Supplementary Materials: Factorized Context Aggregation for Robust Cancer Risk Estimation via Soft Re-Ranked Retrieval and Hierarchical Anchors

Puria Azadi Moghadam
puria.azadi@ubc.ca

Ali Khajegili Mirabadi
ali.mirabadi@ubc.ca

Behnam Maneshgar
behnam.maneshgar@ubc.ca

Hossein Farahani
University of British Columbia
h.farahani@ubc.ca

Ali Bashashati
University of British Columbia
ali.bashashati@ubc.ca

1. Biological Motivation for WSI-Space Retrieval

Using WSIs as the retrieval anchor is biologically motivated. We perform retrieval in WSI embedding space because histology similarity often correlates with both report content and underlying genomic status. Prior histo–genomic studies have shown that specific mutations and molecular alterations leave recognizable morphological signatures [5, 13, 17, 18]. Well-known examples include IDH1 and IDH2 mutations in glioma, which visibly alter cellular structure and can guide pathologists even in the absence of molecular tests. With the advance of deep learning in microscopic image analysis, multiple works have demonstrated that gene mutation status and molecular subtypes can be predicted from histopathology alone. For example, Pizurica et al. [13] showed that gene status can be directly inferred from tissue morphology in different cancer types, and Coudray et al. [5] found that mutations in STK11, EGFR, FAT1, SETBP1, KRAS, and TP53 are predictable from lung cancer pathology images. Moreover, pathology reports are largely driven by pathologists' visual assessment of the slide, combined with other relevant patient information (e.g., additional tests and clinical observations), so similar slides often lead to similar reports. Taken together, these findings suggest that patients with similar slide-level embeddings are likely to share not only visual patterns but also correlated genomic and textual characteristics, making WSI-space retrieval a reasonable proxy for multimodal similarity. While not universal, using histology as the retrieval anchor is therefore a biologically grounded approximation for many cancer types.

2. Foundation Models Details

We employ three modality-specific FMs to extract dense representations that feed into our retrieval and aggregation pipeline. FM models for histopathology slides and gene expression data were developed and pretrained in-house. For pathology reports, we utilized OpenBioLLaMA-7B [1].

- **WSI:** Our in-house WSI FM is based on a large multiple instance learning architecture, inspired by AB-MIL [10]. Following 3-fold cross-validation, the model was trained to classify the primary tissue site across 33 cancer types using slide-level weak supervision. Additionally, we integrate patch-level embeddings extracted from UNI [4], a vision-language FM trained on histopathology patches.
- **Gene Expression (Bulk RNA):** To pretrain our foundation model, we follow the BulkRNABert approach [8], adopting a self-supervised learning framework based on masked language modeling [6]. In this setup, a fixed proportion of gene expression values are randomly masked, shuffled, or left untouched, and the model is trained to reconstruct the masked values. This strategy enables the model to learn contextual dependencies and robust gene-level representations from unlabeled data. Pretraining is conducted on the GTEx dataset [2], which comprises a diverse collection of tissues, conditions, and biological states. A Transformer-based architecture is then trained to minimize the reconstruction loss at the masked positions, enabling the model to capture biologically meaningful transcriptomic patterns.
- **Pathology Reports:** For representing textual pathology reports, we utilize OpenBioLLaMA-7B [1], a biomedical large language model pretrained on diverse medical corpora. The model is used to encode pathology text reports into semantic embeddings.

3. Multimodal Teacher and Multimodal Baseline in More Detail

A central advantage of our framework is that it can be paired with any multimodal teacher model, since it only requires modality-specific and unified multimodal representations for the anchoring process. However, due to limitations in

Table 1. Dataset Summary: Full names, organ origins, and number of patients per cancer type.

Dataset (Acronym)	Origin	# Patients
Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (CESC)	Cervix	241
Colon Adenocarcinoma (COAD)	Colon	376
Head and Neck Squamous Cell Carcinoma (HNSC)	Head & Neck	431
Brain Lower Grade Glioma (LGG)	Brain	442
Liver Hepatocellular Carcinoma (LIHC)	Liver	321
Lung Adenocarcinoma (LUAD)	Lung	430
Lung Squamous Cell Carcinoma (LUSC)	Lung	424
Sarcoma (SARC)	Soft Tissue	236

publicly available multimodal foundation model that jointly incorporates WSI, pathology reports, and gene-expression data for cancer risk estimation, and because public ones may introduce data-leakage concerns with respect to our cohort splits, we construct the teacher model described in the main text using the same foundational components employed in our FM-enabled Multimodal Vectorized Memory. For each modality $j \in \{1, \ldots, M\}$, we extract a modality-specific embedding

$$R^{(j)}_{\text{Teacher},n} = f^j\left(C^{(j)}_n\right),$$

where f^j denotes the corresponding pretrained FM encoder. The modality representations $R^{(j)}_{\text{Teacher},n}$are then passed through learnable projection heads, creating $R^{'(j)}_{\text{Teacher},n}$. The projected representations are concatenated and processed through a multimodal head as follows:

$$R_{\text{Teacher},n} = MM_{head}\left(R^{'(1)}_{\text{Teacher},n}, \ldots, R^{'(M)}_{\text{Teacher},n}\right), \quad (1)$$

where MM_{head} is a lightweight fusion head to aggregate the features from different modalities. We utilized concatenation alongside multiple MLP layers. This design prioritizes simplicity and transferability across datasets, while preventing overfitting to the limited subset of cases where all modalities are jointly available. More expressive fusion architectures, such as multimodal transformers, could be adopted. However, we find that MLP-based fusion is sufficient for providing reliable supervisory signals to the student model.

Because the teacher requires complete multimodal input, it is trained solely on the subset of patients for which all complementary modalities are present. This constraint is inherent to any multimodal teacher but aligns naturally with our setting: the teacher is used only during training, while the student model remains fully operational at inference time, even when other modalities are missing. Additional ablations analyzing the effect of alternative multimodal teachers and different modality-specific encoders are provided in Section 6.

4. Survival-Specific Loss

To estimate the risk, we use the negative partial log-likelihood loss as follows:

$$\mathcal{L}_{\text{cox}} = -\sum_{n:E_n=1}\left[\hat{r}_n - \log\left(\sum_{j\in\mathcal{R}(T_n)} e^{\hat{r}_j}\right)\right], \quad (2)$$

where $\hat{r}_n$ is the predicted risk score for patient n, and $\mathcal{R}(T_n) = \{j \mid T_j \geq T_n\}$ denotes the risk set of patients still at risk at time T_n. This loss encourages higher risk scores for individuals who experience events earlier.

