are one-to-one aligned), which is challenging to fulfill in clinical applications. Several recent efforts can handle incomplete multimodal data (Zhou et al. 2024), but treat missing modality

reconstruction and multimodal alignment as isolated tasks, causing the propagation and accumulation of errors across tasks.

Multimodal Alignment. Alignment is a particularly challenging task in multimodal learning, which can be defined as identifying *coarse-grained* alignments between samples across two or more modalities, or establishing *fine-grained* alignments between their sub-components (Baltrušaitis, Ahuja, and Morency 2018). The integration of histopathology and genomics naturally involves multi-grained alignment tasks (Chen et al. 2024), with fine-grained alignment being especially important due to its potential to reveal precise associations between gene expressions and histological features. Recent alignment methods primarily focus on designing attention mechanisms to simulate soft alignment (Chen et al. 2021; Xu and Chen 2023; Zhou et al. 2023; Ding et al. 2024). However, these attention-based methods lack explicit multi-grained alignment supervision, they decouple the imputation from alignment, leading to scattered attention and weak interpretability. In contrast, CIMA unifies cyclic modality imputation and multi-grained multimodal alignment to achieve the integrative survival prediction.

The Proposed Methodology

Construction of Multimodal Bags

Given the complex structure of histopathology (WSI) data $\mathbf{X}_i^{\mathcal{H}}$ and genomics (gene expression) data $\mathbf{X}_i^{\mathcal{G}}$ for the i-th patient, we model them as bags composed with instances (WSI patches and biological pathways) for fine-grained analysis using the multi-instance learning paradigm (Carbonneau et al. 2018; Wang et al. 2022). Specifically, the histopathology bag $\mathbf{B}_i^{\mathcal{H}} = \{\mathbf{b}_{i,n}^{\mathcal{H}}\}_{n=1}^{N_i^{\mathcal{H}}}$ consists of low-dimensional embeddings of the tissue image patches it contains, where $N_i^{\mathcal{H}}$ denotes the number of non-overlapping patches into which $\mathbf{X}_i^{\mathcal{H}}$ is segmented. Similarly, the genomics data is first organized in the form of biological pathways, guided by the pathway-gene relationship matrix $\mathbf{M}^{\mathcal{G}\to\mathcal{P}} \in \mathbb{R}^{N^{\mathcal{P}}\times N^{\mathcal{G}}}$. Here, $N^{\mathcal{P}}$ and $N^{\mathcal{G}}$ are the number of pathways and genes, respectively. These pathway-organized genomics (referred as pathway expressions $\mathbf{X}_i^{\mathcal{P}}$) are subsequently fed into the pathway encoder $f_{enc}^{\mathcal{P}}(\cdot)$ and constitute the genomics bag $\mathbf{B}_i^{\mathcal{P}}$. The bag formulations for histopathology and genomics are detailed in **Appendix B.2**.

Cyclic Modality Reconstruction

WSIs, as the gold standard for tumor diagnosis, play an indispensable role in clinical practice, while genomic data remains relatively scarce. This disparity significantly limits the clinical applicability of most multimodal survival prediction methods. To address this challenge, we first design a CRM module with the *Genomics Recovery* sub-module to impute the genomics data using available histopathology, and the *Histopathology Reconstruction* sub-module to reconstruct the corresponding pathological features using the imputed genomics. Through this cyclic reconstruction process, our CIMA not only effectively handles the incomplete multimodal data, but also captures richer cross-modal information and minimizes the risk of accumulating errors during the generation process.

A. Genomics Recovery. This sub-module is responsible for imputing genomic data using the available WSI from the patient. It is worth noting that we do not aim to directly recover gene expressions; instead, considering the subsequent alignment tasks, we focus on reconstructing pathway expressions. We model the pathway expressions $\mathbf{X}_i^{\mathcal{P}}$ as the outcome of the interplay among pathway sequences $\mathbf{h}_m^{\mathcal{P}-seq}$, individual states $\mathbf{h}_i^{\mathcal{I}}$ aggregated from WSI patches, and the broader environmental context ϵ_i (a more detailed explanation of these factors can be found in **Appendix B.3**) as follows:

$$\widetilde{\mathbf{x}}_{i,m}^{\mathcal{P}} = g_{pred}^{\mathcal{P}}(\mathbf{h}_i^{\mathcal{I}} + \mathbf{h}_m^{\mathcal{P}-seq} + \epsilon_i). \tag{1}$$

Due to the scarcity of genomics data, $g_{pred}^{\mathcal{P}}(\cdot)$ typically cannot receive sufficient supervision from limited modality-complete samples. Inspired by the concept of shape prior knowledge (Kuo et al. 2019), we introduce the modality priors learned from N^{mc} modality-complete samples as additional supervision to guide $g_{pred}^{\mathcal{P}}(\cdot)$. Formally, the modality prior $\mathcal{S}_{mp}$ is a set of key-value pairs, where each pair consists of the individual state embedding and pathway expressions from a modality-complete sample, namely $\mathcal{S}_{mp} = \{(\mathbf{h}_j^{\mathcal{I}}, \mathbf{X}_j^{\mathcal{P}})\}_{j=1}^{N^{mc}}$. Thus, we directly adopt the original data of modality-complete samples as their supervision and use the weighted sum of modality priors as pseudo-supervision for modality-incomplete samples, and define the genomic recovery loss as follows:

