MUST: Modality-Specific Representation-Aware Transformer for Diffusion-Enhanced Survival Prediction with Missing Modality

Supplementary Material

The Supplementary Material provides detailed implementation settings, training configurations, and additional experimental results.

1. Implementation Details

1.1. Datasets

We evaluate our method on five cancer datasets from The Cancer Genome Atlas (TCGA): Bladder Urothelial Carcinoma (BLCA, $N = 347$), Breast Invasive Carcinoma (BRCA, $N = 899$), Glioblastoma and Low-Grade Glioma (GBMLGG, $N = 546$), Lung Adenocarcinoma (LUAD, $N = 415$), and Uterine Corpus Endometrial Carcinoma (UCEC, $N = 444$). For genomic data, we integrate three molecular data types: RNA-seq expression profiles, copy number variations (CNV), and single nucleotide variations (SNV). Samples with missing values in any of these genomic modalities are excluded from our study to ensure fair comparison across methods. Following the cancer hallmarks framework [2], we organize genes into six biologically meaningful functional groups: Tumor suppression, Oncogenesis, Protein kinases, Cellular differentiation, Transcription factors, and Cytokines and growth factors.

1.2. Survival Time Discretization

For the discrete-time survival model described in Sec. 3.1 of the main paper, we partition the continuous time axis into $K = 4$ intervals. The interval boundaries are determined by ensuring that uncensored samples are distributed uniformly across bins, with censored samples subsequently assigned to the appropriate bins based on their observation times. This binning strategy balances computational efficiency with sufficient temporal resolution for survival prediction.

1.3. Feature Extraction

For pathological data, we employ the UNI2-h histopathology foundation model [1] to extract patch-level features from whole slide images at 20× magnification. Each patch is encoded into a 1536-dimensional feature vector, which is then projected to the common embedding dimension of $D = 256$ through a linear layer. For genomic data, we do not use pretrained encoders. Instead, each of the six functional gene groups is processed by a group-specific multi-layer perceptron (MLP) that learns to encode the corresponding gene expression patterns into 256-dimensional embeddings. These genomic token encoders are trained end-to-end as part of the survival prediction task.

1.4. Architecture Details

As described in Sec. 3 (Method), our framework processes pathological and genomic tokens differently. For pathological tokens $P = \{p_1, \ldots, p_{N_P}\}$, we apply TransMIL [6] with learnable class tokens to obtain global representations g_P, modality-specific components u_P, and cross-attended components $c_{P \leftarrow G}$. For genomic tokens $G = \{g_1, \ldots, g_{N_G}\}$, we prepend modality-specific class tokens and apply multi-head self-attention layers to extract g_G, u_G, and $c_{G \leftarrow P}$. The low-rank shared subspace projection $P_\cap = B_\cap B_\cap^T$ uses rank $r = 64$ as determined by ablation studies in Sec. 4 (Experiments).

2. Training Configurations

2.1. Main Training

Due to the variable number of patches in whole slide images (ranging from 10K to 200K patches per slide), we set the batch size to 1 for all experiments. To ensure training stability, we employ gradient accumulation with 32 accumulation steps, effectively simulating a batch size of 32. Our two-stage progressive training strategy trains Stage 1 (warm-up) for 30 epochs using only the survival loss with Gaussian noise injection ($\sigma = 0.1$), followed by Stage 2 (main) for 30 epochs with the full objective including decomposition, shared consistency, and orthogonality losses. We use the AdamW optimizer with weight decay 1×10^{-5} for both stages. In Stage 1, we set the learning rate to 1×10^{-3}, and in Stage 2, we reduce it to 2×10^{-4} to facilitate stable learning of the algebraic constraints. Loss weights are set to $\lambda_{\text{dec}} = 1.0$, $\lambda_{\text{sh}} = 1.0$, and $\lambda_{\text{orth}} = 0.5$ as specified in the main paper.