5. Datasets

We conduct our experiments on eight cancer types from The Cancer Genome Atlas (TCGA), spanning diverse organs and tissue origins: CESC, COAD, HNSC, LGG, LIHC, LUAD, LUSC, and SARC. These datasets cover a broad spectrum of cancer phenotypes, contributing to the robustness and generalizability of our survival prediction framework. For each cancer type, we included patients who had at least one histopathology slide, while an associated pathology report and matched bulk RNA-seq gene expression data were available. In addition, overall survival information was required to be available for inclusion. Each WSI is a high-resolution image of size approximately 100,000 $\times$ 100,000 pixels. We extract all tissue-containing patches of a crop size 224 $\times$ 224 at 20$\times$ magnification. We utilized 3-fold cross-validation for all experiments to avoid any data leakage. We applied Principal Component Analysis (PCA), fitted on the training set, to project both gene expression and pathology report embeddings into a 128-dimensional space. Before pretraining our transcriptomic foundation model on the external GTEx dataset [2], expression profiles are first normalized using transcripts per million (i.e., TPM) and subsequently discretized into expression bins to reduce noise and promote generalization. In the experiments for the Generalizability and Robustness to Different Missing Rates section, the CESC and SARC datasets were excluded due to their limited sample sizes, which could result in the number of available gene expression profiles or pathology reports falling below the dimensionality required for PCA. Clinical metadata associated with each patient were retrieved from cBioPortal. Table 1 contains the total number of patients and organ origin for each survival prediction cancer dataset

Dataset	Mode	Histo@train&test	Ours (orig.)	SurvPath	Ours (with SurvPath as Teacher)	MCAT	Ours (with MCAT as Teacher)
COAD	H+G†	0.613 ± 0.121	0.673 ± 0.105	0.733 ± 0.073	0.754 ± 0.213	0.667 ± 0.107	0.730 ± 0.197
HNSC	H+G†	0.561 ± 0.076	0.600 ± 0.058	0.564 ± 0.090	0.596 ± 0.059	0.578 ± 0.074	0.583 ± 0.069

Table 2. C-index (↑) comparison for histopathology-only training and different teacher models.

6. Ablation on the Histo-Gene Specific Teacher Models

Our framework is compatible with arbitrary teacher models. To assess this, we perform an ablation with two histopathology–transcriptomic multimodal teachers, SurvPath [11] and MCAT [3]. We focus on the COAD and HNSC cohorts, where Jaume et al. [11] released training configurations for both SurvPath and MCAT, enabling fair comparison. As shown in Table 2, our distilled models achieve C-index values that are close to those of their multimodal teachers, despite not having access to gene expression at inference time and despite SurvPath and MCAT being trained end-to-end on real omic features rather than frozen foundation-model embeddings. At the same time, all of our variants substantially outperform the histology-only baseline, confirming that our distillation strategy consistently boosts performance over uni-modal models.

We also observe that the performance trend of our approach is stable across teachers. All versions of our model obtain similar C-index values on both COAD and HNSC, indicating that the framework does not rely on a specific teacher architecture. Finally, the results suggest that stronger teachers yield stronger model for handling missing modalities. On COAD, where SurvPath achieves the best multimodal performance, the corresponding distilled variant of our method also attains the highest C-index among our teacher choices, supporting the intuition that our framework can effectively inherit the strengths of a well-performing teacher.

7. Comparison Against Memory Bank

We further compare against the recent Memory-Bank of the M2Surv framework [14], showing our method outperforms it under identical missing-modality settings (Table 3).

Model	Mode	CESC	COAD	HNSC	LGG	LIHC	LUAD	LUSC	SARC
M2Surv	H+G†	0.542	0.641	0.562	**0.717**	0.548	**0.570**	0.555	0.543
Ours	H+G†	**0.615**	**0.673**	**0.600**	0.690	**0.625**	0.567	**0.604**	**0.559**
M2Surv	H+R†	0.561	0.633	0.554	0.704	0.495	**0.599**	0.556	0.516
Ours	H+R†	**0.585**	**0.682**	**0.584**	**0.733**	**0.614**	0.569	**0.580**	**0.557**
M2Surv	H+G†+R†	0.608	0.627	0.569	0.716	0.565	0.562	0.597	0.503
Ours	H+G†+R†	**0.625**	**0.674**	**0.582**	**0.728**	**0.619**	**0.568**	**0.612**	**0.556**

Table 3. Model performance compared to a memory bank-based framework

8. Computational Comparison

Tables 4 and 5 highlight the substantial computational advantages of our model compared to existing multimodal and missing-modality handling approaches. Despite relying solely on histopathology information, our method uses orders of magnitude fewer parameters than large multimodal models such as MCAT and SurvPath as Histo-Gene methods, and the original Multimodal (Teacher) framework, which integrates both gene-expression and text foundation models with a large number of parameters. Remarkably, even with this drastically reduced model size, our approach achieves risk-prediction and patient-stratification performance that is competitive with these far heavier multimodal baselines. Furthermore, when compared to other missing-modality techniques (e.g., CycleR, LDVAE, and AcMAE), our model remains significantly more efficient, requiring far fewer parameters and FLOPs while avoiding the overfitting and architectural overhead introduced by these methods. Notably, recent foundation model–based methods like LDVAE and AcMAE exhibit parameter counts and computational FLOPs many times larger than ours, while they do not provide commensurate gains in predictive accuracy. These results collectively demonstrate that our method not only can eliminate the dependency on expensive multimodal foundation models but also offers a computationally lightweight and practically deployable alternative that is especially well-suited for low-resource clinical environments.

Model	Head Model	Histo FM		+ Gene FM	+ Text FM
		UNI	AB-MIL		
Histo@train&test	66.6K	303M	266K	–	–
Shaspec	364K	303M	266K	–	–
EgoKD	134K	303M	266K	–	–
CrossKD	66K	303M	266K	–	–
CycleR	859K	303M	266K	–	–
AcMAE	8.69M	303M	266K	–	–
LDVAE	2.65M	303M	266K	–	–
Ours	759K	303M	266K	–	–
MCAT	4.17M	303M	–	–	–
SurvPath	92.9M	303M	–	–	–
Multimodal@train&test	199K	303M	266K	7M	8B

Table 4. Parameter comparison of all models with UNI (303M) and DeepMIL (266K) grouped under Histo FM.

