Di-PACT: Discrete Prognosis via Aligned Co-Topology of Pathways and Tumor Microenvironment

Zhiheng Cai[1], Fulin Li[1], Qing Cai[2⋆], and Shugang Zhang[1]

[1] Ocean University of China, Qingdao, 266100, China
Caizhiheng@stu.ouc.edu.cn, lifulin@stu.ouc.edu.cn, zsg@ouc.edu.cn
[2] Hefei University of Technology, Hefei, 230009, China
caiqing@hfut.edu.cn

Abstract. In precision oncology, multimodal survival analysis is challenged by overfitting from high-dimensional vision foundation model (VFM) features and limited cross-modal interaction in unidirectional fusion. We propose Di-PACT, a multimodal discrete survival framework that integrates whole-slide images with pathway-level transcriptomic representations. Di-PACT combines distance-aware stratified sampling, direct-bottleneck encoding, bidirectional pathway co-attention, and three-branch residual fusion to reduce VFM redundancy, preserve tumor microenvironment heterogeneity, and jointly model WSI-specific, pathway-specific, and cross-modal prognostic information. This design facilitates robust multimodal coordination under modality imbalance and provides multi-level interpretability cues linking prognostic pathways with morphological patterns. A discrete-time survival objective captures non-linear survival dynamics beyond proportional hazards assumptions. Experiments on three TCGA cohorts and two independent CPTAC cohorts demonstrate stable prognostic performance and favorable zero-shot cross-domain generalization compared with evaluated multimodal baselines. The source code is available at `https://github.com/Cai-Zhiheng/Di-PACT`.

Keywords: Multimodal Survival Analysis · Discrete Survival Prediction · Vision Foundation Models · Bidirectional Co-Attention · Pathway-Level Interpretability

1 Introduction

In precision oncology, integrating whole-slide images (WSIs) and genomic data is essential for disease diagnosis and survival prognosis [1–3]. Vision foundation models (VFMs, e.g., UNI2 [4, 22]) provide powerful pathological representations, while recent generative and fine-grained visual learning studies further highlight the value of semantic visual representation modeling [30, 21]. However, their

⋆ Corresponding author.

high-dimensional features can cause overfitting in limited clinical cohorts. Simple pooling or random compression may also discard rare but prognostically relevant tumor microenvironment (TME) phenotypes, such as invasive margins and heterogeneous regions. Thus, a key challenge is to reduce VFM redundancy while retaining TME phenotypic heterogeneity and feature-space diversity, rather than explicitly reconstructing physical tissue topology.

Existing cross-modal fusion methods also remain limited. Global RNA concatenation methods, such as Pathomic Fusion [5], MCAT [6, 7], and CMTA [9], provide limited pathway-level interpretability, while pathway-aware methods such as PAMT [31] mainly follow a unidirectional paradigm in which genes query image features. Such one-way interaction may underuse the independent WSI context and aggravate modality dominance. In addition, Cox-based survival models rely on proportional hazards assumptions and may be insufficient for modeling non-linear survival dynamics [18].

To address these issues, we propose Di-PACT (*Discrete Prognosis via Aligned Co-Topology of Pathways and Tumor Microenvironment*), where co-topology refers to aligned feature-space organization between TME phenotypes and biological pathways, not explicit two-dimensional tissue adjacency. Di-PACT uses distance-aware stratified sampling and direct-bottleneck encoding as parameter-efficient infrastructure for VFM compression. Its core component is a bidirectional pathway co-attention mechanism, coupled with three-branch residual fusion, to jointly preserve WSI-specific, pathway-specific, and cross-modal prognostic information. This design mitigates modality dominance and provides exploratory multi-level interpretability cues linking prognostic pathways with morphological patterns. Finally, Di-PACT formulates survival prediction as a discrete-time classification task to capture non-linear survival dynamics beyond Cox assumptions.