$$\begin{aligned}\mathcal{L}_{\mathcal{P}}^{REC} = &\frac{1}{N^{mc}}\sum\nolimits_{i=1}^{N}\frac{\Lambda_i}{N^{\mathcal{P}}}\left\|\mathbf{X}_i^{\mathcal{P}} - \widetilde{\mathbf{X}}_i^{\mathcal{P}}\right\|_2^2 \\ &+ \frac{1}{N^{mic}}\sum\nolimits_{i=1}^{N}\frac{1-\Lambda_i}{N^{\mathcal{P}}}\left\|\frac{\sum_j s_{ij}\mathbf{X}_j^{\mathcal{P}} - \widetilde{\mathbf{X}}_i^{\mathcal{P}}}{\sum_{j\in\mathcal{S}_{mp}} s_{ij}}\right\|_2^2,\end{aligned} \tag{2}$$

where N^{mic} represents the number of modality-incomplete samples. The binary variable $\Lambda_i \in \{0, 1\}$ indicates whether the multimodal data for the i-th patient is complete or not. s_{ij} represents the cosine similarity between $\mathbf{h}_i^{\mathcal{I}}$ for the i-th patients and $\mathbf{h}_j^{\mathcal{I}}$ from $\mathcal{S}_{mp}$.

B. Histopathology Reconstruction. To avoid accumulating excessive errors when imputing the missing genomics data, we make efforts in two aspects. On one hand, we introduce Eq. (2) to constrain the generated expressions to fall within the authentic genomics characteristics space. On the other hand, we expect the reconstructed modality to retain as much information as possible with the available ones. Here, we quantify this expectation using the generated pathway expressions to restore the histopathology.

The histopathology is the biological outcome of pathway regulatory effects and environmental context. In other words, the aggregation of all pathway effects ultimately leads to the histopathological phenotype of a given WSI patch. Moreover, to distinguish different patches, we incorporate the positional information of each patch within the WSI during the reconstruction procedure, and restore the n-th patch as follows:

$$\widetilde{\mathbf{b}}_{i,n}^{\mathcal{H}} = \frac{1}{N^{\mathcal{P}}}\sum\nolimits_{m=1}^{N^{\mathcal{P}}} g_{pred}^{\mathcal{H}}\left(f_{enc}^{\mathcal{P}}(\widetilde{\mathbf{x}}_{i,m}^{\mathcal{P}}) + \mathbf{h}_{i,n}^{pos} + \epsilon_i\right), \tag{3}$$

where $g^{\mathcal{H}}_{pred}(\cdot)$ represents the patch predictor, and $\mathbf{h}^{pos}_{i,n}$ denotes the 2D absolute position embedding (Dosovitskiy 2021) for the n-th patch (illustrated in **Appendix B.4**).

Reinterpreting Eq. (3), we view the reconstruction of the n-th patch by the m-th pathway to reflect the activation level of that pathway on the patch. Therefore, we define the activation interaction $a^{AI}_{i,nm}$ between the patch and pathway as the cosine similarity $\cos(\widetilde{\mathbf{b}}^{\mathcal{H}}_{i,nm}, \mathbf{b}^{\mathcal{H}}_{i,n})$, where $\widetilde{\mathbf{b}}^{\mathcal{H}}_{i,nm}$ is the partial reconstruction of the patch from the m-th pathway. Consequently, we can construct the activation interaction map $\mathbf{A}^{AI}_i = [a^{AI}_{i,nm}]^{N^{\mathcal{H}} \times N^{\mathcal{P}}}_{n=1,m=1}$, which will be utilized to guide the multimodal alignment and integration. Reconstructing all patches from each individual pathway to obtain the activation interaction map is computationally prohibitive. To simplify this procedure, we reconstruct the histopathological modality using genomic information only once, and reformulate the generation of activation interaction as a mutual information estimation between different modalities. More details are provided in **Appendix B.5**.

Considering the widespread availability of WSI, the histopathology reconstruction loss can be calculated as:

$$\mathcal{L}^{REC}_{\mathcal{H}} = \frac{1}{N}\sum\nolimits_{i=1}^{N} \frac{1}{N^{\mathcal{H}}_i} \left\| \mathbf{B}^{\mathcal{H}}_i - \widetilde{\mathbf{B}}^{\mathcal{H}}_i \right\|^2_2. \tag{4}$$

To sum up, the training loss for CMR module is:

$$\mathcal{L}_{recon} = (\mathcal{L}^{REC}_{\mathcal{P}} + \mathcal{L}^{REC}_{\mathcal{H}})/2. \tag{5}$$

Multi-grained Multimodal Alignment

To more realistically model the interactions between histopathology and genomics, we explore different levels of interactions between modalities. We first construct the modality-shared projector to map the imputed WSI and genomics bags $\widehat{\mathbf{B}}^{\mathcal{H}}_i$ and $\widehat{\mathbf{B}}^{\mathcal{P}}_i$ into the high-level semantic space, respectively. Next, we design two levels of alignment units to modulate cross-modal bags in the semantic space: (i) *Bag-level Coarse-grained Alignment* (BCA) focuses on exploring the shared semantics between different modalities. (ii) *Instance-level Fine-grained Alignment* (IFA) aims to explicitly capture many-to-many interactions between WSI patches and biological pathways.