Head Model	MACs	FLOPs
Histo@train&test	33.5K	67K
Shaspec	297K	594K
EgoKD	133K	266K
CrossKD	33.5K	67K
CycleR	1.525M	3.04M
AcMAE	30.85M	61.70M
LDVAE	70.81M	141.62M
Ours	1.79M	3.58M
MCAT	409.82M	819.64M
SurvPath	1.84B	3.68B
Multimodal@train&test	199K	398K

Table 5. Head-level computational cost comparison.

9. Kaplan-Meier Curves for Survival Stratification

In Figures 2 to 9, we present Kaplan-Meier survival curves to evaluate patient stratification performance under for three methods: baseline model trained and tested with histopathology only (top rows), models with access to complete multi-modal data (middle rows), and our method using only histopathology at inference time while approximating auxiliary modalities (bottom rows). As shown in Figure 4, our approach consistently achieves stratification performance comparable to fully multimodal models across all datasets. These results highlight that the modality proxy estimations produced by our model are strongly associated with survival outcomes and can effectively support clinical risk prediction. Notably, our method achieves significant separation in 5 cases for H+G†, 5 settings for H+R†, and 6 cases for H+G†+R†, totaling 16 out of 24 (67%) evaluated scenarios, surpassing the best SOTA model, which achieved significance in only 12 out of 24 cases (50% success rate).

10. Semantic Alignment

To assess semantic preservation without pathology reports (**H+R**†), we compute the Boltzmann Semantic Score (BSS)[12] using Gemma-7B language embeddings[16] on CESC, HNSC, LGG, LIHC, LUAD, LUSC, and SARC. Our model achieves a BSS of 0.3549, outperforming baselines by $\sim$ 5–25% relative improvement, indicating superior retrieval of clinically meaningful semantics from histology alone.

11. Full Results on Generalizability and Robustness Against Different Missing Rates

Table 6 reports performance under varying training-time missing rates of the auxiliary modality, with the test set fully missing. Even with 40% missing, our model surpasses the strongest SOTA compensation method trained with complete modalities, indicating its robustness to missing data during both training and inference.

12. Ablation on an Alternative Histopathology Foundation model

To demonstrate the generalizability of our approach, we evaluate its effectiveness across another histopathology framework. While the main results presented in the paper (Tables 1 and 2 in the main text) are based on AB-MIL [10], we additionally train and evaluate our model, along with all baselines, using the framework proposed by Shao et al. [15] (TransMIL). Our model achieves the highest average C-index of 0.551, compared to 0.529 for the histopathology-only model and 0.598 for the multimodal model. Moreover, it outperforms all baselines, the best of which is AcMAE with an average C-index of 0.543, demonstrating its robustness and consistent performance gains across backbone architectures, and highlighting its effectiveness in enhancing cancer survival prediction.

13. Limitations and Future Directions

Although our method performs strongly overall, this framework appears to benefit most in cancers with stronger genotype- or clinical-morphological coupling, where retrieved multimodal context is more likely to reflect the true signal. By contrast, it seems that highly heterogeneous cancers may be more challenging, as weak or inconsistent correspondence across morphology, molecular profiles, and reports can hinder reliable retrieval and proxy estimation. In such settings, the retrieved auxiliary context may only partially capture the true patient-specific biology. Future work should therefore investigate cancer-specific retrieval strategies, more uncertainty-aware proxy estimation, and broader validation on increasingly heterogeneous cohorts.

14. Implementation Details

Table 8 includes the complete training settings of models. All experiments are conducted on a Slurm-managed GPU cluster. The gene-expression foundation model is trained using four NVIDIA RTX 3090 or A6000 GPUs with a batch size of 64 per GPU for 900 epochs, with Flash-Attention enabled to improve GPU memory efficiency. The histopathology foundation model is trained for 100 epochs on a single RTX 3090 or A6000 GPU. All code is implemented in PyTorch, and supports both CPU and GPU training. For downstream survival prediction, we use a learning rate of 1×10^{-3}. We set $k = 17$ for gene-specific experiments and $k = 10$ for report-specific experiments, following the ablation studies reported in the main paper. All survival models are trained for 100 epochs with the Adam optimizer (betas $= 0.9, 0.999$) and a maximum batch size of 512. C-index evaluations are performed using the `torchsurv` li-

Table 6. Performance of our model (average of c-indices) under varying missing rates compared with the best SOTA trained with the full dataset

Dataset	Mode	Best SOTA	Missing rate 0%	Missing rate 10%	Missing rate 20%	Missing rate 30%	Missing rate 40%
COAD	Gene	0.596	0.673	0.669	0.656	0.685	0.650
COAD	Report	0.591	0.682	0.681	0.649	0.645	0.653
COAD	Gene+Report	0.623	0.674	0.687	0.665	0.690	0.667
HNSC	Gene	0.564	0.600	0.567	0.585	0.581	0.574
HNSC	Report	0.584	0.584	0.579	0.575	0.567	0.563
HNSC	Gene+Report	0.568	0.582	0.578	0.577	0.587	0.565
LGG	Gene	0.713	0.690	0.707	0.711	0.709	0.692
LGG	Report	0.705	0.733	0.724	0.698	0.704	0.720
LGG	Gene+Report	0.716	0.728	0.721	0.715	0.701	0.700
LIHC	Gene	0.579	0.625	0.604	0.590	0.583	0.594
LIHC	Report	0.561	0.614	0.616	0.618	0.622	0.618
LIHC	Gene+Report	0.617	0.619	0.624	0.630	0.633	0.629
LUAD	Gene	0.575	0.567	0.575	0.564	0.554	0.554
LUAD	Report	0.565	0.569	0.564	0.537	0.553	0.544
LUAD	Gene+Report	0.548	0.568	0.564	0.531	0.555	0.538
LUSC	Gene	0.587	0.604	0.602	0.613	0.611	0.606
LUSC	Report	0.578	0.580	0.582	0.586	0.579	0.602
LUSC	Gene+Report	0.579	0.612	0.625	0.602	0.595	0.582
Average	Gene	0.602	0.626	0.621	0.620	0.621	0.612
Average	Report	0.597	0.627	0.624	0.611	0.611	0.617
Average	Gene+Report	0.609	0.630	0.633	0.620	0.627	0.613
Average	Overall	0.603	0.628	0.626	0.617	0.620	0.614

Table 7. Top-5 Boltzmann Semantic Score for H+R† settings using Gemma-7B as the reference LLM

Model	Boltzmann Semantic Score
CycleR (TMI'20)	0.3378
EgoKD(ICCV'23)	0.3316
CrossKD (ISBI'23)	0.3342
AcMAE (AAAI'23)	0.2830
Shaspec (CVPR'23)	0.3309
LDVAE (CVPR'25)	0.3092
Ours	<u>0.3549</u>

Table 8. Summary of training and optimization hyperparameters.

