

CIGTSurv: Clinical Information Guided Tri-modal Survival Prediction with Local Prototype Association and Global Feature Alignment

Jing Dai^{1,2}, Qibin Zhang², Weiwei Zhou², Mingde Xu², Jingsong Liu^{3,4},
Jingdong Zhang^{1,2}, and Hongming Xu^{1,2,5}✉

¹ Cancer Hospital of Dalian University of Technology, Shenyang, China

² School of Biomedical Engineering, Faculty of Medicine,
Dalian University of Technology, Dalian, China

³ Institute of Pathology, Technical University of Munich,
TUM School of Medicine and Health, Munich, Germany

⁴ Munich Center for Machine Learning (MCML), Munich, Germany

⁵ Key Laboratory of Integrated Circuit and Biomedical Electronic System,
Dalian University of Technology, Dalian, China
{mxu}@dlut.edu.cn

Abstract. Multimodal learning has significantly advanced survival prediction by integrating pathology images with genomic data. However, clinical information, despite its critical role in reflecting a patient’s overall health, remains underutilized due to its discrete, sparse, and low-dimensional nature. Furthermore, the inherent heterogeneity across these modalities pose significant challenges in modeling cross-modal interactions. In this paper, we propose CIGTSurv, a Clinical Information Guided Tri-modal framework for Survival prediction. Specifically, we first design a holistic text template and use pretrained foundation models to transform clinical tabular data into high-dimensional tokenized embeddings. Using clinical information as an anchor, we then introduce a dual-level interaction mechanism: 1) a local prototype association (LPA) module based on cross-attention to explicitly learn token-level correspondences between different modalities, and 2) a global feature alignment (GFA) loss based on Maximum Mean Discrepancy (MMD) to implicitly enhance cross-modal distribution consistency. Extensive experiments on five TCGA cancer cohorts demonstrate that CIGTSurv achieves state-of-the-art (SOTA) survival prediction performance. Our source code is publicly available at <https://github.com/Daijing-ai/CIGT-Surv.git>.

Keywords: Multimodal Learning · Computational Pathology · Survival Prediction · Text Encoding.

1 Introduction

Survival prediction is a cornerstone of clinical oncology, dedicated to estimating the time-to-event for critical outcomes. Accurate prediction provides essential

insights into disease progression, treatment response, and patient prognosis, ultimately guiding personalized therapeutic decision making [1].

With the rapid advancement of digital pathology, whole-slide images (WSIs) have been increasingly used for survival prediction. WSIs provide a comprehensive view of morphological changes at the cellular and tissue levels, offering a critical basis for survival assessment. Given the large number of patches extracted from WSIs, many methods [2,3,4,5] adopt Multiple Instance Learning (MIL) [6] for efficient processing and analysis. However, pathology images alone cannot comprehensively reflect the entire process of cancer occurrence and development. Therefore, recent advances [7,8,9,10,11,12] that integrate genomic data with WSIs have demonstrated improved performance, as these complementary modalities provide a more comprehensive understanding of the disease.

Nevertheless, clinical information, which provides complete context for survival prediction by enhancing the model’s understanding of a patient’s overall health [13], have not been fully explored. This is mainly due to the discrete and low-dimensional nature of clinical tabular data, which limits their effectiveness in deep learning models that typically excel with high-dimensional features [14]. Traditional encoding techniques, such as one-hot encoding [15], struggle with discrete clinical features like T, N, and M stages, treating different categorical variables as entirely independent entities and overlooking potential intrinsic relationships between them. Although the recent SurvPGC framework [16] demonstrated the importance of incorporating clinical information to improve survival prediction via text templates and foundation models, its variable-wise prompting strategy generates isolated embedding vectors for each indicator, which overlooks the intrinsic semantic dependencies among clinical indicators. Furthermore, the inherent heterogeneity across modalities (e.g., WSIs, genomics and texts) creates a significant gap that hinders effective cross-modal integration. The overwhelming amount of WSI data can easily obscure information from other modalities, ultimately leading to information loss and suboptimal predictive performance.