A. Bag-level Coarse-grained Alignment. This alignment aims to constrain different modalities into the shared semantic space, thereby capturing cross-modal consistency and facilitating subsequent fine-grained alignment. BCA first applies global max pooling to multimodal bags in the semantic space to obtain bag-level representations. Then, it formulates the bag-level contrastive learning task within the space with the goal of maximizing the consistency between different modalities of the same patient. Specifically, given the WSI bag $\widehat{\mathbf{B}}^{\mathcal{H}}_i$ and genomics bag $\widehat{\mathbf{B}}^{\mathcal{P}}_i$, BCA treats them as a positive pair and combines them with other patient bags ($\widehat{\mathbf{B}}^{\mathcal{P}}_j$ or $\widehat{\mathbf{B}}^{\mathcal{H}}_j$) as negative pairs. Then it defines the bag-level contrastive loss anchored by histopathology for the i-th patient as:

$$\ell^{BCA}_{\mathcal{H},i} = -\log \frac{\exp\left(\cos\left(\rho(\widehat{\mathbf{B}}^{\mathcal{H}}_i), \rho(\widehat{\mathbf{B}}^{\mathcal{P}}_i)\right)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\cos\left(\rho(\widehat{\mathbf{B}}^{\mathcal{H}}_i), \rho(\widehat{\mathbf{B}}^{\mathcal{P}}_j)\right)/\tau\right)}, \tag{6}$$

where ρ denotes the global max pooling operation, and τ is a temperature parameter. In a symmetrical manner, we can obtain the bag-level loss $\ell^{BCA}_{\mathcal{P},i}$ with genomics as the anchor.

B. Instance-level Fine-grained Alignment. Based on the coarse-grained alignment, we aim to further identify fine-grained alignment among instances of histopathology and genomics bags. For this purpose, IFA needs to quantify the improvement that the interaction between a pair of instances brings to the entire system. The Shapley interaction in game theory provides a solution to this quantification and has been widely applied (Li et al. 2022; Luo et al. 2024). It measures the additional contribution brought by a coalition compared with the case when the players work alone. Here, IFA treats each instance in the multimodal bags of the i-th patient as a player. Given Shapley interaction $a^{SI}_{i,nm}$ for the instance pair $\mathcal{B}^{sub}_{i,nm} = (\mathbf{b}^{\mathcal{H}}_{i,n}, \mathbf{b}^{\mathcal{P}}_{i,m})$, we define the Shapley Interaction map $\mathbf{A}^{SI}_i = [a^{SI}_{i,nm}]^{N^{\mathcal{H}} \times N^{\mathcal{P}}}_{n=1,m=1}$. Details about the Shapley Interaction and its calculation pipeline can be found in **Appendix B.6**.

Next, we design a specialized guidance to encourage fine-grained alignment between instances from different modalities. Unlike the one-to-one coarse-grained alignment between modalities of the same patient in Eq. (6), there is a clear many-to-many relationship between WSI patches and biological pathways. For example, a pathway may be expressed to varying degrees across different patches, while a single patch may exhibit abnormal expression regulated by several pathways. In other words, within the given multimodal bags, there are inevitably multiple positive instance pairs. In this case, simply assigning one positive pair while treating other combinations as negative pairs would result in insufficient or even erroneous alignments. Inspired by supervised contrastive learning (Khosla et al. 2020), IFA utilizes the fine-grained interaction $a^{SI}_{i,nm}$ to weigh the contrastive loss term for each instance pair as follows:

$$\begin{aligned} \ell^{IFA}_{\mathcal{H},i} &= \frac{1}{N^{\mathcal{H}}_i} \sum\nolimits_{n=1}^{N^{\mathcal{H}}_i} \frac{1}{\sum_{m=1}^{N^{\mathcal{P}}} a^{SI}_{i,nm}} \sum_{m=1}^{N^{\mathcal{P}}} a^{SI}_{i,nm} \cdot \ell^{CL}_{i,nm}, \\ \ell^{CL}_{i,nm} &= -\log \frac{\exp\left(\cos\left(\widehat{\mathbf{b}}^{\mathcal{H}}_{i,n}, \widehat{\mathbf{b}}^{\mathcal{P}}_{i,m}\right)/\tau\right)}{\sum_{k=1}^{N^{\mathcal{P}}} \exp\left(\cos\left(\widehat{\mathbf{b}}^{\mathcal{H}}_{i,n}, \widehat{\mathbf{b}}^{\mathcal{P}}_{i,k}\right)/\tau\right)}. \end{aligned} \tag{7}$$

It is worth noting that for modality-complete samples, IFA uses Eq. (7) to promote fine-grained alignment within the multimodal bags. However, for modality-incomplete samples, the learned fine-grained interactions may become biased or even erroneous due to the inevitable biases introduced during the reconstruction process. Therefore, we replace the Shapley interaction $a^{SI}_{i,nm}$ with the activation interaction $a^{AI}_{i,nm}$ for weighting in such cases. In this way, IFA overcomes the drawback of attention-based alignments

(Chen et al. 2021; Ding et al. 2024), which are often guided by downstream task-specific losses, without explicit guidance for alignments between histopathology and genomics.

Similar to the coarse-grained alignment module, we can also derive the instance-level contrastive loss term $\ell_{\mathcal{P},i}^{IFA}$ anchored by genomics here. Finally, by combining the coarse- and fine-granularity, the overall loss of the multimodal alignment module is:

$$\mathcal{L}_{align} = \frac{1}{2N}\sum_{i=1}^{N}\left(\ell_{\mathcal{H},i}^{BCA} + \ell_{\mathcal{P},i}^{BCA}\right) + \left(\ell_{\mathcal{H},i}^{IFA} + \ell_{\mathcal{P},i}^{IFA}\right). \quad (8)$$

Adaptive Multimodal Fusion

After meticulously modeling the multimodal bags, we aim to design an effective fusion mechanism that integrates the thousands of instances from different bags into a compact multimodal representation. To this end, we design the interaction map-guided adaptive fusion process. In concrete terms, given the quadruple $\{\mathbf{B}_i^{\mathcal{H}}, \mathbf{B}_i^{\mathcal{P}}, \mathbf{A}_i^{AI}, \mathbf{A}_i^{SI}\}$ for the i-th patient, we sequentially apply bilinear pooling layers over the Activation Interaction map $\mathbf{A}_i^{AI}$ and the Shapley Interaction map $\mathbf{A}_i^{SI}$.