2.2. Latent Diffusion Model Training

After the main network converges, we freeze all encoder and decomposition parameters and train conditional latent diffusion models [5] to reconstruct missing modality-specific components. The denoising network is implemented as a 4-layer transformer architecture operating on 256-dimensional feature embeddings. We train separate diffusion models for pathology-specific and genomic-specific component generation, each trained for 1 million steps using the AdamW optimizer with learning rate 1×10^{-4} and batch size 32. During inference, we perform 50-step DDIM [7] sampling to generate missing components. To account for the stochastic nature of the sampling process, we generate 5 samples per test instance and average the resulting features before

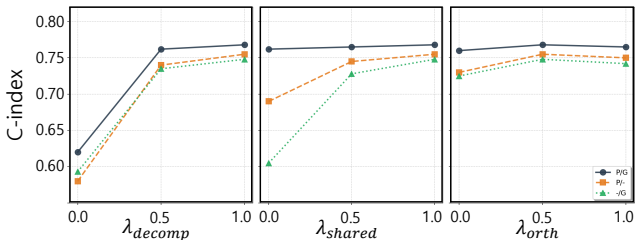

Figure 1. Hyperparameter sensitivity analysis on the UCEC dataset. Each loss weight (λ_{decomp}, λ_{shared}, λ_{orth}) is varied independently while keeping others at their default values. P/G denotes complete data, P/- denotes missing genomics, and -/G denotes missing pathology.

feeding them to the prediction head, as described in Sec. 4 (Experiments).

2.3. Comparison Methods

For fair comparison, we implement all baseline methods following their original papers while adapting them to our experimental setup. All experiments are conducted with batch size 1 due to the variable number of patches in whole slide images. SurvPath [3] is trained for 30 epochs using the Adam optimizer with learning rate 1×10^{-3}. CMTA [10] is trained for 30 epochs using SGD with learning rate 1×10^{-3}. For methods handling missing modalities, SMIL [4] is trained for 50 epochs with Adam optimizer (learning rate 1×10^{-3}), M^{3}Care [9] for 30 epochs with Adam (learning rate 1×10^{-3}), ShaSpec [8] for 30 epochs with Adam (learning rate 1×10^{-3}), and LD-CVAE [11] for 30 epochs with Adam (learning rate 2×10^{-4}). All methods are trained until the training C-index fully converges to ensure optimal performance.

2.4. Evaluation Protocol

We perform 5-fold cross-validation for all experiments and report the concordance index (C-index) along with standard deviations across folds. For each fold, we split the data at the patient level to ensure no data leakage. Statistical significance of risk group stratification is assessed using the log-rank test on Kaplan-Meier survival curves with the median predicted risk score as the threshold for defining high-risk and low-risk groups.

3. Additional Experimental Results

3.1. Hyperparameter Sensitivity Analysis

Fig. 1 evaluates the sensitivity of MUST to the three loss weights λ_{dec}, λ_{sh}, and λ_{orth} on the UCEC dataset. Each weight is varied independently while keeping others at their default values ($\lambda_{\text{dec}} = 1.0$, $\lambda_{\text{sh}} = 1.0$, $\lambda_{\text{orth}} = 0.5$).

$\mathcal{L}_{\text{decomp}}$ has the most pronounced effect: setting $\lambda_{\text{dec}} = 0$ causes severe degradation across all scenarios, as the algebraic decomposition structure breaks down entirely. With non-zero values, performance stabilizes rapidly, confirming that the decomposition constraint is essential but not sensitive to precise tuning. $\mathcal{L}_{\text{shared}}$ primarily affects missing modality scenarios: removing it ($\lambda_{\text{sh}} = 0$) leads to substantial drops under missing genomics and missing pathology, since cross-modal recovery relies on aligned shared components. Complete-data performance remains largely unaffected, as the shared component is not directly used during complete-data inference. $\mathcal{L}_{\text{orth}}$ shows the gentlest effect: even at $\lambda_{\text{orth}} = 0$, performance degrades only moderately, though with notable modality collapse where cross-modal recovery becomes less reliable. Our selected values ($\lambda_{\text{dec}} = 1.0$, $\lambda_{\text{sh}} = 1.0$, $\lambda_{\text{orth}} = 0.5$) balance complete-data performance with missing-modality robustness across all three loss terms.

3.2. Training-Time Missingness

While MUST is primarily designed for paired training with complete modalities, its architectural structure—decomposing representations into shared and modality-specific components—naturally suggests a pathway for handling training-time missingness as well. Specifically, the shared consistency constraint ($\mathcal{L}_{\text{shared}}$) aligns shared components across modalities, meaning that even when one modality is absent during training, the available modality's shared component can still provide a meaningful surrogate for cross-modal recovery. This structural property motivates us to explore whether MUST can be adapted to scenarios where complete pairs are not always available.