The main contributions are:

1. We introduce a distance-aware sampling and direct-bottleneck strategy that compresses high-dimensional VFM features while preserving TME phenotypic heterogeneity and feature-space diversity.
2. We propose bidirectional pathway co-attention with three-branch residual fusion as the core mechanism to reduce unidirectional fusion blind spots, mitigate modality dominance, and provide exploratory pathway-morphology interpretability cues.
3. We adopt discrete-time survival modeling and evaluate Di-PACT on three TCGA cohorts and two independent CPTAC cohorts, showing stable prognostic performance and favorable zero-shot cross-domain generalization compared with evaluated multimodal baselines.

2 Methods

2.1 Overall Architecture and Data Preprocessing

As shown in Fig. 1, Di-PACT integrates multimodal features via an end-to-end workflow, comprising: (a) **Heterogeneity-Preserving WSI Encoding**; (b)

Biology-Guided Genomic Encoding; (c) **Bidirectional Residual Synergy and Fusion**; and (d) **Discrete Survival Prediction** [20].

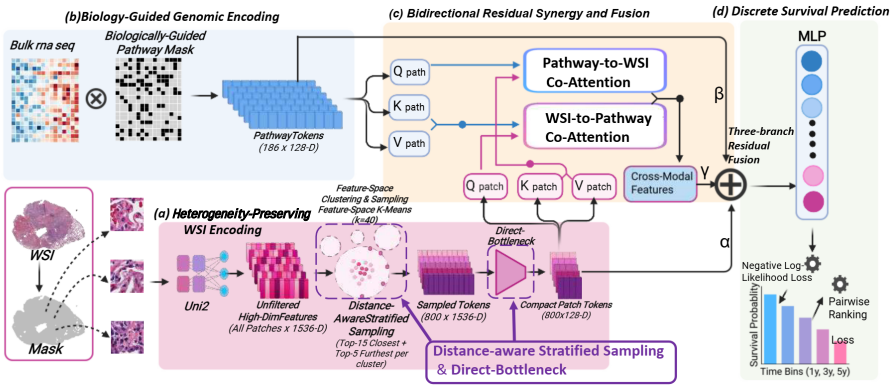

Fig. 1. Overall architecture of Di-PACT, illustrating the end-to-end prognostic workflow from feature-space heterogeneity-aware WSI sampling and biological pathway mapping of genomics to cross-modal bidirectional synergy and residual fusion.

During preprocessing, UNI2 extracts 1,536-dimensional WSI features [4], and RNA-seq profiles are mapped to pathways [16, 27]. To prevent overfitting [29], **Pathway-level Masking** randomly masks $10\% \sim 50\%$ of pathway signals ($p = 0.4$) during training to enforce generalizable global feature learning.

2.2 Heterogeneity-Preserving WSI Encoding

To alleviate VFM-induced overfitting while retaining tumor microenvironment (TME) phenotypic heterogeneity and feature-space diversity [10], we design a dual-compression mechanism. **Distance-aware Stratified Sampling** first clusters UNI2 patch features via K-Means ($K = 40$), selecting the 15 closest and 5 furthest patches per cluster to distill each WSI into $\mathbf{X}_{wsi} \in \mathbb{R}^{N\times 1536}$ ($N = 800$). This center–margin sampling preserves representative phenotypes and rare potentially prognostic patterns in the latent feature space, rather than explicitly reconstructing two-dimensional tissue adjacency. The choice of $K = 40$ and the 15/5 sampling ratio was determined from early validation experiments to balance heterogeneity coverage and computational efficiency. **Direct-Bottleneck Encoding** then linearly compresses $\mathbf{X}_{wsi}$ into $\mathbf{H}_{wsi} \in \mathbb{R}^{N\times d}$ ($d = 128$). Finally, a Positional Encoding Generator (PPEG) and a TransMIL module [28, 25] model global contextual dependencies among sampled tokens:

$$\mathbf{H}_{seq} = \text{TransMIL}(\mathbf{H}_{wsi} + \text{PPEG}(\mathbf{H}_{wsi})) \in \mathbb{R}^{N\times d}. \tag{1}$$

The representation $\mathbf{f}_{wsi} \in \mathbb{R}^d$ is aggregated via attention pooling [23, 13].