We first perform bilinear pooling over $\mathbf{A}_i^{AI}$ to obtain the intermediate representation $\mathbf{z}_i' \in \mathbb{R}^d$, where the k-th element can be computed as follows:

$$z_{i,k}' = \left(\mathbf{B}_i^{\mathcal{H}}\mathbf{W}^{\mathcal{H}}\right)_{[:,k]}^{\top} \mathbf{A}_i^{AI} \left(\mathbf{B}_i^{\mathcal{P}}\mathbf{W}^{\mathcal{P}}\right)_{[:,k]} \quad (9)$$

where $\mathbf{W}^{\mathcal{H}}, \mathbf{W}^{\mathcal{P}} \in \mathbb{R}^{d\times d}$ are learnable transformation matrices, $(\cdot)_{[:,k]}$ denotes the column index for the inner matrix. For convenience, we denote Eq. (9) as $\mathbf{z}_i' = f_{pool}(\mathbf{B}_i^{\mathcal{H}}, \mathbf{B}_i^{\mathcal{P}}; \mathbf{A}_i^{AI})$, where $f_{pool}(\cdot)$ is the pooling function that integrates two multi-instance inputs over the given map. Next, we apply pooling over the $\mathbf{A}_i^{SI}$ map, which indicates the fine-grained interactions between patches and pathways, giving the final representation $\mathbf{z}_i$ as follows:

$$\mathbf{z}_i = f_{pool}\left(\mathbf{B}_i^{\mathcal{H}} + \mathbf{z}_i'(\mathbf{1}_i^{\mathcal{H}})^{\top}, \mathbf{B}_i^{\mathcal{P}} + \mathbf{z}_i'(\mathbf{1}^{\mathcal{P}})^{\top}; \mathbf{A}_i^{SI}\right), \quad (10)$$

where $\mathbf{1}_i^{\mathcal{H}} \in \mathbb{R}^{N_i^{\mathcal{H}}}$ and $\mathbf{1}^{\mathcal{P}} \in \mathbb{R}^{N^{\mathcal{P}}}$ are vectors of ones. The first two inputs of $f_{pool}(\cdot)$ can be intuitively viewed as establishing residual connections between the multimodal bags and $\mathbf{z}_i'$. Notably, the transformation matrices $\mathbf{W}^{\mathcal{H}}$ and $\mathbf{W}^{\mathcal{P}}$ are *shared* across these pooling operations to reduce the number of parameters and alleviate overfitting.

Given the multimodal representation $\mathbf{z}_i$ for the i-th patient, the goal is to estimate the risk probability of an outcome event occurring before a certain time. As discussed in prior works (Zadeh and Schmid 2020; Chen et al. 2022), this objective can be formulated as a classification task with censorship information, where the i-th patient is represented by the triplet $\{\mathbf{z}_i, c_i, y_i^{surv}\}$. Here, c_i is a binary indicator, with $c_i = 0$ denoting an observed death, and $c_i = 1$ indicating an unknown outcome. The variable $y_i^{surv} \in \{1, 2, \cdots, N^t\}$ is the discrete time label, meaning that the survival time of the patient falls within the y_i^{surv}-th time interval. These N^t intervals are derived from the evenly divided quantiles of survival times for uncensored patients. Thus, the survival loss can be defined as:

$$\begin{aligned}\mathcal{L}_{surv} = -\sum_{i=1}^{N} & c_i \log\left(f_{surv}(y_i^{surv}|\mathbf{z}_i)\right) \\ & + (1-c_i)\log\left(f_{surv}(y_i^{surv}-1|\mathbf{z}_i)\right) \\ & + (1-c_i)\log\left(f_{hazard}(y_i^{surv}|\mathbf{z}_i)\right),\end{aligned} \quad (11)$$

where $f_{hazard}(y_i^{surv}|\mathbf{z}_i)$ is the conditional hazard function, it estimates the probability of the death event occurring within the y_i^{surv}-th time interval. The discrete survival function $f_{surv}(y_i^{surv}|\mathbf{z}_i)$ measures the probability that the patient survives until the y_i^{surv}-th time interval. More details for the survival prediction procedure are given in **Appendix B.7**.

Overall Loss

To sum up, CIMA is an end-to-end architecture involving three key modules that are optimized simultaneously in uniform, and the total loss can be calculated based on the individual loss of each module as:

$$\mathcal{L} = \lambda_1\mathcal{L}_{recon} + \lambda_2\mathcal{L}_{align} + \mathcal{L}_{surv}, \quad (12)$$

where λ_1, and λ_2 are trade-off parameters among three individual losses. In practice, we set both of them to 1 by default. The whole procedure of CIMA is summarized in Algorithm 1 in **Appendix B.1**. It is worth noting that although we integrate the loss from each module into $\mathcal{L}$, each loss only affects its corresponding sub-networks during the backpropagation process. For instance, $\mathcal{L}_{align}$ updates only the encoders of each modality without affecting other modules.

