Case-specific Priors-guided Multimodal Fusion for Prognosis Prediction in Head and Neck Cancer

Yue Lin[1,2][0009−0008−7758−4530], Shuchang Ye[3][0009−0006−1935−1953], Shenghai Wu[2][0009−0001−9473−136X], Dashun Zheng[4][0009−0005−7878−978X], and Mingyuan Meng[1][0000−0002−9562−1613]✉

[1] Zhongguancun Academy, Beijing, China
[2] Research Institute, Newland Digital Technology, Fuzhou, China
[3] School of Computer Science, The University of Sydney, Sydney, Australia
[4] Faculty of Applied Sciences, Macao Polytechnic University, Macao SAR, China
mengmingyuan@bza.edu.cn

Abstract. Accurate prognosis prediction for head and neck cancer typically necessitates integrating highly heterogeneous clinical and pathological evidence by sophisticated multimodal fusion. However, existing fusion techniques are usually grounded in case-generic clinical rationales about complementary synergies across modalities, underestimating case-specific variance. In this study, we propose a prognosis prediction framework that explicitly incorporates **C**ase-specific **P**riors to guide **M**ultimodal fusion (named **CPM**). It leverages broad medical knowledge of LLMs to characterize each case's (case-specific) intra-modal contexts and inter-modal relations, as well as case-generic multimodal rationale, as priors, enabling three-stage priors-guided multimodal fusion: (i) Intra-modal Context-guided Calibration (IntraCC) that leverages case-specific intra-modal priors to guide feature calibration within each modality; (ii) Inter-modal Relation-guided Interaction (InterRI) that leverages case-specific inter-modal priors to guide pair-wise feature interaction in modality pairs; and (iii) Multimodal Rationale-guided Aggregation (MultiRA) that leverages case-generic multimodal priors to guide final feature aggregation over all modalities. On the well-established HANCOCK challenge dataset involving four distinct medical modalities, our CPM achieved a mean AUC of 83.98% for predicting 5-year overall survival and 2-year recurrence-free survival, improving upon the baseline by 7.58%, pioneeringly demonstrating the effectiveness of LLM-derived medical priors in guiding multimodal fusion for prognosis prediction. The code is available at https://github.com/yuelin421/CPM.

Keywords: Prognosis · Multimodal Fusion · Case-specific Priors

1 Introduction

Head and neck cancer (HNC) remains a major global health burden, with approximately 890,000 new cases and 450,000 deaths worldwide each year [20]. Because patient outcomes vary substantially and clinical decisions rely on risk

stratification, accurate prognosis prediction is essential for guiding risk-adapted therapy and follow-up management [11]. To achieve this, prognosis prediction for HNC typically necessitates integrating clinical and pathological evidence via multimodal fusion [22]. However, such evidence is inherently heterogeneous across modalities, posing substantial challenges for multimodal fusion due to ambiguous cross-modal correspondence and mismatched representation granularity [3].

Early multimodal fusion techniques for prognosis prediction explore cross-modal correspondence in a data-driven manner via end-to-end deep learning. Concretely, many methods adopt generic fusion operators (e.g., attention-based interaction [6, 13, 17] or bilinear-style interaction [5]) and rely on survival objectives to induce multimodal synergies during training [23]. However, these methods failed to optimize multimodal fusion based on valuable clinical rationale, which exists as medical common sense and can provide explicit insight into cross-modal correspondence. To attain clinically grounded multimodal fusion, recent studies have started to develop specialized multimodal fusion techniques based on modality-specific clinical rationale, for example, by designing a PET/CT fusion encoder around the anatomical-metabolic complementarity [16, 14, 18], or by introducing pathway-structured representations to mitigate the granularity mismatch between gigapixel histology and global omics profiles [10]. Nevertheless, these designs are inherently coupled to specific modality scenarios and, more importantly, they are anchored to case-generic clinical priors, resulting in limited adaptability to case-specific variance in cross-modal correspondence.