2.3 Biology-Guided Genomic Encoding

Instead of flattening raw RNA profiles, a biology-guided sparse linear layer based on KEGG priors [16] decouples gene expressions into P independent pathway tokens. Given a binarized pathway-gene mask matrix $\mathbf{M} \in \{0,1\}^{P\times M}$ (where P is the total number of pathways and M is the number of genes), the k-th decoupled pathway token p_k is calculated as:

$$p_k = (x_{rna} \odot \mathbf{M}_k)\mathbf{W}_k \quad \forall k \in \{1, 2, \ldots, P\}. \tag{2}$$

A Self-Normalizing Neural Network (SNN) [19, 11] further processes these concatenated tokens to generate the pathway sequence matrix $\mathbf{H}_{rna} \in \mathbb{R}^{P\times d}$ and molecular representation $\mathbf{f}_{rna} \in \mathbb{R}^d$, effectively embedding clinical priors into the latent space.

2.4 Bidirectional Residual Synergy and Fusion

This module performs reciprocal pathway–morphology decoding [6, 14]. **Bidirectional Pathway Co-Attention:** The "Pathway-to-WSI" branch uses pathway tokens to query WSI sequences and highlight pathway-relevant morphological patterns; symmetrically, the "WSI-to-Pathway" branch uses WSI context to query pathway tokens and identify associated molecular signals:

$$\mathbf{A}_{p\to w} = \text{Softmax}\left(\frac{(\mathbf{H}_{rna}\mathbf{W}_{q1})(\mathbf{H}_{seq}\mathbf{W}_{k1})^\top}{\sqrt{d}}\right)(\mathbf{H}_{seq}\mathbf{W}_{v1}), \tag{3}$$

$$\mathbf{A}_{w\to p} = \text{Softmax}\left(\frac{(\mathbf{H}_{seq}\mathbf{W}_{q2})(\mathbf{H}_{rna}\mathbf{W}_{k2})^\top}{\sqrt{d}}\right)(\mathbf{H}_{rna}\mathbf{W}_{v2}). \tag{4}$$

The outputs are normalized, concatenated, and projected into a cross-modal feature $\mathbf{f}_{cross} \in \mathbb{R}^d$.

Three-branch Residual Fusion: To reduce modality collapse [29], the final representation preserves WSI-specific, RNA-specific, and cross-modal information [12]:

$$\mathbf{f}_{final} = \alpha \cdot \mathbf{f}_{wsi} + \beta \cdot \mathbf{f}_{rna} + \gamma \cdot \mathbf{f}_{cross}. \tag{5}$$

The weights (α, β, γ) are selected by grid search with step 0.1 on the TCGA validation set. The optimal setting is $\alpha = 0.2$, $\beta = 0.4$, and $\gamma = 0.4$, assigning larger weights to genomic and cross-modal features while retaining a non-zero WSI branch to preserve macroscopic TME heterogeneity. During CPTAC zero-shot testing, these weights are strictly frozen without using target-cohort data. This residual design also provides exploratory multi-level interpretability cues across histological, transcriptomic, and cross-modal levels [24].

2.5 Discrete Survival Prediction

To relax the proportional hazards assumption [18], survival prediction is formulated as a 5-interval discrete-time classification task [20]. Interval cut points are

determined by training-set survival quantiles to balance interval sizes without using validation or test distributions. For a true interval y and event indicator c, the NLL loss is:

$$\mathcal{L}_{nll} = -c \log P(\hat{y} = y) - (1 - c) \log \sum_{j \geq y} P(\hat{y} = j). \tag{6}$$

The final objective integrates a Pairwise Ranking Loss $\mathcal{L}_{rank}$ [26] and auxiliary NLL losses for deep supervision:

$$\mathcal{L}_{total} = \mathcal{L}_{nll}^{final} + \lambda_{rank}\mathcal{L}_{rank} + \alpha_{decay} \sum_{m \in \{wsi, rna, cross\}} w_m \mathcal{L}_{nll}^{m}, \tag{7}$$

where α_{decay} decreases during training, and w_m denotes the loss weight for branch $m \in \{wsi, rna, cross\}$.