Experiments Results and Analysis

Experiment Settings

To validate the effectiveness of our proposed CIMA in integrating histopathology and genomics for cancer survival prediction, we conducted extensive experiments on five benchmark cancer datasets from TCGA (Weinstein et al. 2013): breast invasive carcinoma (BRCA), colon & rectum adenocarcinoma (COADREAD), glioblastoma multiforme & low-grade glioma (GBMLGG), lung adenocarcinoma & squamous cell carcinoma (LUADLUSC), and stomach adenocarcinoma (STAD). For each patient, we collected the WSI used for primary diagnosis, RNA-Seq expression, and corresponding clinical data from TCGA and cBioPortal (Cerami et al. 2012) (the overview of these datasets and details of data preprocessing can be found in **Appendix C.1** and **C.2**). Among them, we filtered and processed the expression data using the collected biological pathways (Jaume et al. 2024; Zhang et al. 2024). We employed 5-fold cross-validation on each cancer dataset and evaluated the model using the Concordance Index (Harrell Jr, Lee, and Mark 1996) (C-index) and its standard deviation (std), thereby quantifying the performance of correctly ranking the predicted patient risk scores for disease-specific survival. We also employed Kaplan-Meier (KM) curves (Kaplan and Meier 1958) to visualize the survival probabilities of different risk groups.

We compare CIMA against thirteen representative and competitive algorithms, which could be categorized into

Type	Method	BRCA (N=1017)	COADREAD (N=534)	GBMLGG (N=567)	LUADLUSC (N=824)	STAD (N=326)	Overall
Unimodal (Genomics)	MLP	.586±.011	.517±.020	.615±.013	.518±.004	.526±.023	.552
	SNN	.600±.017	.529±.019	.594±.008	.525±.008	.547±.020	.559
	SNNTrans	.593±.017	.524±.017	.609±.021	.523±.014	.568±.022	.563
Unimodal (WSI)	TransMIL	.640±.014	.541±.021	.630±.008	.553±.019	.560±.020	.585
	DTFD-MIL	.637±.011	.571±.025	.612±.016	.545±.017	.567±.023	.586
	WiKG	.683±.015	.562±.010	.629±.015	.535±.013	.576±.023	.597
Multimodal	Porpoise(Cat)	.674±.018	.556±.029	.622±.012	.545±.010	.566±.014	.593
	Porpoise(KP)	.687±.016	.563±.014	.637±.012	.558±.016	.577±.026	.604
	MOTCat	.694±.023	.579±.017	.625±.014	.563±.035	.574±.031	.607
	PIBD	<u>.712±.022</u>	.583±.022	.640±.010	<u>.568±.023</u>	<u>.592±.016</u>	.619
	SurvPath	.701±.010	<u>.607±.014</u>	<u>.659±.018</u>	.555±.015	.583±.026	<u>.621</u>
Incomplete Multimodal	DCP	.696±.015	.587±.007	.621±.013	.553±.010	**.595±.022**	.610
	HGCN	.681±.019	.596±.013	.643±.030	.541±.008	.580±.026	.608
	CIMA (ours)	**.714±.011**	**.622±.013**	**.673±.017**	**.580±.017**	.585±.013	**.635**

Table 1: The results (C-index, mean±std) of unimodal, multimodal, and incomplete multimodal methods over five cancer datasets under complete multimodal. The best and the second-best results are highlighted in **bold** and <u>underline</u>.

three groups: (a) **Unimodal methods**. For genomics, we adopt MLP, SNN (Klambauer et al. 2017), and SNNTrans that incorporates SNN as the feature extractor and TransMIL (Shao et al. 2021) as the aggregation model for multiple instances. For histopathology, we compare with TransMIL, DTFD-MIL (Zhang et al. 2022), and WiKG (Li et al. 2024). (b) **Multimodal methods**. We select four methods for comparison: Porpoise (Chen et al. 2022), MOTCat (Xu and Chen 2023), SurvPath (Jaume et al. 2024), and PIBD (Zhang et al. 2024), where we employ two fusion approaches, including concatenation (Cat) and Kronecker product (KP). (c) **Incomplete Multimodal methods**, including DCP (Lin et al. 2022) and HGCN (Hou et al. 2023). Among them, we adopt AMIL (Ilse, Tomczak, and Welling 2018) for DCP to aggregate modality bags, thereby extending it into multi-instance learning. More details and configurations of these methods are provided in **Appendix C.3**.

Result and Discussion

Result Analysis under Complete Multimodal

The experimental results presented in Table 1 demonstrate that CIMA consistently makes top performance across five cancer datasets and achieves the best overall performance. Specifically, CIMA gains improvements of 7.2% (compared to SNNTrans for Genomics), 3.8% (compared to WiKG for WSI), 1.4% (compared to PIBD for Multimodal), and 2.5% (compared to DCP for Incomplete Multimodal), respectively. Here, the superiority of CIMA can be attributed to: (i) Different from unimodal baselines, CIMA employs multiple modal perspectives and effectively fuses them, allowing for a more comprehensive understanding of patients' condition. (ii) Compared to multimodal baselines, CIMA explicitly explores and encourages alignment between different modalities, providing more reliable guidance for subsequent integration. (iii) Compared to counterparts using incomplete multimodal data, CIMA not only accounts for the entire reconstruction procedure from a biological perspective, but also achieves sensible survival analysis by multi-instance learning to model complex WSI and genomics data. A more detailed analysis of the results can be found in **Appendix C.4**. For further insights into the effectiveness and efficiency of CIMA, we report the ablation study, hyperparameter and runtime analysis in **Appendix C.5-C.7**.