Parameter	Value
Batch size (Head models)	512
Batch size per GPU (Gene FM)	64
Batch size per GPU (WSI FM)	1
Optimizer	Adam
Adam β_1, β_2	0.9, 0.999
Learning rate	1e$-$3
K (gene expression)	17
K (text reports)	10
$\lambda_{\text{inter}}, \lambda_{\text{intra}}$	3.0, 3.0
α (scaling factor)	0.1
C-index method	Harrell's estimator (via `torchsurv`)
Ties handling (survival loss)	Efron's method
Epochs (main models and Histopathology FM)	100
Epochs (gene expression FM)	900
GPU	4 $\times$ RTX 3090, A6000
Attention optimization (gene expression FM)	Flash attention (enabled)
Framework	PyTorch
Seeds	3 (1001, 1002, 1003)
Folds	3
Total runs	3 $\times$ 3 = 9 runs/model

brary, which implements Harrell et al.'s nonparametric concordance index estimator [9]. Efron's method is used within the negative partial log-likelihood survival loss [7] to handle ties in event times. We use $\lambda_{\text{inter}} = \lambda_{\text{intra}} = 3.0$ to control the contributions of the inter- and intra-modality anchor losses, and soft-ranks scaling factor is set to 0.1. Our model and all baseline models are trained under identical settings and on the same embeddings to ensure fair comparison. All models (ours and baselines) are trained using 3-fold cross-validation with 3 random initialization seeds, resulting in 9 training runs per experimental setting.

15. Qualitative Analysis: t-SNE Visualization of Representations

To qualitatively assess the learned representations, we extract the final-layer embeddings of each patient in a fold and visualize them using t-SNE. We compare the baseline model trained only on histopathology, the fully multimodal model, and our method under the H + G† + R† setting. As shown in Figure 1, the t-SNE plots, especially for CESC, HNSC, LUAD, LUSC, and SARC datasets, reveal that the

baseline model fails to separate low- and high-risk patients into distinct clusters. In contrast, the multimodal model yields well-separated clusters, validating the utility of incorporating auxiliary modalities. Notably, our method, despite relying only on histopathology at inference, achieves comparable clustering of risk groups. It suggests that our proxy representations effectively encode survival-relevant distinctions and exhibit stronger semantic separation.

References

[1] Malaikannan Sankarasubbu Ankit Pal. Openbiollms: Advancing open-source large language models for healthcare and life sciences. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B, 2024. 1

[2] Latarsha J Carithers and Helen M Moore. The genotype-tissue expression (gtex) project. *Biopreservation and biobanking*, 13(5):307, 2015. 1, 2

[3] 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. 3

[4] 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. 1

[5] Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature medicine*, 24(10):1559–1567, 2018. 1

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 1

[7] Bradley Efron. The efficiency of cox's likelihood function for censored data. *Journal of the American statistical Association*, 72(359):557–565, 1977. 5

[8] Maxence Gélard, Guillaume Richard, Thomas Pierrot, and Paul-Henry Cournède. Bulkrnabert: Cancer prognosis from bulk rna-seq based language models. *bioRxiv*, pages 2024–06, 2024. 1

[9] Frank E Harrell Jr, Kerry L Lee, and Daniel B Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387, 1996. 5

[10] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 1, 4

[11] 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. 3

[12] Ali Khajegili Mirabadi, Katherine Rich, Hossein Farahani, and Ali Bashashati. Boltzmann semantic score: A semantic metric for evaluating large vision models using large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. 4

[13] Marija Pizurica, Yuanning Zheng, Francisco Carrillo-Perez, Humaira Noor, Wei Yao, Christian Wohlfart, Antoaneta Vladimirova, Kathleen Marchal, and Olivier Gevaert. Digital profiling of gene expression from histology images with linearized attention. *Nature Communications*, 15(1):9886, 2024. 1

[14] Mingcheng Qu, Guang Yang, Donglin Di, Yue Gao, Tonghua Su, Yang Song, and Lei Fan. Memory-augmented incomplete multimodal survival prediction via cross-slide and gene-attentive hypergraph learning. *arXiv preprint arXiv:2506.19324*, 2025. 3

[15] 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. 4

[16] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024. 4

[17] Bo-Han Wei, Xavier Cheng-Hong Tsai, Kuo-Jui Sun, Min-Yen Lo, Sheng-Yu Hung, Wen-Chien Chou, Hwei-Fang Tien, Hsin-An Hou, and Chien-Yu Chen. Annotation-free deep learning for predicting gene mutations from whole slide images of acute myeloid leukemia. *NPJ precision oncology*, 9(1):35, 2025. 1

[18] Yu Zhao, Shan Xiong, Qin Ren, Jun Wang, Min Li, Lin Yang, Di Wu, Kejing Tang, Xiaojie Pan, Fengxia Chen, et al. Deep learning using histological images for gene mutation prediction in lung cancer: a multicentre retrospective study. *The Lancet Oncology*, 26(1):136–146, 2025. 1