To address the above challenges, we propose CIGTSurv, a Clinical Information Guided Tri-timodal framework for Survival prediction. Our main contributions are: (1) We design a holistic text template that synthesizes discrete variables into a coherent clinical description and tokenize clinical text into high-dimensional embeddings to support survival prediction. (2) We introduce a local prototype association (LPA) based on cross-attention for explicit learning of token-level correspondences, complemented by a global feature alignment (GFA) based on Maximum Mean Discrepancy (MMD) for implicit distribution alignment. The dual-level interaction mechanism ensures model to exploit both local and global relationships between modalities, thereby enabling comprehensive cross-modal interaction. (3) Extensive experiments across five cancer data cohorts demonstrate the superiority and efficiency of the proposed method compared to SOTA baselines in survival prediction.

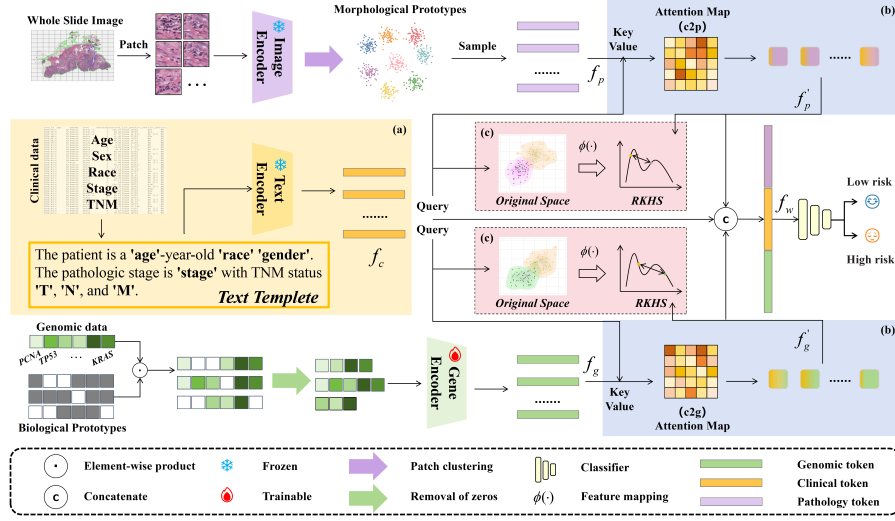


Fig. 1. Overview of our proposed CIGTSurv framework featured by (a) text encoding from clinical information, (b) local prototype association and (c) global feature alignment.

2 Methods

2.1 Overview

The overall architecture of the proposed CIGTSurv is shown in Fig. 1. The framework first encodes raw signals from each modality through modality-specific encoders to generate corresponding unimodal embedding sequences f_p , f_g , and f_c . It then employs a dual-level interaction mechanism consisting of a LPA module for capturing token-level correspondences and a GFA loss for enforcing distribution-level consistency. Using the clinical modality as the anchor, this mechanism yields aligned embedding sequences f'_p and f'_g . These aligned unimodal embeddings, enriched with cross-modal correspondence information, are subsequently concatenated with clinical embedding f_c to form a fused feature representation f_w . Finally, the fused tri-modal features are utilized for survival prediction.

2.2 Text Encoding from Clinical Information

Clinical data are inherently low-dimensional and tabular, making unified feature encoding challenging. To address this, we first design a holistic description to transform clinical information into descriptive sentences that capture contextual semantics. To ensure completeness and prognostic relevance, key patient characteristics including age, race, gender, tumor stage, and tumor TNM status are selected. For every patient, we construct a text template "The patient is a

{age}-year-old {race} {gender}. The pathologic stage is {stage} with TNM status {T}, {N}, and {M}". The resulting text is encoded using Biomed-BERT [17], resulting in a clinical embedding $I_c \in \mathbb{R}^{n_c \times d_c}$, where n_c denotes the number of clinical tokens. Finally, a Multi-Layer Perceptron (MLP) projects this representation into a clinical embedding $f_c \in \mathbb{R}^{n_c \times d}$, with $d = 256$.