Result Analysis under Incomplete Multimodal

Due to factors such as limitations in inspection equipment and patient objections, incomplete multimodal clinical data is a typical scenario, particularly for the more costly genomic data. Therefore, we further evaluate the robustness of multimodal methods when confronted with incomplete data. Specifically, we randomly mask the genomic data of patients at ratios of $\{0.25, 0.5, 0.75, 0.9\}$, and then measure the performance of the models. Taking the STAD dataset as an example, we analyze the results of the above experiment.

Figure 2 illustrates the predictive performance of tested methods under different missing rates, as well as the performance decline compared to using complete multimodal data. It is obvious that, as the missing rate rises, the performance of each method declines consistently. When genomics data experiences severely missingness (with a missing rate exceeding 75%), the multimodal baselines become inferior to several unimodal rivals, while the incomplete multimodal methods still maintain considerable competitiveness. Among these, CIMA is least impacted by data missingness, which can primarily be attributed to its cyclic reconstruction design. Since the imputed modality is tasked with recovering the available ones, when the missing rate is high, the generated data will approximate the latent representation of the available modality. This ensures that CIMA does not introduce excessive bias during the reconstruction pro-

cedure, thereby roughly maintaining the performance lower bound similar to that of using unimodal data. Although we mainly focus on imputing missing genomics data, our CRM module can be readily applied to other data missing scenarios.

To further study the robustness of CIMA under incomplete multimodal conditions, we assess its imputation performance against several representative baselines and test CIMA under a severity-dependent missingness setting, and report the results in **Appendix C.8 and C.9**.

Survival Analysis

To validate the effectiveness of CIMA in survival analysis, we divide patients from the testing cohort into high- and low-risk groups based on the median risk scores predicted by the model, and visualize the results using Kaplan-Meier (KM) curves (Kaplan and Meier 1958). Additionally, we employ the log-rank test to assess the statistical significance between different risk groups. As illustrated in Figure 3 (with more results on other datasets can be found in **Appendix C.10**), CIMA can more confidently distinguish patients in different risk groups than the best baselines (WiKG for unimodal, PIBD for multimodal, and DCP for incomplete multimodal), expressing its practical value in cancer prognosis prediction.

Interpretability of Multimodal Alignment

The explored multi-grained alignments of CIMA can provide novel insights into the relationship between biological pathways and histological phenotypes that are crucial for identifying cancer patient risk factors. Here, we compare and analyze the differences between low-risk and high-risk gastric cancer (STAD) cases.

We focus on the biological pathways that contribute significantly to the prediction results (quantified by comparing the change in survival loss before and after masking the pathway) and visualize their interactions with WSI patches. In Figure 4, we present the results for the PI3K/Akt/mTOR signaling pathway (Zhao et al. 2020) and

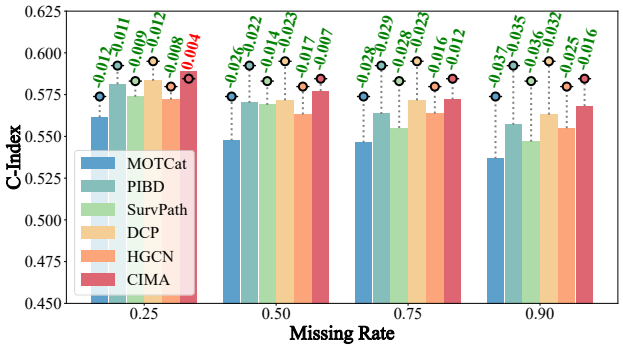

Figure 2: **Prediction performance of tested methods on the STAD dataset**. Solid dots represent the performance of the corresponding method under complete multimodal data, serving as the baseline. We have annotated the performance degradation of each method under different missing rates.

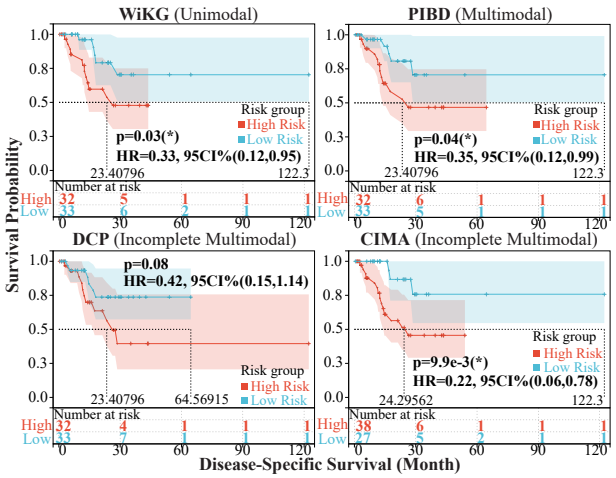

Figure 3: **Kaplan Meier curves of CIMA, compared against unimodal, multimodal, and incomplete multimodal baselines on the STAD dataset**. A p-value < 0.05 (marked with *) indicates statistical significance. HR and 95CI are abbreviations of Hazard Ratio and 95% Confidence Interval. The shaded area represents the confidence interval.