The above limitation suggests a potential direction. If we could obtain case-specific prior knowledge about cross-modal correspondence for each case, multimodal fusion could be guided on a per-case basis rather than being constrained by fixed, case-generic interaction patterns in general clinical rationale. Unfortunately, obtaining case-specific priors at scale remains difficult as it requires expert case-by-case annotation that jointly considers multiple modalities to summarize cross-modal correspondence for each patient. Therefore, such priors are rarely available in existing datasets and have not been explored for multimodal fusion. Recently, large language models (LLMs) have made it feasible to derive such priors by automatically analyzing each case's multimodal evidence and generating clinically plausible summaries based on their broad medical knowledge [21, 19]. Several recent studies have incorporated LLM-derived priors or clinical reports into survival prediction [12, 7, 24, 26, 15]. However, they merely exploited LLM to augment the evidence within the modality, where the generated text is treated as either supplementary evidence for a single modality [7, 24, 12, 15], or as a refined version of the textual modality [26]. In contrast, leveraging LLMs to explicitly derive case-specific cross-modal correspondence and further use it to guide multimodal fusion across heterogenous modalities has yet to be explored.

In this study, we propose CPM, a prognosis prediction framework that explicitly incorporates Case-specific Priors to guide Multimodal fusion. CPM leverages the broad medical knowledge of LLMs to derive case-specific priors that summarize intra-modal contexts and inter-modal relations, together with case-generic multimodal rationale, and then uses them to guide multimodal fusion in

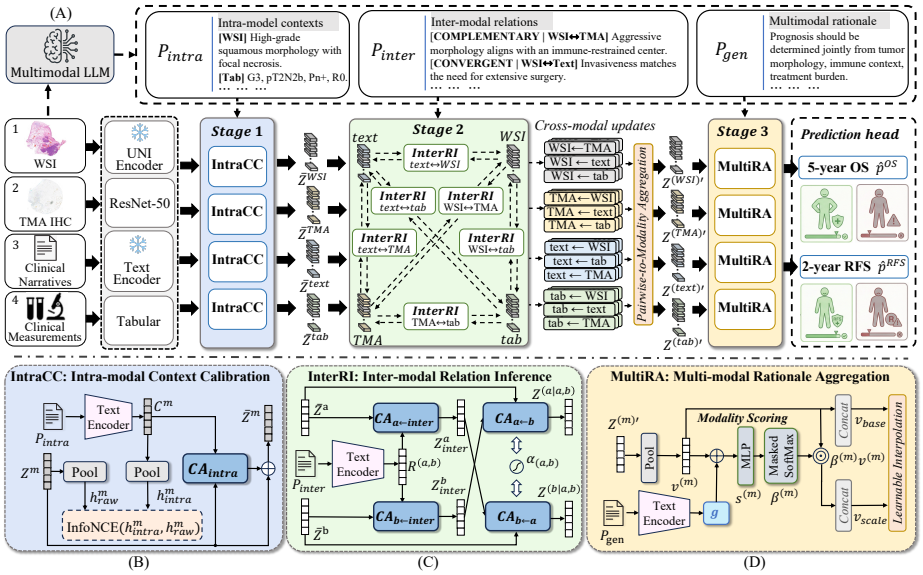

Fig. 1. Overview of CPM. (A) Overall framework. LLM-derived clinically grounded priors guide multimodal fusion through three stages: (B) IntraCC for intra-modal context calibration, (C) InterRI for relation-guided pairwise interaction, and (D) MultiRA for rationale-guided aggregation.

a three-stage manner. Specifically, CPM performs priors-guided multimodal fusion via: (i) Intra-modal Context-guided Calibration (IntraCC), which leverages case-specific intra-modal contexts to calibrate each modality stream and aligns modality representations with their corresponding priors through contrastive learning; (ii) Inter-modal Relation-guided Interaction (InterRI), which leverages case-specific inter-modal relations to guide pair-wise feature interaction by conditioning each modality-pair exchange on the relation evidence derived for the current case; and (iii) Multimodal Rationale-guided Aggregation (MultiRA), which leverages the case-generic multimodal rationale to learn rationale-conditioned importance weights over modality pairs and integrates all pairwise interactions into a unified risk representation for final prognosis prediction.