To evaluate this, we conduct experiments on the UCEC dataset by varying the proportion of unpaired samples (20% and 50% missing rate). For unpaired data, we substitute zero-tensors for the missing modality-specific component $\hat{u}$ and apply only the survival loss $\mathcal{L}_{\text{surv}}$, while training the LDM with the available $\tilde{c}$ as a surrogate condition.

Table 1. Training-time missingness evaluation on UCEC. Values show C-index without $\rightarrow$ with LDM-based inference.

Missing Rate	Missing G	Missing P
0% (baseline)	0.755	0.748
20%	0.716 $\rightarrow$ 0.735	0.606 $\rightarrow$ 0.724
50%	0.713 $\rightarrow$ 0.724	0.601 $\rightarrow$ 0.716

Tab. 1 shows that MUST remains robust under training-time missingness, and the performance gain when replacing zero-tensors with LDM-sampled residuals at inference confirms that $\mathcal{L}_{\text{shared}}$ effectively aligns shared components for cross-modal recovery even with incomplete training pairs. While our framework prioritizes paired training to establish strict algebraic constraints, these results demonstrate practi-

cal adaptability to realistic clinical scenarios where complete data may not always be available during model development.

3.3. Kaplan-Meier Survival Analysis

To provide comprehensive validation of our method's robustness under missing modality scenarios, we present detailed Kaplan-Meier survival curves comparing MUST against representative baseline methods: ShaSpec [8] and SMIL [4]. We focus on these two methods as LD-CVAE [11]'s unidirectional architecture only supports missing genomics scenarios. We stratify patients into high-risk and low-risk groups at the median predicted risk score and assess statistical significance using the log-rank test across all five TCGA datasets.

When both modalities are available as shown in Fig. 2, all three methods achieve statistically significant risk stratification ($p < 0.05$) across all datasets, demonstrating comparable discriminative power. The clear separation between high-risk and low-risk groups indicates that all methods effectively learn prognostic patterns when full multimodal information is available.

When genomic data is unavailable as shown in Fig. 3, substantial differences in robustness emerge. MUST maintains statistically significant stratification across all five datasets, demonstrating effective reconstruction of missing genomic-specific components. In contrast, both baselines show notable deterioration. ShaSpec loses statistical significance on BRCA ($p = 1.338\text{e-}01$) and LUAD ($p = 2.762\text{e-}01$), indicating weakened discriminative capability. SMIL exhibits severe vulnerability, failing to maintain significance on BLCA ($p = 6.266\text{e-}01$), BRCA ($p = 1.111\text{e-}01$), and LUAD ($p = 8.771\text{e-}01$). This catastrophic degradation confirms that alignment-based implicit imputation cannot adequately compensate for missing information without explicit modeling of modality-specific structures.

In Fig. 4, MUST continues to exhibit remarkable robustness when pathological images are missing, maintaining statistically significant stratification across all datasets with minimal degradation from complete data performance. This validates that our diffusion-based reconstruction of pathology-specific components preserves essential prognostic information. ShaSpec also maintains statistical significance across all datasets in this scenario, though with moderate performance degradation compared to MUST. SMIL shows intermediate vulnerability, losing significance on UCEC ($p = 1.475\text{e-}01$) while maintaining significance on other datasets.

3.4. Decomposition Analysis

Fig. 5 presents pairwise cosine similarities between global representations and their algebraic decompositions across all five datasets. The analysis validates that our decomposition successfully reconstructs the original global representations while maintaining their semantic structure.

The key validation comes from comparing global representations with their decomposed counterparts. For pathology, the decomposition achieves high reconstruction fidelity with cosine similarities ranging from 0.82 (GBMLGG) to 0.94 (BRCA), confirming that $g_P \approx \hat{u}_P + \hat{c}_{G \leftarrow P}$ holds with high precision. For genomics, the reconstruction quality is similarly high with similarities from 0.80 (GBMLGG) to 0.93 (BLCA), validating that $g_G \approx \hat{u}_G + \hat{c}_{P \leftarrow G}$.

Cross-modality similarities between g_P and g_G are notably low (0.03-0.31), confirming that the two modalities capture complementary rather than redundant information. Similarly, the low similarities between g_P and $\hat{u}_G + \hat{c}_{P \leftarrow G}$ (0.14-0.21) and between g_G and $\hat{u}_P + \hat{c}_{G \leftarrow P}$ (0.06-0.25) further confirm that modality-specific components retain distinctive characteristics even after decomposition.