3 Experiments and Results

3.1 Datasets and Implementation Details

We evaluated Di-PACT on three internal cohorts from The Cancer Genome Atlas (TCGA) [15]: TCGA-BLCA (372 patients), TCGA-LUAD (404 patients), and TCGA-LUSC (446 patients), using patient-level 5-fold cross-validation. To assess zero-shot cross-domain generalization, we used two independent external cohorts from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) [8]: CPTAC-LUAD (211 patients) and CPTAC-LSCC (208 patients), filtered by complete WSI/RNA-seq/survival intersections. CPTAC WSIs followed the TCGA preprocessing pipeline, including tissue masking, 256×256 patch extraction, and UNI2 feature extraction. RNA-seq data were transformed by $\log_2(x+1)$ and Z-score normalized using only TCGA training-set statistics to avoid target-cohort leakage.

Implementation Details: For the histopathological modality, UNI2 features with 1,536 dimensions were compressed to a 128D latent space via Direct-Bottleneck encoding. The genomic branch processed RNA-seq data into KEGG-based pathway tokens. Survival times were partitioned into 5 non-overlapping intervals using training-set quantiles. Performance was evaluated by C-Index, time-dependent AUC, BS, and IBS.

3.2 Comparisons with Advanced Methods

As shown in Tables 1 and 2, Di-PACT achieves favorable performance across TCGA and CPTAC cohorts with generally smaller standard deviations. In TCGA internal validation, it improves over PAMT [31] by C-Index margins of 0.009, 0.011, and 0.003 on BLCA, LUAD, and LUSC, respectively, and outperforms MCAT [6] by up to 11.9%. Although the LUSC gain is modest, paired t-tests over 5-fold cross-validation support its significance over PAMT ($p = 0.034 < 0.05$).

These results suggest that bidirectional co-attention helps capture complementary prognostic information between TME phenotypes and molecular pathways. In CPTAC zero-shot validation, Di-PACT exceeds PAMT by 0.015 and 0.012 on LUAD and LSCC, indicating favorable cross-domain generalization under external domain shifts.

Table 1. C-Index comparison on internal TCGA cohorts.

Method	**Modality**	**BLCA**	**LUAD**	**LUSC**
SNN [19]	Genomic	0.611±0.035	0.620±0.036	0.602±0.031
Cox-PASNet [11]	Genomic	0.631±0.045	0.601±0.042	0.596±0.043
ABMIL [13]	WSI-only	0.578±0.042	0.572±0.034	0.564±0.037
TransMIL [25]	WSI-only	0.611±0.025	0.602±0.031	0.612±0.021
UNI2-Direct [4]	WSI-only	0.653±0.056	0.626±0.042	0.612±0.047
MCAT [6]	Multimodal	0.628±0.036	0.623±0.032	0.618±0.023
SurvPath [14]	Multimodal	0.624±0.032	0.638±0.041	0.662±0.045
PAMT [31]	Multimodal	0.738±0.036	0.721±0.043	0.707±0.023
Di-PACT (Ours)	**Multimodal**	**0.747±0.026**	**0.732±0.033**	**0.710±0.015**

Table 2. C-Index comparison on external CPTAC cohorts (Zero-shot generalization).

Method	**TransMIL** [25]	**MCAT** [6]	**PAMT** [31]	**Di-PACT (Ours)**
Modality	WSI-only	Multimodal	Multimodal	Multimodal
CPTAC-LUAD	0.537±0.026	0.542±0.043	0.626±0.037	0.641±0.031
CPTAC-LSCC	0.559±0.033	0.548±0.037	0.620±0.031	0.632±0.022

Table 3. Comparison of time-dependent AUC at 1, 3, and 5-year intervals on TCGA cohorts.