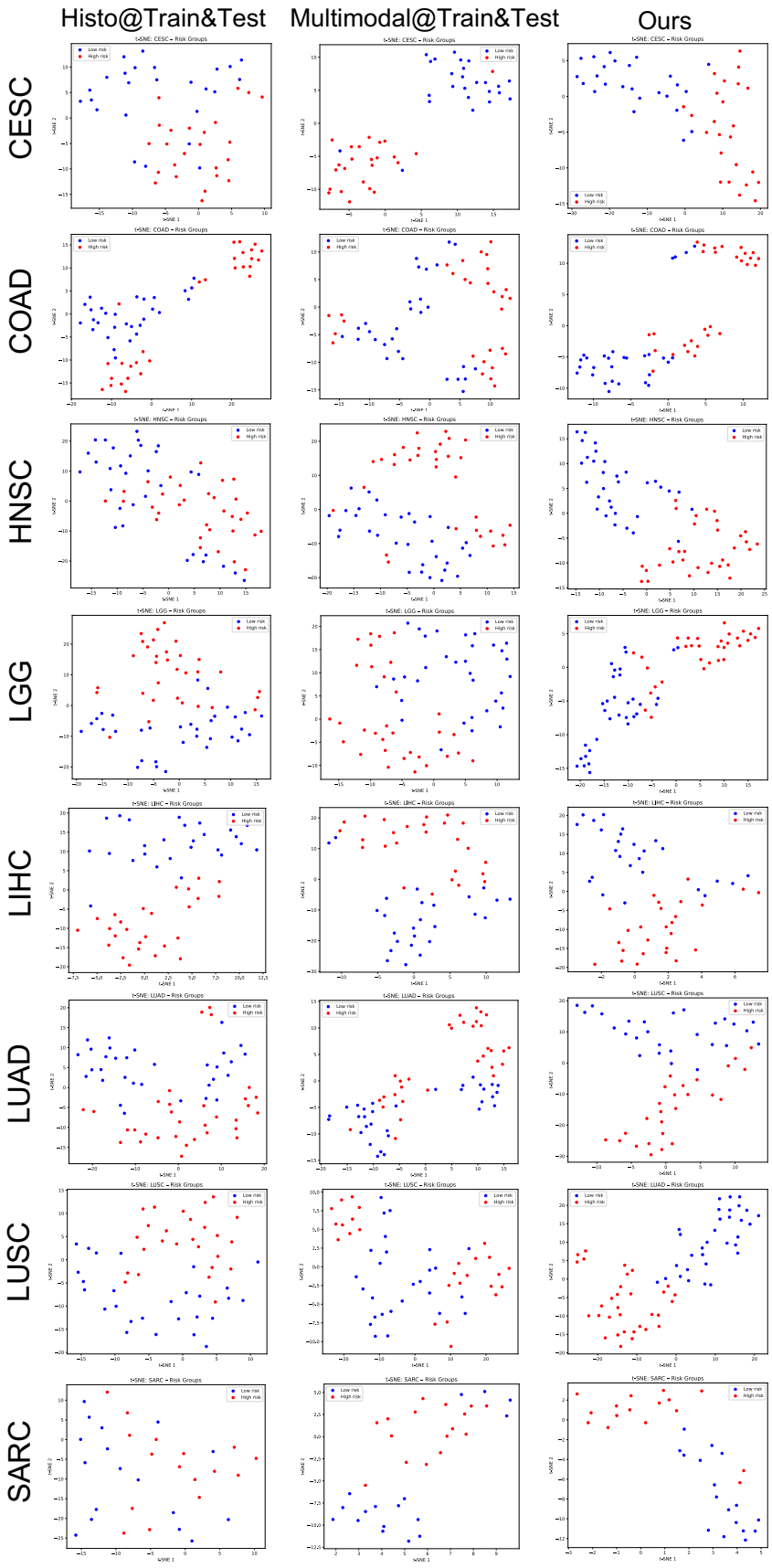

Figure 1. t-SNE visualization of patient embeddings

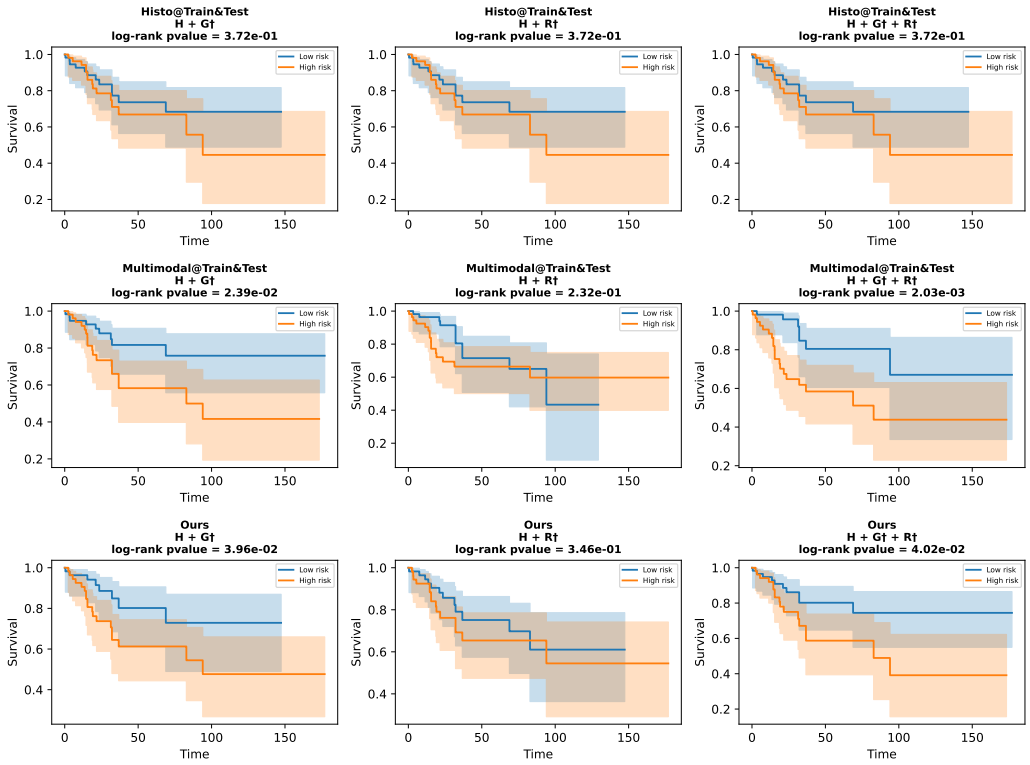

Figure 2. KM curves on the **CESC** dataset. Top: baseline (H only), middle: multimodal, bottom: Ours using only H at inference.