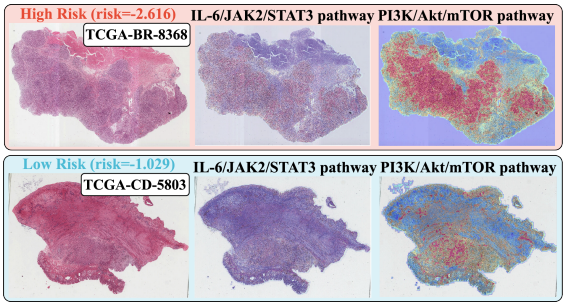

Figure 4: **Interaction between several pathways and WSI patches.** Red indicates the pathway being highly expressed on the patch, while blue indicates a low interaction.

the IL-6/JAK2/STAT3 pathway (Liu et al. 2022), which are recognized to regulate the onset and progression of gastric cancer. We can observe that these two pathways exhibit lower interactions with patches from low-risk patients, but higher interactions with the high-risk group. This shows the authenticity of the learned interaction map and the interpretability of CIMA.

Conclusion

We present CIMA, an end-to-end framework for multimodal cancer survival prediction that harmonizes the missing modality generation and multi-grained alignment. CIMA not only flexibly addresses typical multimodal data incompleteness scenarios in clinical practice but also effectively identifies key interactions between biological pathways and histopathology patches, providing valuable insights for exploring the mechanisms underlying cancer initiation and progression. Experimental results on cancer datasets validate its effectiveness and highlight its authenticity.

Acknowledgments

This work is supported by National Key Research and Development Program of China (No. 2023YFF0725500), NSFC (62531013, 62031003 and 62272276), Shandong Provincial Natural Science Foundation (No. ZR2024JQ001), Taishan Scholars Program (No. tsqn202306007 and tsqn202408317).

References

Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2018. Multimodal machine learning: a survey and taxonomy. *TPAMI*, 41(2): 423–443.

Boehm, K. M.; Khosravi, P.; Vanguri, R.; Gao, J.; and Shah, S. P. 2022. Harnessing multimodal data integration to advance precision oncology. *Nat. Rev. Cancer*, 22(2): 114–126.

Bray, F.; Laversanne, M.; Sung, H.; Ferlay, J.; Siegel, R. L.; Soerjomataram, I.; and Jemal, A. 2024. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.*, 74(3): 229–263.

Bu, Y.; Liang, J.; Li, Z.; Wang, J.; Wang, J.; and Yu, G. 2024. Cancer molecular subtyping using limited multi-omics data with missingness. *PLOS Comp. Biol.*, 20(12): e1012710.

Carbonneau, M.-A.; Cheplygina, V.; Granger, E.; and Gagnon, G. 2018. Multiple instance learning: a survey of problem characteristics and applications. *Pattern Recog.*, 77: 329–353.

Cerami, E.; Gao, J.; Dogrusoz, U.; Gross, B. E.; Sumer, S. O.; Aksoy, B. A.; Jacobsen, A.; Byrne, C. J.; Heuer, M. L.; Larsson, E.; et al. 2012. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.*, 2(5): 401–404.

Chen, R. J.; Lu, M. Y.; Weng, W.-H.; Chen, T. Y.; Williamson, D. F.; Manz, T.; Shady, M.; and Mahmood, F. 2021. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *ICCV*, 4015–4025.

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. 2022. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, 40(8): 865–878.

Chen, Y.; Xie, J.; Lin, Y.; Song, Y.; Yang, W.; and Yu, R. 2024. Survmamba: state space model with multi-grained multi-modal interaction for survival prediction. *arXiv preprint arXiv:2404.08027*.

David, C. R.; et al. 1972. Regression models and life tables (with discussion). *J. R. Stat. Soc.*, 34(2): 187–220.

Ding, S.; Li, J.; Wang, J.; Ying, S.; and Shi, J. 2024. Multimodal co-attention fusion network with online data augmentation for cancer subtype classification. *TMI*, 43(11): 3977–3989.

Dosovitskiy, A. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

Dwivedi, C.; Nofallah, S.; Pouryahya, M.; Iyer, J.; Leidal, K.; Chung, C.; Watkins, T.; Billin, A.; Myers, R.; Abel, J.; et al. 2022. Multi stain graph fusion for multimodal integration in pathology. In *CVPR*, 1835–1845.

Harrell Jr, F. E.; Lee, K. L.; and Mark, D. B. 1996. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. in Med.*, 15(4): 361–387.

Hou, W.; Lin, C.; Yu, L.; Qin, J.; Yu, R.; and Wang, L. 2023. Hybrid graph convolutional network with online masked autoencoder for robust multimodal cancer survival prediction. *TMI*, 42(8): 2462–2473.

Ilse, M.; Tomczak, J.; and Welling, M. 2018. Attention-based deep multiple instance learning. In *ICML*, 2127–2136.

Jaume, G.; Vaidya, A.; Chen, R. J.; Williamson, D. F.; Liang, P. P.; and Mahmood, F. 2024. Modeling dense multimodal interactions between biological pathways and histology for survival prediction. In *CVPR*, 11579–11590.

Kaplan, E. L.; and Meier, P. 1958. Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.*, 53(282): 457–481.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. *NeurIPS*, 18661–18673.

Klambauer, G.; Unterthiner, T.; Mayr, A.; and Hochreiter, S. 2017. Self-normalizing neural networks. *NeurIPS*, 971–980.

Kuo, W.; Angelova, A.; Malik, J.; and Lin, T.-Y. 2019. Shapemask: learning to segment novel objects by refining shape priors. In *ICCV*, 9207–9216.

Li, J.; Chen, Y.; Chu, H.; Sun, Q.; Guan, T.; Han, A.; and He, Y. 2024. Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis. In *CVPR*, 11323–11332.