2.3 Prototype-based Encoding

Morphological Prototypes from Pathology Images. Given a WSI, we extract N_p patches of size 256×256 at $20\times$ magnification using the CLAM toolbox [3]. Each patch is encoded into a 1024-dimensional feature vector using a pre-trained pathology foundation model UNI [18], yielding a patch-level feature set $I_p \in \mathbb{R}^{N_p \times d_p}$. To reduce feature redundancy and computational overhead while preserving representative tissue information, we adopt a clustering-based patch selection strategy. Following prior work [19], all patch embeddings are grouped into $K = 50$ clusters using K-means. From each cluster, $S = 10$ patches are randomly sampled, resulting in $n_p = 500$ patches per WSI. The selected patch features are then projected via a MLP to obtain the final pathology embedding $f_p \in \mathbb{R}^{n_p \times d}$.

Biological Prototypes from Genomic Data. The genomic data for each sample can be represented by a gene-expression vector $I_g \in \mathbb{R}^{N_G}$, where N_G is the total number of genes. We aim to group these genes into $n_g = 6$ pre-defined biological function sets [8], each acting as a prototype. For each set i we define a binary mask $B_i \in \{0, 1\}^{N_G}$, where each element indicates whether a gene is included in a given set (1) or not (0). To construct the biological prototypes, we perform an element-wise product of the gene expression vector I_g with the binary set vectors (B) [11]. This operation generates a prototype representation:

$$Z_{g,i} = I_g \odot B_i, \quad \forall i \in [1, n_g] \quad (1)$$

The resulting vector $Z_{g,i}$ is then reduced to remove any zero entries, yielding a dense but variable-length representation (due to different gene counts across function sets). To ensure a fixed size embedding, each set is transformed into a feature by two cascaded Self-normalizing Neural Network (SNN) [20] layers. These grouped sets are stacked to obtain genomic embedding $f_g \in \mathbb{R}^{n_g \times d}$.

2.4 Local Prototype Association

Using clinical embedding as anchors, We employ cross-attention to model fine-grained, token-level association across modalities. Specifically, we replace the original sequences used in traditional attention mechanisms [8] with prototypes to compute interactive attention, improving efficiency and reducing computational complexity. For clinic-to-pathology association (c2p), clinical embeddings guide the aggregation of pathology features into clinic-aware pathological concepts $f'_p \in \mathbb{R}^{n_c \times d}$. Formally, let $f_c \in \mathbb{R}^{n_c \times d}$ and $f_p \in \mathbb{R}^{n_p \times d}$ denote the clinical

and pathology features, respectively. The cross-attention mechanism is formulated as:

$$f'_p = \text{CrossAttn}(f_c, f_p) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \quad (2)$$

where the Query (Q), Key (K), and Value (V) matrices are linearly projected from the corresponding modalities:

$$Q = f_c \mathbf{W}_q, \quad K = f_p \mathbf{W}_k, \quad V = f_p \mathbf{W}_v, \quad (3)$$

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{d \times d_k}$ are learnable projection matrices, and d_k denotes the scaling dimension. A similar cross-attention operation is applied for clinic-to-genomic association (c2g), producing fused genomic features $f'_g \in \mathbb{R}^{n_c \times d}$.