Our CPM was evaluated on the public HANCOCK challenge dataset [9], a well-established multimodal HNC dataset with four distinct medical modalities, including (i) whole-slide images (WSIs), (ii) tumor-center tissue microarrays (TMAs), (iii) clinical narratives from medical histories and surgery reports, and (iv) structured clinicopathologic and laboratory measurements [8]. Our CPM achieved a mean AUC of 83.98% for predicting 5-year Overall Survival (OS) and 2-year Recurrence-Free Survival (RFS), improving by 7.58% over the baseline methods that were not guided by case-specific priors.

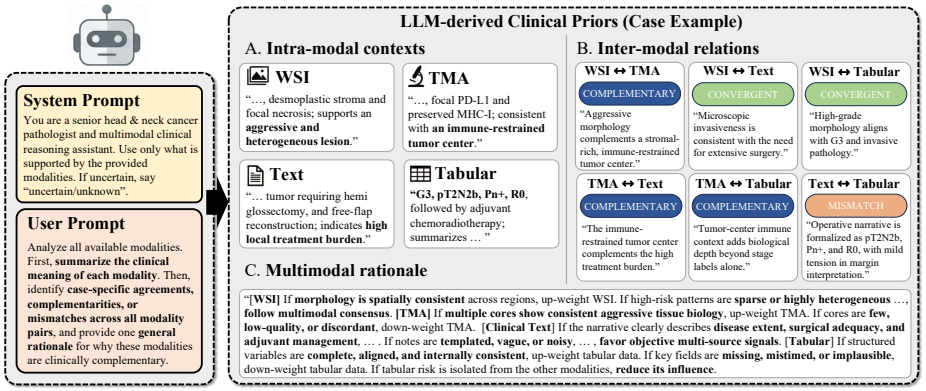

Fig. 2. Example prompt and output of the LLM-derived clinical priors, including intra-modal contexts, inter-modal relations, and a multimodal rationale.

2 Methods

We study multimodal prognosis prediction with M heterogeneous modalities ($M = 4$ in this study). For patient i, the model takes the multimodal evidence $\{x_i^{(m)}\}_{m=1}^{M}$ as input and predicts endpoint probabilities $(\hat{p}_i^{\mathrm{OS}}, \hat{p}_i^{\mathrm{RFS}})$ for OS and RFS, respectively. We optimize the model with a masked multi-task binary cross-entropy (BCE) objective, where each task loss is computed only when the corresponding label is available.

The overall framework of CPM is shown in Fig. 1(A). Given a patient case, we use an LLM to derive clinically grounded priors, as described in Section 2.1, including case-specific intra-modal contexts, case-specific inter-modal relations, and a case-generic multimodal rationale. These priors then guide fusion in a three-stage pipeline, including Intra-modal Context-guided Calibration (IntraCC, Section 2.2), Inter-modal Relation-guided Interaction (InterRI, Section 2.3), and Multimodal Rationale-guided Aggregation (MultiRA, Section 2.4).

2.1 LLM-derived Clinical Priors

CPM uses a multimodal LLM to derive clinical priors that guide multimodal fusion with case-specific evidence. For each patient i, we ask Qwen3-VL-8B-Instruct[2] to conduct pairwise cross-modal assessments over all modality pairs. Each assessment follows the logic of multidisciplinary case review: it first summarizes intra-modal findings and then characterizes case-level inter-modal agreements or conflicts. To ensure evidence grounding and consistent formatting, we require anchor-cited statements, allow explicit "uncertain/unknown" when evidence is insufficient, and restrict outputs to predefined relation types with confidence scores. The final prompt templates, output schema, and LLM generation settings are released with the code to support reproducibility.

From these responses, CPM derives three complementary clinical priors. The case-specific intra-modal contexts $\mathcal{P}_i^{\text{intra}}$ provide modality-level LLM interpretations for the case, where each context describes the clinical and pathologic meaning. The case-specific inter-modal relations $\mathcal{P}_i^{\text{inter}}$ capture pairwise agreements or conflicts to represent cross-modal correspondence within the case. The case-generic multimodal rationale $\mathcal{P}^{\text{gen}}$ summarizes clinical rationales about how different modalities complement each other for prognosis prediction. Fig. 2 shows representative prompts and outputs.