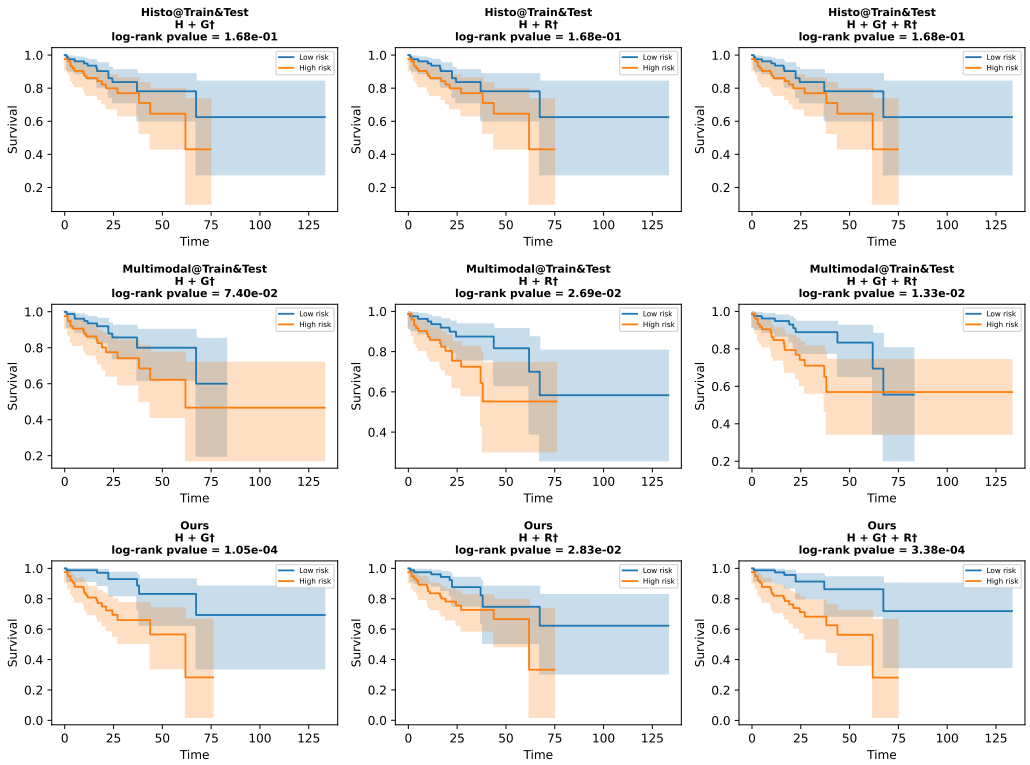

Figure 3. KM curves on the **COAD** dataset. Top: baseline (H only), middle: multimodal, bottom: Ours using only H at inference.

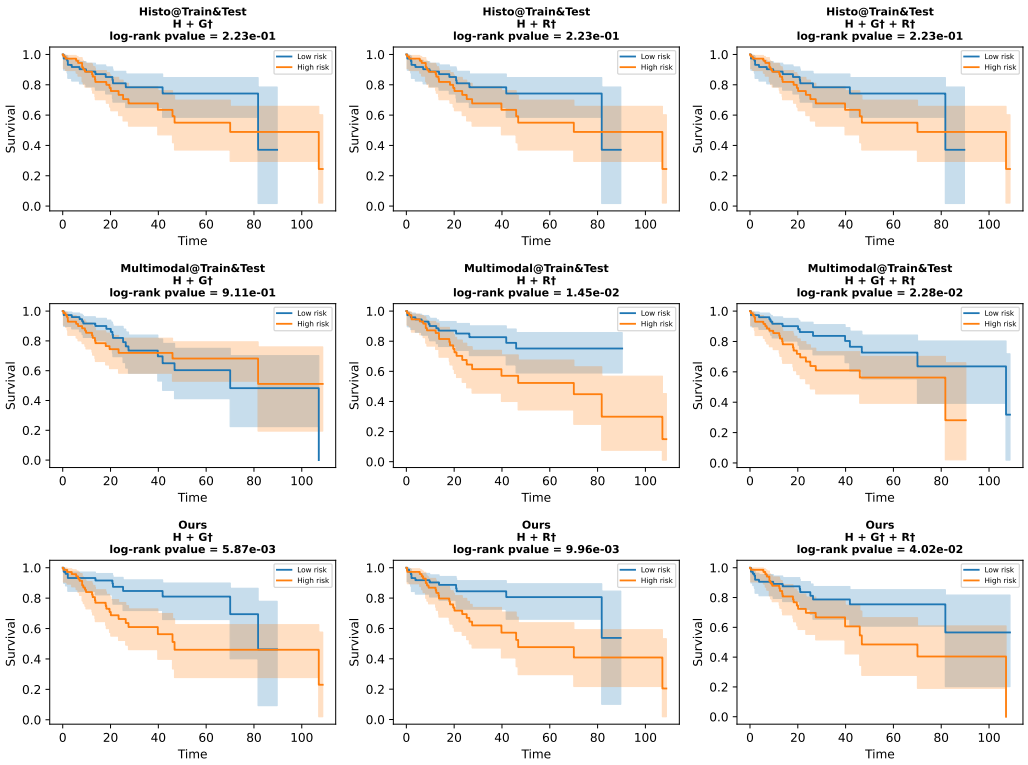

Figure 4. KM curves on the **LIHC** dataset. Top: baseline (H only), middle: multimodal, bottom: Ours using only H at inference.