2.5 Global Feature Alignment

To implicitly align the global distributions of locally fused features with anchor features, we employ Maximum Mean Discrepancy (MMD) [21]. MMD quantifies the distributional divergence between different modalities by comparing their statistics in a high-dimensional Reproducing Kernel Hilbert Space (RKHS). Specifically, given two sets of feature tokens $X = \{x_i\}_{i=1}^N$ and $Y = \{y_j\}_{j=1}^N$, where N denotes the sequence length, the squared MMD distance is defined as:

$$\text{MMD}^2(X, Y) = \left\| \frac{1}{N} \sum_{i=1}^N \phi(x_i) - \frac{1}{N} \sum_{j=1}^N \phi(y_j) \right\|_{\mathcal{H}}^2, \quad (4)$$

where $\phi(\cdot)$ represents the feature mapping to the RKHS $\mathcal{H}$. In practice, this distance is efficiently computed via the kernel trick:

$$\text{MMD}^2(X, Y) = \frac{1}{N^2} \sum_{i=1}^N \sum_{i'=1}^N k(x_i, x_{i'}) + \frac{1}{N^2} \sum_{j=1}^N \sum_{j'=1}^N k(y_j, y_{j'}) - \frac{2}{N^2} \sum_{i=1}^N \sum_{j=1}^N k(x_i, y_j), \quad (5)$$

where we employ a Gaussian kernel $k(x, y) = \exp\left(-\frac{\|x-y\|_2^2}{2\sigma^2}\right)$ with bandwidth σ .

For the fused pathology features f'_p , genomic features f'_g , and clinical anchor features f_c , we define the global alignment loss as the cumulative of MMD distances across modality pairs:

$$L_{\text{align}} = \text{MMD}^2(f'_p, f_c) + \text{MMD}^2(f'_g, f_c). \quad (6)$$

Minimizing this loss forces the model to bridge the distribution gap between the sets of local tokens across modalities. By integrating explicit local token-level associations with implicit global feature alignment, our framework facilitates multi-granular fusion and effective cross-modal interaction.

2.6 Feature Aggregation and Survival Prediction

For a patient, we apply mean pooling to the clinical features f_c , aligned pathology features f'_p and aligned genomic features f'_g and concatenate these three feature sequences to get the final tri-modal feature vector f_w . This vector is subsequently passed into a linear classifier to generate the final survival prediction. Following prior studies [8,10], we employ the Negative Log-Likelihood (NLL) loss function, as L_{surv} . The total loss for end-to-end model training is then calculated with a weight factor λ :

$$L = L_{\text{surv}} + \lambda L_{\text{align}}. \quad (7)$$

3 Experiments

3.1 Datasets and Implementation Details

We conducted experiments on five cancer cohorts derived from The Cancer Genome Atlas (TCGA)⁶, including Bladder Urothelial Carcinoma (BLCA, $n=337$), Breast Invasive Carcinoma (BRCA, $n=968$), Colon and Rectum Adenocarcinoma (COADREAD, $n = 252$), Liver Cancer (LIHC, $n = 311$) and Kidney Papillary Cell Carcinoma (KIRP, $n = 261$). During preprocessing, incomplete cases were excluded; therefore, all included cases contain complete age, race, gender, stage, and TNM information. We employed 5-fold cross-validation to evaluate our model and other compared methods. The mean Concordance index (C-index) with its standard deviation (std) are reported for a comprehensive performance comparison.

During each fold of cross-validation, models were trained for up to 100 epochs with early stopping based on validation C-index. A batch size of one was used, and gradients were accumulated over 32 iterations prior to backpropagation. Optimization was performed using the Adam optimizer with a learning rate of 2×10^{-4} and a weight decay of 1×10^{-5} . The Gaussian kernel bandwidth σ is computed using the median heuristic. The alignment weight λ is selected from 0.01, 0.05, 0.1, 0.5, 1, with dataset-specific values: BLCA 0.01, BRCA 1, COADREAD 0.1, LIHC 0.01, and KIRP 0.5. To address class imbalance across tasks, a weighted sampling strategy was applied during training. All experiments were implemented in PyTorch and executed on a single NVIDIA RTX 4090 GPU.