2.2 Intra-modal Context-guided Calibration

IntraCC uses the case-specific intra-modal contexts $\mathcal{P}_i^{\text{intra}}$ to calibrate each modality before cross-modal fusion. As shown in Fig. 1(B), for patient i and modality m, let $Z_i^{(m)}$ denote modality token embeddings extracted from $x_i^{(m)}$. The modality-specific context in $\mathcal{P}_i^{\text{intra}}$ is encoded into context token embeddings $C_i^{(m)}$ using a text encoder. A residual cross-attention block is applied so that modality tokens $Z_i^{(m)}$ can selectively attend to $C_i^{(m)}$. Specifically, cross-attention produces a context-aware update, and residual connection preserves the original modality evidence. As a result, we obtain the calibrated embeddings $\bar{Z}_i^{(m)}$.

To further align raw evidence with the LLM-derived interpretation, an intra-modal contrastive objective is introduced. Pooled embeddings $h_{i,\text{raw}}^{(m)}$ and $h_{i,\text{intra}}^{(m)}$ are obtained from $Z_i^{(m)}$ and $C_i^{(m)}$, respectively. Within each batch, we minimize an InfoNCE loss where same-patient pairs are positives and cross-patient pairs are negatives. This loss is computed only for available modalities and added to the masked multi-task BCE loss during training.

2.3 Inter-modal Relation-guided Interaction

InterRI captures case-specific cross-modal correspondence through pairwise interactions guided by the inter-modal relations $\mathcal{P}_i^{\text{inter}}$. This pairwise design serves as a tractable simplification of full multimodal interaction, since directly characterizing case-specific correspondence across all modalities is highly complex and becomes increasingly difficult as the number of modalities grows, both for modeling and for reliable LLM characterization. Accordingly, InterRI decomposes the full multimodal interaction into a set of relation-guided pairwise exchanges.

As illustrated in Fig. 1(C), for patient i, consider any available modality pair (a, b). The IntraCC-calibrated tokens $\bar{Z}_i^{(a)}$ and $\bar{Z}_i^{(b)}$ are used as inputs and the corresponding relation in $\mathcal{P}_i^{\text{inter}}$ is encoded into relation tokens $R_i^{(a,b)}$. First, each modality is conditioned on $R_i^{(a,b)}$ via cross-attention so that both modalities are guided by the same pair-specific relation description. Next, symmetric cross-attention updates both modalities by incorporating relation-conditioned evidence from the other modality. A learnable pair gate $\alpha_{a,b}$ modulates the interaction strength, and residual connections preserve the IntraCC-calibrated tokens to stabilize fusion.

For each available modality m, residual deltas from all relation-conditioned pairs involving m are averaged to form its final update. As a result, the InterRI-updated token embeddings are denoted as $Z_i^{(m)\prime}$.

2.4 Multimodal Rationale-guided Aggregation

MultiRA aggregates modality representations by using the case-generic clinical rationales in $\mathcal{P}^{\text{gen}}$ to score modalities and scale their representations before fusion. As shown in Fig. 1(D), for patient i and modality m, attentive pooling maps its InterRI-updated tokens $Z_i^{(m)\prime}$ to a modality vector $v_i^{(m)}$. The case-generic clinical rationale in $\mathcal{P}^{\text{gen}}$ are encoded into a shared embedding g. Let $\mathcal{M}_i$ denote the set of available modalities for patient i. Conditioned on g, MultiRA assigns an importance score to each modality and normalizes it across available modalities:

$$s_i^{(m)} = \text{MLP}\Big([v_i^{(m)} \| g]\Big), \qquad \beta_i^{(m)} = \frac{\exp(s_i^{(m)})}{\sum_{k \in \mathcal{M}_i} \exp(s_i^{(k)})}. \tag{1}$$

Unlike weighted-sum fusion that collapses all modalities into a single vector, MultiRA scales each modality vector by $\beta_i^{(m)}$ and then concatenates them to form the fused representation. This keeps a dedicated subspace for each modality while allowing the rationale-guided weights to control how much each modality contributes. To make fusion less sensitive to noisy predicted weights, the scaled concatenation is softly mixed with the unscaled concatenation using a learnable coefficient. The resulting fused vector is then fed to the prediction head to produce $(\hat{p}_i^{\text{OS}}, \hat{p}_i^{\text{RFS}})$.