Method	**TCGA-BLCA**			**TCGA-LUAD**			**TCGA-LUSC**		
	1-yr	3-yr	5-yr	1-yr	3-yr	5-yr	1-yr	3-yr	5-yr
MCAT [6]	0.778	0.689	0.732	0.698	0.710	0.664	0.664	0.697	0.722
SurvPath [14]	0.769	0.768	0.742	0.724	0.785	0.621	0.645	0.726	0.745
PAMT [31]	0.832	0.846	0.834	0.732	0.867	0.798	0.675	0.789	0.714
Di-PACT (Ours)	**0.866**	**0.854**	**0.852**	**0.802**	**0.864**	**0.806**	**0.720**	**0.792**	**0.784**

As shown in Table 3, Di-PACT also obtains competitive or improved 1-, 3-, and 5-year time-dependent AUCs compared with multimodal baselines, supporting stable prognostic performance across survival stages.

Calibration Evaluation. Using the original discrete survival probabilities, we report Brier Score (BS) and Integrated Brier Score (IBS), where lower values indicate better calibration. Di-PACT achieved the lowest IBS (0.118) among evaluated methods, compared with MCAT (0.311), SurvPath (0.223), and PAMT (0.175), with 1-, 3-, and 5-year BS values of 0.082, 0.125, and 0.158.

Survival Stratification Evaluation. Kaplan-Meier analysis (Fig. 2) [17] and log-rank tests show that Di-PACT significantly stratifies high- and low-risk patients across cohorts, supporting its potential for risk stratification. KM curves visualize survival differences, while log-rank p-values quantify group separation.

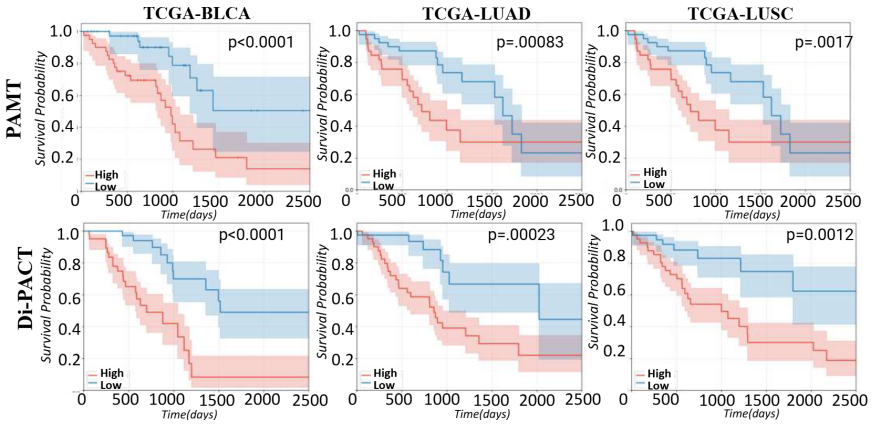

Fig. 2. Kaplan-Meier survival analysis of patients stratified into high-risk and low-risk groups based on the median predicted risk score. The log-rank test confirms statistically significant stratification.

3.3 Ablation Study

We conducted removal ablations on TCGA datasets (Table 4) to assess each module. Removing distance-aware sampling or Direct-Bottleneck encoding caused C-Index drops of up to 3.7% and 2.5%, suggesting their roles in preserving TME phenotypic diversity and reducing VFM-induced overfitting. Removing pathway masking, bidirectional attention, or discrete NLL loss led to drops of up to 1.3%, 3.5%, and 4.5%, supporting the benefits of biological priors, reciprocal cross-modal interaction, and discrete survival modeling [29, 20]. The unidirectional attention variant was also inferior to the full model, indicating that bidirectional co-attention reduces one-way fusion blind spots. Removing residual fusion decreased C-Index by 2.1%–2.9%, showing its contribution to preserving modality-specific and cross-modal information [12].