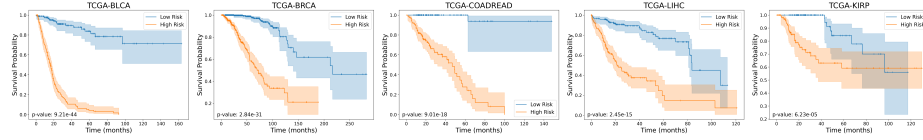
3.2 Performance Evaluation

Comparisons with SOTA methods. We compared our model in five settings: clinic-only (Cox model [22]), pathology-only (ABMIL [2], CLAM-MB [3], TransMIL [4], and RRTMIL [5]), genomic-only (MLP, SNN [20], and SNNTrans [20]), dual-modal (Porpoise [7], MCAT [8], MOTCat [9], SurvPath [10], MMP_{Trans} . [11] and CCL [12]) and tri-modal (Concat, Pathomic [23] and SurvPGC [16]). Table 1

⁶ <https://portal.gdc.cancer.gov/>

Table 1. C-index ($\uparrow$) performance (mean $\pm$ std) over five datasets. The **best** and the **second best** results are highlighted.

Model	BLCA	BRCA	COADREAD	LIHC	KIRP	Mean
Cox(<i>age,sex,stage</i>) [22]	0.645 $\pm$ 0.054	0.772 $\pm$ 0.046	0.724 $\pm$ 0.138	0.627 $\pm$ 0.044	0.832 $\pm$ 0.089	0.720
Geno. MLP	0.657 $\pm$ 0.008	0.647 $\pm$ 0.060	0.604 $\pm$ 0.098	0.621 $\pm$ 0.023	0.823 $\pm$ 0.135	0.670
SNN [20]	0.686 $\pm$ 0.036	0.666 $\pm$ 0.027	0.608 $\pm$ 0.072	0.630 $\pm$ 0.032	0.866 $\pm$ 0.102	0.691
SNNTrans [20]	0.683 $\pm$ 0.018	0.674 $\pm$ 0.048	0.716 $\pm$ 0.121	0.632 $\pm$ 0.058	0.862 $\pm$ 0.109	0.713
Patho. ABMIL [2]	0.688 $\pm$ 0.047	0.711 $\pm$ 0.046	0.764 $\pm$ 0.063	0.733 $\pm$ 0.044	0.873 $\pm$ 0.063	0.754
CLAM-MB [3]	0.664 $\pm$ 0.039	0.697 $\pm$ 0.040	0.804 $\pm$ 0.088	0.728 $\pm$ 0.024	0.847 $\pm$ 0.067	0.753
TransMIL [4]	0.699 $\pm$ 0.049	0.674 $\pm$ 0.024	0.727 $\pm$ 0.107	0.721 $\pm$ 0.036	0.786 $\pm$ 0.136	0.721
RRTMIL [5]	0.639 $\pm$ 0.030	0.702 $\pm$ 0.013	0.783 $\pm$ 0.088	0.722 $\pm$ 0.038	0.870 $\pm$ 0.068	0.743
Dual. Porpoise [7]	0.690 $\pm$ 0.041	0.695 $\pm$ 0.056	0.765 $\pm$ 0.106	0.740 $\pm$ 0.046	0.872 $\pm$ 0.090	0.752
MCAT [8]	0.670 $\pm$ 0.020	0.706 $\pm$ 0.057	0.788 $\pm$ 0.072	0.735 $\pm$ 0.047	0.905 $\pm$ 0.038	0.761
MOTCat [9]	0.702 $\pm$ 0.016	0.720 $\pm$ 0.054	<u>0.815 $\pm$ 0.068</u>	0.736 $\pm$ 0.046	0.901 $\pm$ 0.060	<u>0.775</u>
SurvPath [10]	0.712 $\pm$ 0.037	0.707 $\pm$ 0.028	0.774 $\pm$ 0.087	0.730 $\pm$ 0.049	0.891 $\pm$ 0.065	0.763
MMP _{Trans.} [11]	0.704 $\pm$ 0.036	0.701 $\pm$ 0.024	0.738 $\pm$ 0.124	0.688 $\pm$ 0.044	0.887 $\pm$ 0.086	0.744
CCL [12]	0.691 $\pm$ 0.022	0.660 $\pm$ 0.036	0.763 $\pm$ 0.063	0.630 $\pm$ 0.062	0.814 $\pm$ 0.155	0.712
Tri. Concat	0.701 $\pm$ 0.014	0.683 $\pm$ 0.021	0.782 $\pm$ 0.094	<u>0.747 $\pm$ 0.034</u>	0.873 $\pm$ 0.067	0.757
Pathomic [23]	0.699 $\pm$ 0.037	0.693 $\pm$ 0.045	0.701 $\pm$ 0.065	0.725 $\pm$ 0.041	0.885 $\pm$ 0.087	0.741
SurvPGC [16]	0.689 $\pm$ 0.021	0.722 $\pm$ 0.040	0.792 $\pm$ 0.070	0.724 $\pm$ 0.055	0.881 $\pm$ 0.075	0.762
CIGTSurv (Ours)	0.713 $\pm$ 0.026	<u>0.748 $\pm$ 0.040</u>	0.816 $\pm$ 0.081	0.749 $\pm$ 0.038	0.913 $\pm$ 0.045	0.788