3 Experimental Setup

3.1 Dataset

CPM was evaluated on the public HANCOCK challenge dataset [8, 9], which comprises 764 HNC patients and provides four distinct medical modalities. WSIs capture fine-grained tumor morphology, and HANCOCK provides pre-extracted UNI patch embeddings [4]. In parallel, TMA cores with immunohistochemistry staining capture immune and tumor phenotypes. These pathology modalities are complemented by clinical narratives that summarize patient history and treatments, together with structured clinicopathologic and laboratory measurements that reflect disease characteristics. All experiments followed the official HANCOCK challenge split, comprising 611 training patients and 152 test patients.

3.2 Implementation Details

Training scheme. CPM was implemented in PyTorch and trained on 2 NVIDIA GeForce RTX 5090 GPUs for 50 epochs with a batch size of 12 using the AdamW

optimizer. We set the learning rate to 1×10^{-5} and used a weight decay of 1×10^{-2}. The objective combined a masked multi-task BCE loss with an InfoNCE loss used in IntraCC, weighted by 0.1 and using a temperature of 0.07. Model selection was performed on the training split via 5-fold cross-validation, where checkpoints were chosen by the best mean validation AUC across folds.

Feature extraction. For WSIs, we directly used the pre-extracted UNI patch embeddings provided by HANCOCK [4, 8]. For TMAs, we resized each core image to 512×512 and encoded it with an ImageNet-pretrained ResNet-50 backbone. For clinical narratives, we concatenated them and encoded with ClinicalBERT [1] using a maximum length of 512, and froze the text encoder during training. For tabular data, we standardized numerical variables by z-score normalization and encoded categorical variables by one-hot encoding to construct tabular features. Finally, we tokenized the LLM-derived clinical priors with a maximum length of 512 and encoded them with the same frozen ClinicalBERT to obtain prior features that guide IntraCC, InterRI, and MultiRA.

3.3 Experiment Designs

Comparisons were designed to cover the two main fusion paradigms discussed in Section 1, including data-driven and clinically grounded fusion. As reference baselines, we reported unimodal results for each modality to quantify standalone performances, and included a simple feature concatenation model as well as a Random Forest classifier to contextualize multimodal gains. We further included recent LLM-assisted prognosis methods for comparison. For all compared methods, we reused the same extracted modality features described above to ensure a controlled comparison. Top-performing methods from HANCOCK challenge were not included since they were not publicly released and the hidden test set is not accessible for direct evaluation.[5]

We also performed stage-wise ablations by removing one stage of CPM at a time to evaluate the contribution of each stage. Statistical significance of pairwise AUC differences between methods was assessed using DeLong's test.

4 Results and Discussion

Table 1 reports the comparison results on HANCOCK. Using the tabular modality alone outperforms several multimodal methods, suggesting that effective fusion across the four distinct modalities is highly challenging. This result indicates that many existing fusion methods fail to fully exploit cross-modal synergy in HANCOCK. Data-driven fusion methods rely mainly on end-to-end supervision to discover multimodal correspondence, but they do not leverage valuable clinical priors. Clinically grounded fusion methods incorporate prior knowledge, yet their designs are tailored to modality-specific scenarios, which limits their adaptability in this setting. LLM-assisted methods mainly use LLM outputs

[5] `https://hancothon25.grand-challenge.org/`

Table 1. Comparison on HANCOCK. Best results are highlighted in bold. * indicates $p < 0.05$ for AUC compared with CPM using DeLong's test.