Table 4. Ablation study of Di-PACT on the TCGA datasets evaluated by C-Index.

Model Variant	TCGA-BLCA	TCGA-LUAD	TCGA-LUSC
Full Model	0.747 ± 0.026	0.732 ± 0.033	0.710 ± 0.015
w/o Dist. Sampling	0.710 ± 0.032	0.712 ± 0.041	0.694 ± 0.012
w/o Bottleneck	0.722 ± 0.045	0.718 ± 0.044	0.685 ± 0.032
w/o Pathway Mask	0.734 ± 0.037	0.723 ± 0.035	0.702 ± 0.024
w/o Bi-Attention	0.712 ± 0.035	0.722 ± 0.041	0.692 ± 0.034
w/ Unidirectional Att.	0.718 ± 0.044	0.707 ± 0.039	0.694 ± 0.036
w/o Discrete Loss	0.702 ± 0.021	0.721 ± 0.037	0.687 ± 0.018
w/o Res. Fusion	0.718 ± 0.024	0.706 ± 0.031	0.689 ± 0.022

3.4 Exploratory Multimodal Interpretability

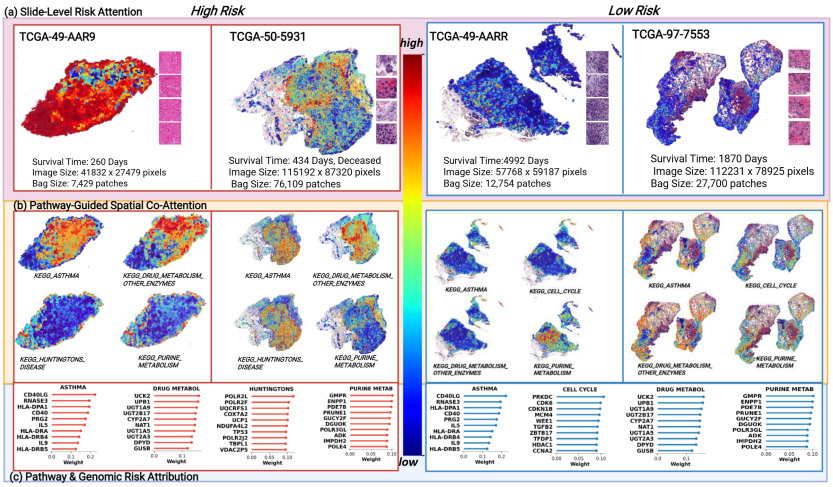

Fig. 3. Exploratory triple visualization for multimodal interpretability: **(a)** slide-level risk attention maps, **(b)** pathway-guided co-attention, and **(c)** gene attribution of key pathways.

Di-PACT provides an exploratory **"Triple Visualization"** scheme across pathology, transcriptomic, and cross-modal levels (Fig. 3) [24]. Slide-level risk heatmaps highlight risk-contributing regions, gene attribution identifies prognosis-related pathways and genes, and pathway-guided co-attention links prognostic pathways with morphological patterns. These views provide multi-level interpretability cues, with high-attribution pathways such as EMT and apoptosis aligning with prior biological knowledge and supporting biomarker hypothesis generation.

4 Conclusion

This study presents **Di-PACT**, a multimodal discrete survival framework for pathology–genomic fusion. By integrating distance-aware sampling, bottleneck encoding, bidirectional co-attention, and residual fusion, Di-PACT reduces VFM redundancy and enhances pathway–morphology interaction under modality imbalance. Experiments on TCGA and CPTAC cohorts show stable prognostic performance and favorable zero-shot generalization. Its exploratory **"Triple Visualization"** scheme provides multi-level cues for risk stratification and biomarker hypothesis generation.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant No. 62471448; in part by the Shandong Provincial Natural Science Foundation under Grant No. ZR2024YQ004; and in part by the Taishan Scholars Youth Expert Program of Shandong Province under Grant No. tsqn202312109.