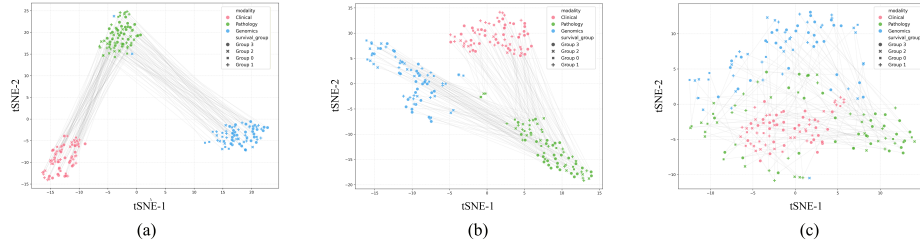
**Fig. 2.** Kaplan-Meier (KM) curves of the proposed model on five cancer datasets, where patient stratification of low risk (blue) and high risk (orange) are presented.

presents the comparative results, where our proposed CIGTSurv achieves an average C-index of 78.8%, surpassing all other comparative methods. The superior performance is mainly attributed to its dual-level mechanism, which effectively bridges the modality gap between clinical information, pathology images, and genomic data via the combined use of cross-attention and MMD loss. While most multimodal approaches surpass unimodal baselines, CIGTSurv further excels by utilizing clinical prompts to guide the extraction of prognostic cues from high-dimensional pathology and genomic signatures. Furthermore, compared to existing tri-modal architectures, our model significantly enhances interaction efficiency by prototype construction, achieving a superior balance between computational overhead and predictive accuracy.

Patient Stratification. To validate the discriminative ability of our model, we performed Kaplan-Meier survival analysis and log-rank test by stratifying patients into high- and low-risk groups according to the median predicted risk scores, as illustrated in Fig. 2. Statistically significant differences ($p < 0.05$) were observed across all datasets, confirming the effectiveness of our model in patient risk stratification.