Method	5-year OS		2-year RFS		Average	
	AUC	Acc	AUC	Acc	AUC	Acc
Reference baselines						
WSI-only	0.7648*	0.6479	0.7049*	0.7593	0.7349	0.7036
TMA-only	0.6573*	0.4444	0.6309*	0.6818	0.6441	0.5631
Text-only	0.6301*	0.5417	0.6935*	0.7364	0.6618	0.6391
Tabular-only	0.7615*	0.6528	0.7782*	0.7545	0.7699	0.7037
Random Forest	0.7078*	0.6389	0.7248*	0.7364	0.7163	0.6877
Concat + MLP	0.7695*	0.7465	0.7580*	**0.7870**	0.7640	0.7668
Data-driven fusion methods						
MCAT [6]	0.6754*	0.6338	0.7795*	0.7315	0.7275	0.6827
MOTCat [25]	0.8070*	0.6901	0.7080*	0.6759	0.7575	0.6830
Clinically grounded fusion methods						
SurvPath [10]	0.8101*	0.7165	0.6942*	0.7407	0.7522	0.7286
LLM-assisted methods						
TMSE [26]	0.7959*	0.6901	0.6737*	0.7407	0.7348	0.7154
LLM-guided MIL [12]	0.7368*	0.6901	0.7852*	0.7200	0.7610	0.7051
CPM (Ours)	**0.8461**	**0.7746**	**0.8335**	**0.7870**	**0.8398**	**0.7808**

Table 2. Ablations on HANCOCK. Best results are highlighted in bold. * indicates $p < 0.05$ for AUC compared with CPM using DeLong's test.

Variant	5-year OS		2-year RFS		Average	
	AUC	Acc	AUC	Acc	AUC	Acc
CPM (Ours)	**0.8461**	**0.7746**	**0.8335**	**0.7870**	**0.8398**	**0.7808**
w/o IntraCC	0.8381*	0.7606	0.8013*	0.7315	0.8197	0.7461
w/o InterRI	0.8094*	0.7324	0.8299*	**0.7870**	0.8196	0.7597
w/o MultiRA	0.8134*	**0.7746**	0.8116*	**0.7870**	0.8125	**0.7808**

as supplementary evidence for individual modalities, but they do not explicitly exploit case-specific cross-modal correspondence for multimodal fusion.

By contrast, CPM explicitly derives case-specific priors to characterize intra-modal contexts and inter-modal relations, together with a case-generic multimodal rationale, and uses them to guide multimodal fusion through IntraCC, InterRI, and MultiRA. This design enables CPM to more effectively uncover complementary synergies across the four HANCOCK modalities and thereby achieve the best overall performance, surpassing both unimodal settings and existing multimodal methods.

Table 2 reports the ablation results of CPM. Removing IntraCC, InterRI, or MultiRA consistently reduces the overall AUC, confirming that each stage contributes to the effectiveness of CPM. Even after removing a single stage, the results remain competitive with representative methods in Table 1, supporting the value of case-specific LLM-derived priors for guiding multimodal fusion.

5 Conclusion

We presented CPM, a case-specific priors-guided multimodal prognosis prediction framework that uses LLM-derived case-specific priors to guide multimodal fusion for HNC. By modeling intra-modal contexts, inter-modal relations, and multimodal rationales, CPM performs priors-guided multimodal fusion through IntraCC, InterRI, and MultiRA. Experimental results on HANCOCK demonstrated that CPM consistently outperformed representative baselines and competing methods, highlighting the effectiveness of explicitly modeling case-specific cross-modal correspondence for multimodal prognosis prediction.

Acknowledgments. This work was supported by the Zhongguancun Academy (Grant No. XTS0071).