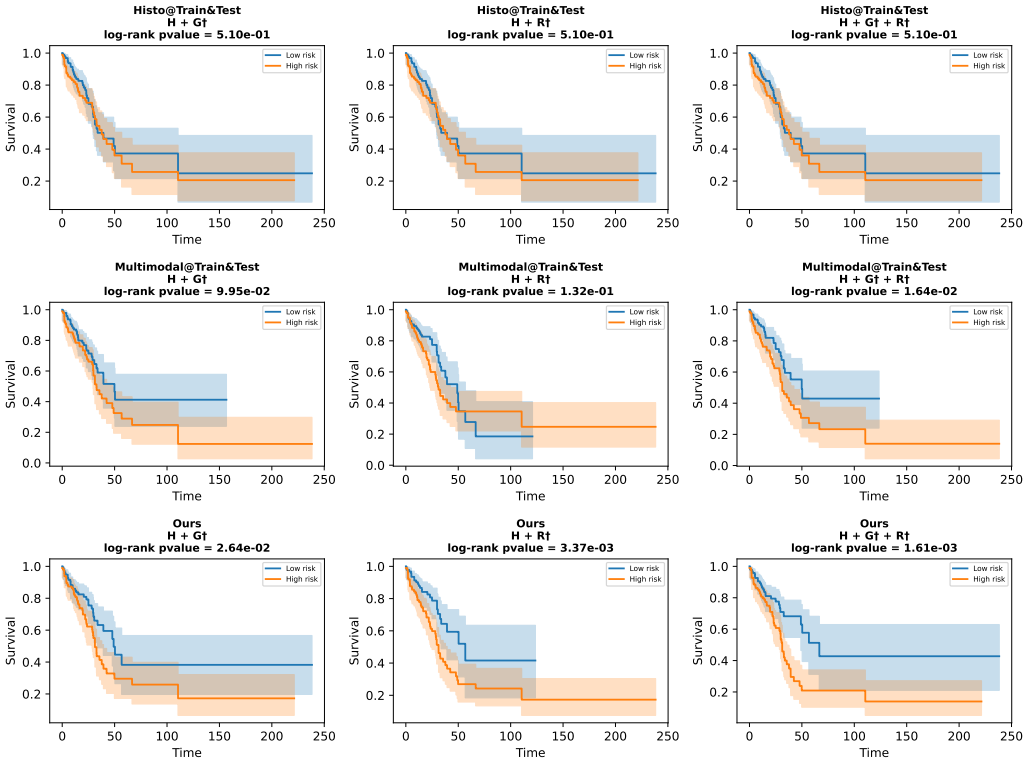

Figure 5. KM curves on the **LUAD** dataset. Top: baseline (H only), middle: multimodal, bottom: Ours using only H at inference.

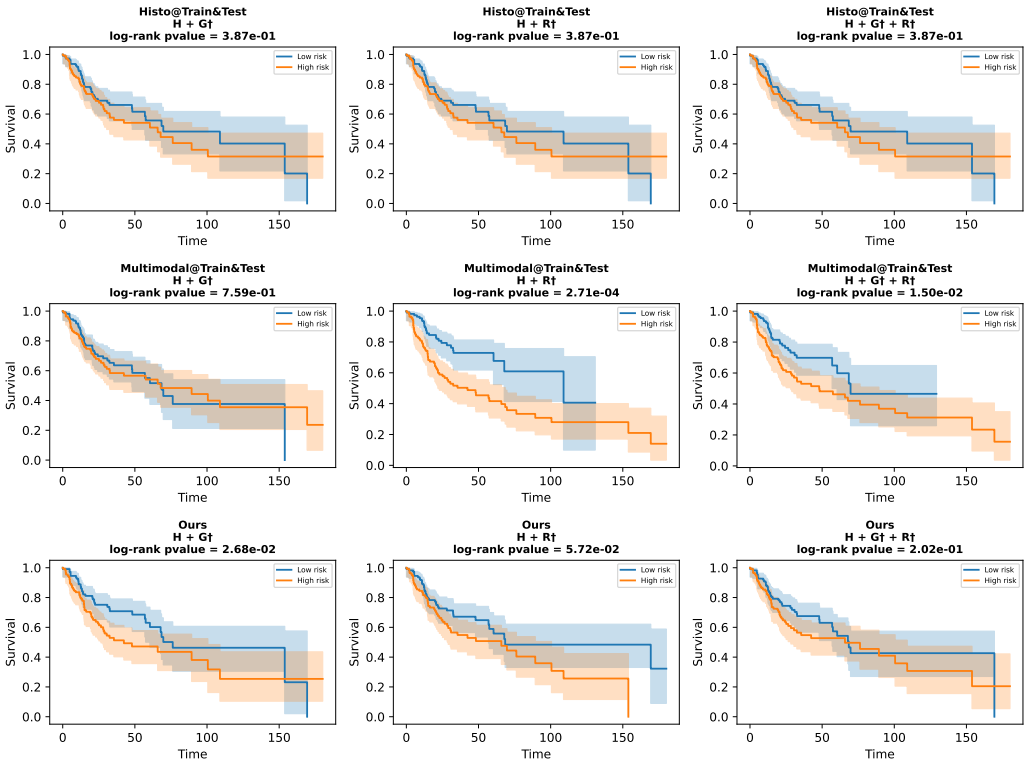

Figure 6. KM curves on the **HNSC** dataset. Top: baseline (H only), middle: multimodal, bottom: Ours using only H at inference.

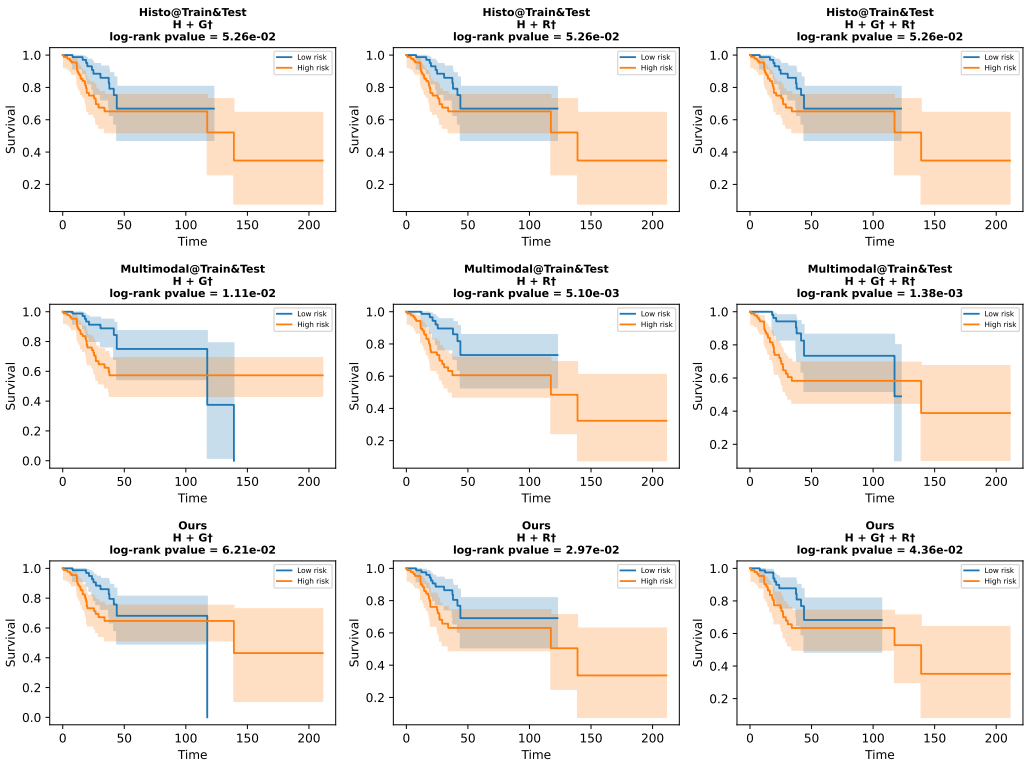

Figure 7. KM curves on the **LGG** dataset. Top: baseline (H only), middle: multimodal, bottom: Ours using only H at inference.