Table 2. Ablation results of our method

Module	BLCA	BRCA	COADREAD	LIHC	KIRP	Mean
w/o Text Embed.	0.704 $\pm$ 0.027	0.744 $\pm$ 0.090	0.779 $\pm$ 0.078	0.730 $\pm$ 0.038	0.909 $\pm$ 0.044	0.773
w/o Proto.	0.691 $\pm$ 0.025	0.654 $\pm$ 0.045	0.706 $\pm$ 0.119	0.707 $\pm$ 0.072	0.907 $\pm$ 0.031	0.733
w/o LPA	0.709 $\pm$ 0.032	0.692 $\pm$ 0.045	0.768 $\pm$ 0.106	0.737 $\pm$ 0.021	0.891 $\pm$ 0.060	0.759
w/o GFA	0.690 $\pm$ 0.032	0.691 $\pm$ 0.049	0.801 $\pm$ 0.085	0.742 $\pm$ 0.030	0.881 $\pm$ 0.070	0.761
w/o Both	0.694 $\pm$ 0.044	0.692 $\pm$ 0.016	0.769 $\pm$ 0.091	0.736 $\pm$ 0.021	0.860 $\pm$ 0.074	0.750
w/o Clinic	0.709 $\pm$ 0.032	0.712 $\pm$ 0.022	0.766 $\pm$ 0.111	0.736 $\pm$ 0.033	0.883 $\pm$ 0.057	0.761
w/o Pathology	0.664 $\pm$ 0.045	0.678 $\pm$ 0.027	0.706 $\pm$ 0.141	0.668 $\pm$ 0.056	0.884 $\pm$ 0.114	0.720
w/o Genomic	0.690 $\pm$ 0.026	0.682 $\pm$ 0.063	0.795 $\pm$ 0.096	0.745 $\pm$ 0.027	0.892 $\pm$ 0.066	0.761
CIGTSurv(Ours)	0.713$\pm$ 0.026	0.748$\pm$ 0.040	0.816$\pm$ 0.081	0.749$\pm$ 0.038	0.913$\pm$ 0.045	0.788

**Fig. 3.** T-SNE visualization of the feature distributions. (a) Feature distributions in the baseline. (b) Feature distributions after incorporating LPA. (c) Feature distributions by our proposed CIGTSurv framework.

3.3 Ablation Studies

We conduct comprehensive ablation studies from four aspects to evaluate our proposed method, as shown in Table 2.

Effectiveness of Text Encoding. To assess the impact of different encoding strategies for clinical information, we transformed all clinical variables into one-hot encoded vectors [24]. This approach (i.e., w/o Text Embed.) yielded an overall C-index of 0.773, which is lower than that achieved using our text-based clinical embeddings. This indicates that directly encoding clinical variables in a one-hot format may hinder the model from learning effective semantic relationships and introduce spurious associations among features.

Effectiveness of Prototypes. We replaced the morphological and biological prototypes with all patch tokens and genomic data (i.e., w/o Proto.). The C-index consistently dropped across all datasets, indicating that compact prototypes improve the efficiency of cross-modal information interaction.

Effectiveness of Proposed Components. We evaluated the contributions of local prototype association (LPA) module and global feature alignment (GFA) loss. Removing either component led to a performance drop (see Table 2). Furthermore, we employed t-SNE [25] to visualize feature distributions at different stages of our framework, as shown in Fig. 3. In the baseline (Fig. 3a), features form well-separated clusters, indicating a severe modality gap. With the gradual

introduction of cross-attention (Fig. 3b) and MMD loss (Fig. 3c), these clusters gradually converge toward a shared latent space.

Effectiveness of Tri-modal Fusion. Lastly, we conducted modality ablation experiments by removing one modality in turn. Excluding the pathology modality caused a marked performance drop, reflecting the strong correlation between histopathology and survival outcomes. The highest performance was achieved when all three modalities were integrated, confirming the effectiveness of our tri-modal fusion approach.

4 Conclusion

In this work, we propose CIGTSurv, a novel Clinical Information Guided Tri-modal framework for cancer survival prediction. By converting clinical tabular data into holistic, semantic-rich text embeddings, our framework achieves seamless integration of clinical insights with pathology and genomic modalities. The proposed dual-level cross-modal interaction mechanism—comprising local prototype association and global distribution alignment—effectively mitigates the heterogeneous modality gap, yielding a comprehensive and highly discriminative representation. The results highlight the framework’s efficiency and potential to guide future multimodal learning research in oncology.

Acknowledgments. This work was supported in part by the Liaoning Province Science and Technology Joint Program (No. 2024-MSLH-065), the Fundamental Research Funds for Central Universities (No. DUT25Z2514).

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

References

1. Huajun Zhou, Fengtao Zhou, Chenyu Zhao, Yingxue Xu, Luyang Luo, and Hao Chen. Multimodal data integration for precision oncology: Challenges and future directions. *arXiv preprint arXiv:2406.19611*, 2024.
2. Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.
3. Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.
4. Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021.