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Alsentzer, E., Murphy, J., Boag, W., Weng, W.H., Jindi, D., Naumann, T., McDermott, M.: Publicly available clinical bert embeddings. In: Proceedings of the 2nd clinical natural language processing workshop. pp. 72–78 (2019)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. 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)
4. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature medicine **30**(3), 850–862 (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. Cui, J., Wen, L., Fei, Y., Liu, B., Zhou, L., Shen, D., Wang, Y.: Hila: Hierarchical vision-language collaboration for cancer survival prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 240–250. Springer (2025)
8. Dörrich, M., Balk, M., Heusinger, T., Beyer, S., Mirbagheri, H., Fischer, D.J., Kanso, H., Matek, C., Hartmann, A., Iro, H., et al.: A multimodal dataset for precision oncology in head and neck cancer. Nature Communications **16**(1), 7163 (2025)
9. HANCOTHON: Hancock dataset (2025), `https://hancothon25.grand-challenge.org/hancock-dataset/`, retrieved February 27, 2026

10. 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)
11. Johnson, D.E., Burtness, B., Leemans, C.R., Lui, V.W.Y., Bauman, J.E., Grandis, J.R.: Head and neck squamous cell carcinoma. Nature reviews Disease primers **6**(1), 92 (2020)
12. Kim, K., Lee, Y., Park, D., Eo, T., Youn, D., Lee, H., Hwang, D.: Llm-guided multi-modal multiple instance learning for 5-year overall survival prediction of lung cancer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 239–249. Springer (2024)
13. Li, Z., Jiang, Y., Lu, M., Li, R., Xia, Y.: Survival prediction via hierarchical multimodal co-attention transformer: A computational histology-radiology solution. IEEE Transactions on Medical Imaging **42**(9), 2678–2689 (2023)
14. Lin, Y., Wu, S., Wang, H., Bi, L., Meng, M.: Multi-stage multimodal progressive learning for coordinated segmentation, diagnosis, and prognosis in head and neck cancer. In: Fourth Head and Neck Cancer Tumor Lesion Segmentation, Diagnosis and Prognosis (2025)
15. Liu, P., Ji, L., Gou, J., Fu, B., Ye, M.: Interpretable vision-language survival analysis with ordinal inductive bias for computational pathology. In: International Conference on Learning Representations (2025)
16. Meng, M., Bi, L., Fulham, M., Feng, D., Kim, J.: Merging-diverging hybrid transformer networks for survival prediction in head and neck cancer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 400–410. Springer (2023)
17. Meng, M., Gu, B., Fulham, M., Song, S., Feng, D., Bi, L., Kim, J.: Adaptive segmentation-to-survival learning for survival prediction from multi-modality medical images. npj Precision Oncology **8**(1), 232 (2024)
18. Saeed, N., Hassan, S., Hardan, S., Cai, L., Liang, X., Mazher, M., Qayyum, A., Bu, Y., Lyu, M., Lin, Y., et al.: Head and neck tumor (hecktor) 2025: Benchmark of segmentation, diagnosis, and prognosis in multimodal pet/ct. arXiv preprint arXiv:2606.20143 (2026)
19. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., Hou, L., Clark, K., Pfohl, S.R., Cole-Lewis, H., et al.: Toward expert-level medical question answering with large language models. Nature medicine **31**(3), 943–950 (2025)
20. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **71**(3), 209–249 (2021)
21. Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.S.W.: Large language models in medicine. Nature medicine **29**(8), 1930–1940 (2023)
22. Tian, R., Hou, F., Zhang, H., Yu, G., Yang, P., Li, J., Yuan, T., Chen, X., Chen, Y., Hao, Y., et al.: Multimodal fusion model for prognostic prediction and radiotherapy response assessment in head and neck squamous cell carcinoma. npj Digital Medicine **8**(1), 302 (2025)
23. Unger, M., Kather, J.N.: A systematic analysis of deep learning in genomics and histopathology for precision oncology. BMC Medical Genomics **17**(1), 48 (2024)
24. Wang, Z., Liu, H., Wang, Z., Li, D., Cen, M., Magnier, B., Liang, L., Wang, L.: Enhancing wsi-based survival analysis with report-auxiliary self-distillation. In:

International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 197–207. Springer (2025)
25. Xu, Y., Chen, H.: Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 21241–21251 (2023)
26. Zhang, R., Fang, M., Liu, S., Wang, Z., Tian, J., Dong, D.: Tmse: Tri-modal survival estimation with context-aware tissue prototype and attention-entropy interaction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 640–650. Springer (2025)