Li, J.; He, X.; Wei, L.; Qian, L.; Zhu, L.; Xie, L.; Zhuang, Y.; Tian, Q.; and Tang, S. 2022. Fine-grained semantically aligned vision-language pre-training. *NeurIPS*, 7290–7303.

Lin, Y.; Gou, Y.; Liu, X.; Bai, J.; Lv, J.; and Peng, X. 2022. Dual contrastive prediction for incomplete multi-view representation learning. *TPAMI*, 45(4): 4447–4461.

Liu, M.; Li, H.; Zhang, H.; Zhou, H.; Jiao, T.; Feng, M.; Na, F.; Sun, M.; Zhao, M.; Xue, L.; et al. 2022. RBMS1 promotes gastric cancer metastasis through autocrine IL-6/JAK2/STAT3 signaling. *Cell Death & Disease*, 13(3): 287.

Luo, W.; Xia, Y.; Tianshu, S.; and Li, S. 2024. Shapley value-based contrastive alignment for multimodal information extraction. In *ACM MM*, 5270–5279.

Miao, X.; Wu, Y.; Chen, L.; Gao, Y.; and Yin, J. 2022. An experimental survey of missing data imputation algorithms. *TKDE*, 35(7): 6630–6650.

Qiu, S.; Wang, M.; Yang, Y.; Yu, G.; Wang, J.; Yan, Z.; Domeniconi, C.; and Guo, M. 2023. Meta multi-instance multi-label learning by heterogeneous network fusion. *Information Fusion*, 94: 272–283.

Shao, Z.; Bian, H.; Chen, Y.; Wang, Y.; Zhang, J.; Ji, X.; and Zhang, Y. 2021. Transmil: Transformer based correlated

multiple instance learning for whole slide image classification. *NeurIPS*, 2136–2147.

Togasaki, K.; Sugimoto, S.; Ohta, Y.; Nanki, K.; Matano, M.; Takahashi, S.; Fujii, M.; Kanai, T.; and Sato, T. 2021. Wnt signaling shapes the histologic variation in diffuse gastric cancer. *Gastroenterology*, 160(3): 823–830.

Unger, M.; and Kather, J. N. 2024. Deep learning in cancer genomics and histopathology. *Genome Med.*, 16(1): 44.

Van der Laak, J.; Litjens, G.; and Ciompi, F. 2021. Deep learning in histopathology: the path to the clinic. *Nat. Med.*, 27(5): 775–784.

van Loon, W.; Fokkema, M.; de Vos, F.; Koini, M.; Schmidt, R.; and de Rooij, M. 2024. Imputation of missing values in multi-view data. *Inf. Fusion*, 102524.

Wang, X.; Yu, G.; Wang, J.; Zain, A. M.; and Guo, W. 2022. Lung cancer subtype diagnosis using weakly-paired multi-omics data. *Bioinfo.*, 38(22): 5092–5099.

Weinstein, J. N.; Collisson, E. A.; Mills, G. B.; Shaw, K. R.; Ozenberger, B. A.; Ellrott, K.; Shmulevich, I.; Sander, C.; and Stuart, J. M. 2013. The cancer genome atlas pan-cancer analysis project. *Nat. Genet.*, 45(10): 1113–1120.

Wiegrebe, S.; Kopper, P.; Sonabend, R.; Bischl, B.; and Bender, A. 2024. Deep learning for survival analysis: a review. *Artif. Intell. Rev.*, 57(3): 65.

Wu, B.; Du, W.; Wang, J.; and Yu, G. 2025. Imputation-free incomplete multi-view clustering via knowledge distillation. In *IJCAI*, 6570–6578.

Xu, H.; Xu, Q.; Cong, F.; Kang, J.; Han, C.; Liu, Z.; Madabhushi, A.; and Lu, C. 2023. Vision transformers for computational histopathology. *Rev. in Biomed. Eng.*, 17: 63–79.

Xu, Y.; and Chen, H. 2023. Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In *ICCV*, 21241–21251.

Zadeh, S. G.; and Schmid, M. 2020. Bias in cross-entropy-based training of deep survival networks. *TPAMI*, 43(9): 3126–3137.

Zhang, C.; Yang, Z.; He, X.; and Deng, L. 2020. Multimodal intelligence: representation learning, information fusion, and applications. *JSTSP*, 14(3): 478–493.

Zhang, H.; Meng, Y.; Zhao, Y.; Qiao, Y.; Yang, X.; Coupland, S. E.; and Zheng, Y. 2022. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *CVPR*, 18802–18812.

Zhang, Y.; Xu, Y.; Chen, J.; Xie, F.; and Chen, H. 2024. Prototypical information bottlenecking and disentangling for multimodal cancer survival prediction. In *ICLR*.

Zhao, Q.; Zhao, Y.; Hu, W.; Zhang, Y.; Wu, X.; Lu, J.; Li, M.; Li, W.; Wu, W.; Wang, J.; et al. 2020. m6A RNA modification modulates PI3K/Akt/mTOR signal pathway in gastrointestinal cancer. *Theranostics*, 10(21): 9528.

Zhou, H.; Zhou, F.; Zhao, C.; Xu, Y.; Luo, L.; and Chen, H. 2024. Multimodal data integration for precision oncology: challenges and future directions. *arXiv preprint arXiv:2406.19611*.

Zhou, H.-Y.; Yu, Y.; Wang, C.; Zhang, S.; Gao, Y.; Pan, J.; Shao, J.; Lu, G.; Zhang, K.; and Li, W. 2023. A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. *Nat. Biomed. Eng.*, 7(6): 743–755.