5. Wenhao Tang, Fengtao Zhou, Sheng Huang, Xiang Zhu, Yi Zhang, and Bo Liu. Feature re-embedding: Towards foundation model-level performance in computational pathology. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11343–11352, 2024.
6. Jaume Amores. Multiple instance classification: Review, taxonomy and comparative study. *Artificial intelligence*, 201:81–105, 2013.
7. Richard J Chen, Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Jana Lipkova, Zahra Noor, Muhammad Shaban, Maha Shady, Mane Williams, Bumjin Joo, et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer cell*, 40(8):865–878, 2022.
8. Richard J Chen, Ming Y Lu, Wei-Hung Weng, Tiffany Y Chen, Drew FK Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4025, 2021.
9. Yingxue Xu and Hao Chen. Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 21241–21251, 2023.
10. Guillaume Jaume, Anurag Vaidya, Richard J Chen, Drew FK Williamson, Paul Pu Liang, and Faisal Mahmood. Modeling dense multimodal interactions between biological pathways and histology for survival prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11579–11590, 2024.
11. Andrew H Song, Richard J Chen, Guillaume Jaume, Anurag J Vaidya, Alexander S Baras, and Faisal Mahmood. Multimodal prototyping for cancer survival prediction. *arXiv preprint arXiv:2407.00224*, 2024.
12. Huajun Zhou, Fengtao Zhou, and Hao Chen. Cohort-individual cooperative learning for multimodal cancer survival analysis. *IEEE Transactions on Medical Imaging*, 44(2):656–667, 2024.
13. Hailong Yang, Jia Wang, Wenyan Wang, Shufang Shi, Lijing Liu, Yuhua Yao, Geng Tian, Peizhen Wang, and Jialiang Yang. Mmsurv: a multimodal multi-instance multi-cancer survival prediction model integrating pathological images, clinical information, and sequencing data. *Briefings in Bioinformatics*, 26(3):bbaf209, 2025.
14. Sandra Steyaert, Marija Pizurica, Divya Nagaraj, Priya Khandelwal, Tina Hernandez-Boussard, Andrew J Gentles, and Olivier Gevaert. Multimodal data fusion for cancer biomarker discovery with deep learning. *Nature machine intelligence*, 5(4):351–362, 2023.
15. Zahra Sedighi-Maman and Alexa Mondello. A two-stage modeling approach for breast cancer survivability prediction. *International journal of medical informatics*, 149:104438, 2021.
16. Jiaxin Hou, Ranran Zhang, Yaoqin Xie, Chao Li, and Wenjian Qin. Multimodal deep learning for cancer prognosis prediction with clinical information prompts integration. *npj Digital Medicine*, 2025.
17. Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.
18. Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban,

- et al. Towards a general-purpose foundation model for computational pathology. *Nature medicine*, 30(3):850–862, 2024.
19. Rui Yan, Xueyuan Zhang, Zihang Jiang, Baizhi Wang, Xiuwu Bian, Fei Ren, and S Kevin Zhou. Pathway-aware multimodal transformer (pamt): Integrating pathological image and gene expression for interpretable cancer survival analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
 20. Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. *Advances in neural information processing systems*, 30, 2017.
 21. Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012.
 22. Norman E Breslow. Analysis of survival data under the proportional hazards model. *International Statistical Review/Revue Internationale de Statistique*, pages 45–57, 1975.
 23. Richard J Chen, Ming Y Lu, Jingwen Wang, Drew FK Williamson, Scott J Rodig, Neal I Lindeman, and Faisal Mahmood. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging*, 41(4):757–770, 2020.
 24. Xinglong Liang, Tianyu Zhang, Miguel Braga, Luyi Han, Maarten Donswijk, Jiaju Huang, Jinhong Song, Chunyao Lu, Xin Wang, Yuan Gao, et al. Multi-omics deep learning improves fdg pet-ct-based long-term prognostication of breast cancer. *npj Precision Oncology*, 2026.
 25. Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.