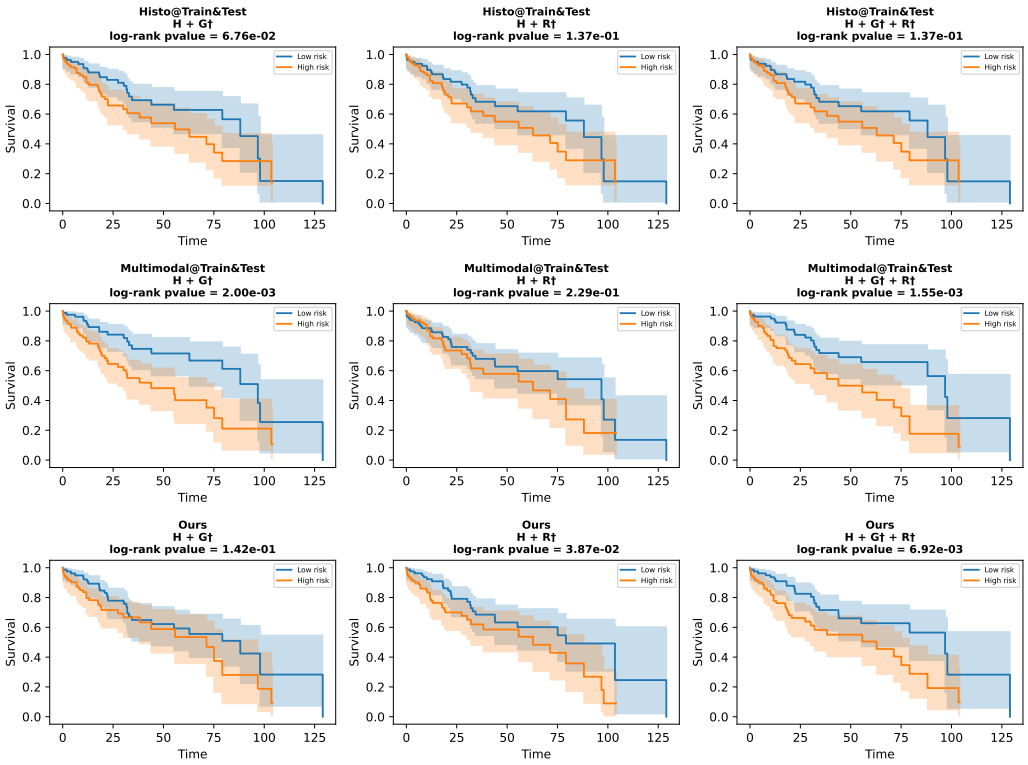

Figure 8. KM curves on the **LUSC** dataset. Top: baseline (H only), middle: multimodal, bottom: Ours using only H at inference.

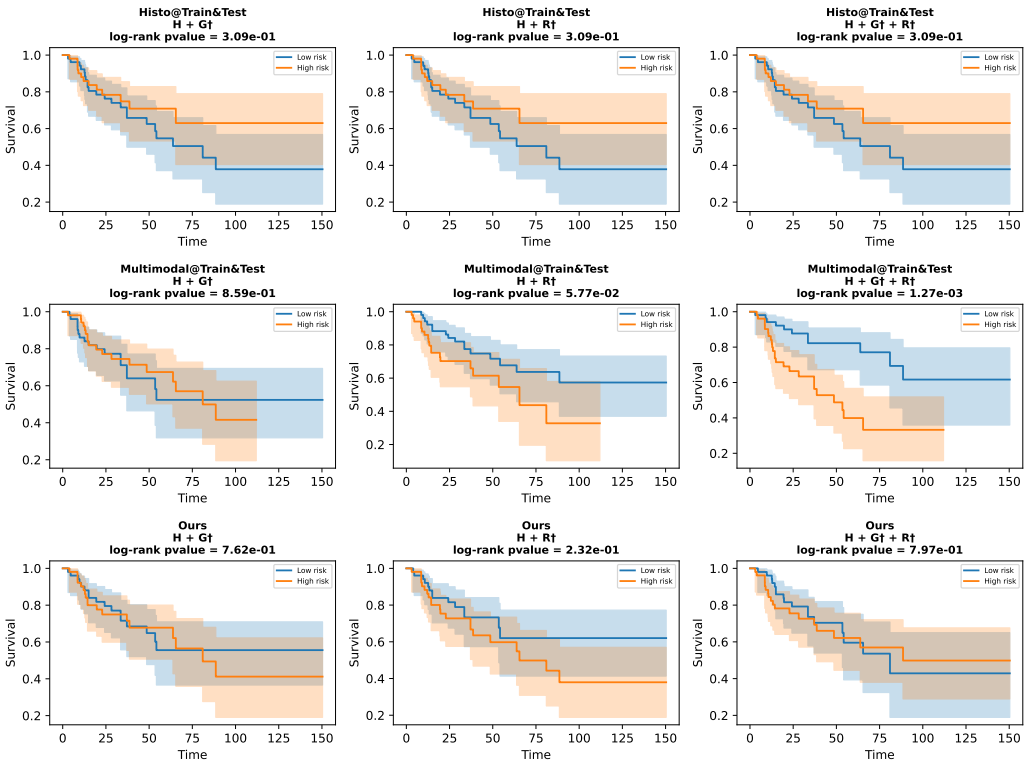

Figure 9. KM curves on the **SARC** dataset. Top: baseline (H only), middle: multimodal, bottom: Ours using only H at inference.