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

References

1. Acosta, J.N., Falcone, G.J., Rajpurkar, P., Topol, E.J.: Multimodal biomedical ai. Nature medicine **28**(9), 1773–1784 (2022)
2. Boehm, K.M., Khosravi, P., Vanguri, R., Gao, J., Shah, S.P.: Harnessing multimodal data integration to advance precision oncology. Nature Reviews Cancer **22**(2), 114–126 (2022)
3. Cai, Q., Wang, S., Zhang, F., Zhang, C., Liu, Y., Li, H., Dong, J., Zhang, D.: Multimodal artificial intelligence for disease diagnosis: Advances, applications, and challenges. Pattern Recognition p. 113456 (2026)
4. 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)
5. 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)
6. 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)
7. 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)
8. Edwards, N.J., Oberti, M., Thangudu, R.R., Cai, S., McGarvey, P.B., Jacob, S., Madhavan, S., Ketchum, K.A.: The cptac data portal: a resource for cancer proteomics research. Journal of proteome research **14**(6), 2707–2713 (2015)

9. Feng, Z., et al.: Cross-modal translation and alignment for survival analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 21485–21494 (2023)
10. Fridman, W.H., Pagès, F., Sautès-Fridman, C., Galon, J.: The immune contexture in human tumours: impact on clinical outcome. Nature Reviews Cancer **12**(4), 298–306 (2012)
11. Hao, J., Kim, Y., Mallavarapu, T., Oh, J.H., Kang, M.: Cox-pasnet: pathway-based sparse deep neural network for survival analysis. In: 2018 IEEE international conference on bioinformatics and biomedicine (BIBM). pp. 381–386. IEEE (2018)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
14. Jaume, G., Vaidya, A., Chen, R.J., Williamson, D.F., Liang, P.P., Mahmood, F.: Modeling dense multimodal interactions between biological pathways and histology for survival prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11579–11590 (2024)
15. Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., et al.: Mutational landscape and significance across 12 major cancer types. Nature **502**(7471), 333–339 (2013)
16. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M.: Kegg as a reference resource for gene and protein annotation. Nucleic acids research **44**(D1), D457–D462 (2016)
17. Kaplan, E.L., Meier, P.: Nonparametric estimation from incomplete observations. Journal of the American statistical association **53**(282), 457–481 (1958)
18. Katzman, J.L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., Kluger, Y.: Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. BMC medical research methodology **18**(1), 24 (2018)
19. Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S.: Self-normalizing neural networks. Advances in neural information processing systems **30** (2017)
20. Lee, C., Zame, W., Yoon, J., Van Der Schaar, M.: Deephit: A deep learning approach to survival analysis with competing risks. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
21. Li, S., Cai, Q., Zhang, F., Shu, Y., Li, G., Dong, J., Liu, L.: LearnMat: Semantic-Aware self-supervision Fine-Grained visual recognition. IEEE Transactions on Image Processing (2026)
22. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al.: A visual-language foundation model for computational pathology. Nature medicine **30**(3), 863–874 (2024)
23. 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)
24. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence **1**(5), 206–215 (2019)
25. 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)

26. Steck, H., Krishnapuram, B., Dehing-Oberije, C., Lambin, P., Raykar, V.C.: On ranking in survival analysis: Bounds on the concordance index. Advances in neural information processing systems **20** (2007)
27. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al.: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the national academy of sciences **102**(43), 15545–15550 (2005)
28. 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)
29. Wang, W., Tran, D., Feiszli, M.: What makes training multi-modal classification networks hard? In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12695–12705 (2020)
30. Xu, M., Wang, J., Tao, M., Bao, B.K., Xu, C.: CookGALIP: Recipe controllable generative adversarial CLIPs with sequential ingredient prompts for food image generation. IEEE Transactions on Multimedia **27**, 2772–2782 (2025)
31. Yan, R., Zhang, X., Jiang, Z., Wang, B., Bian, X., Ren, F., Zhou, S.K.: 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)