The consistently high reconstruction similarities across both modalities explain why MUST can maintain robust performance under missing modality scenarios. When a modality is missing, the decomposition framework enables deterministic reconstruction through the algebraic relation: the shared component can be derived from the available modality ($\tilde{c} = g_{\text{available}} - \hat{u}_{\text{available}}$), and the missing modality-specific component is generated via diffusion conditioning on this shared component.

References

[1] 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

[2] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011. 1

[3] 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. 2

[4] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2302–2310, 2021. 2, 3

[5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[6] 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. 1

[7] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1

[8] Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15878–15887, 2023. 2, 3

[9] Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao. M3care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 2418–2428, 2022. 2

[10] Fengtao Zhou and Hao Chen. Cross-modal translation and alignment for survival analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21485–21494, 2023. 2

[11] Junjie Zhou, Jiao Tang, Yingli Zuo, Peng Wan, Daoqiang Zhang, and Wei Shao. Robust multimodal survival prediction with conditional latent differentiation variational autoencoder. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10384–10393, 2025. 2, 3

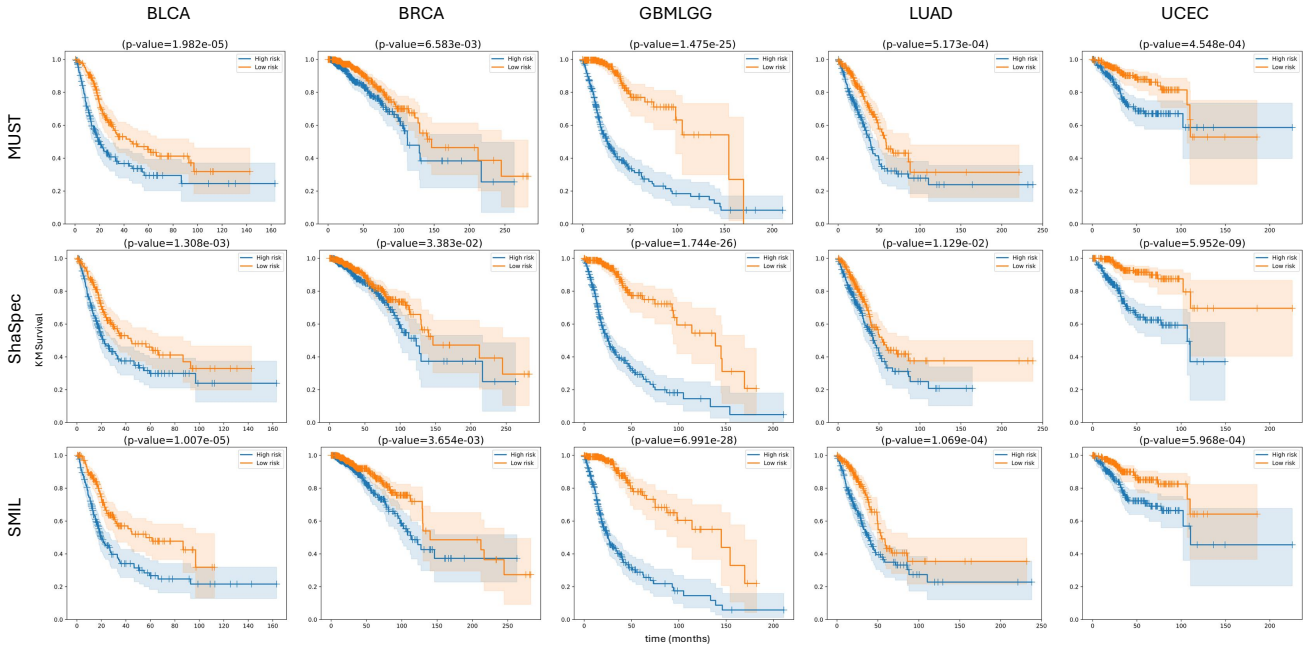

Figure 2. Kaplan-Meier survival curves on five different datasets comparing high-risk and low-risk groups across complete scenario. p-values are computed using the log-rank test. Patients are stratified into high-risk (blue) and low-risk (orange) groups based on predicted risk scores at the median threshold. Shaded regions represent 95% confidence intervals, and censored observations are indicated by vertical tick marks on the curves.

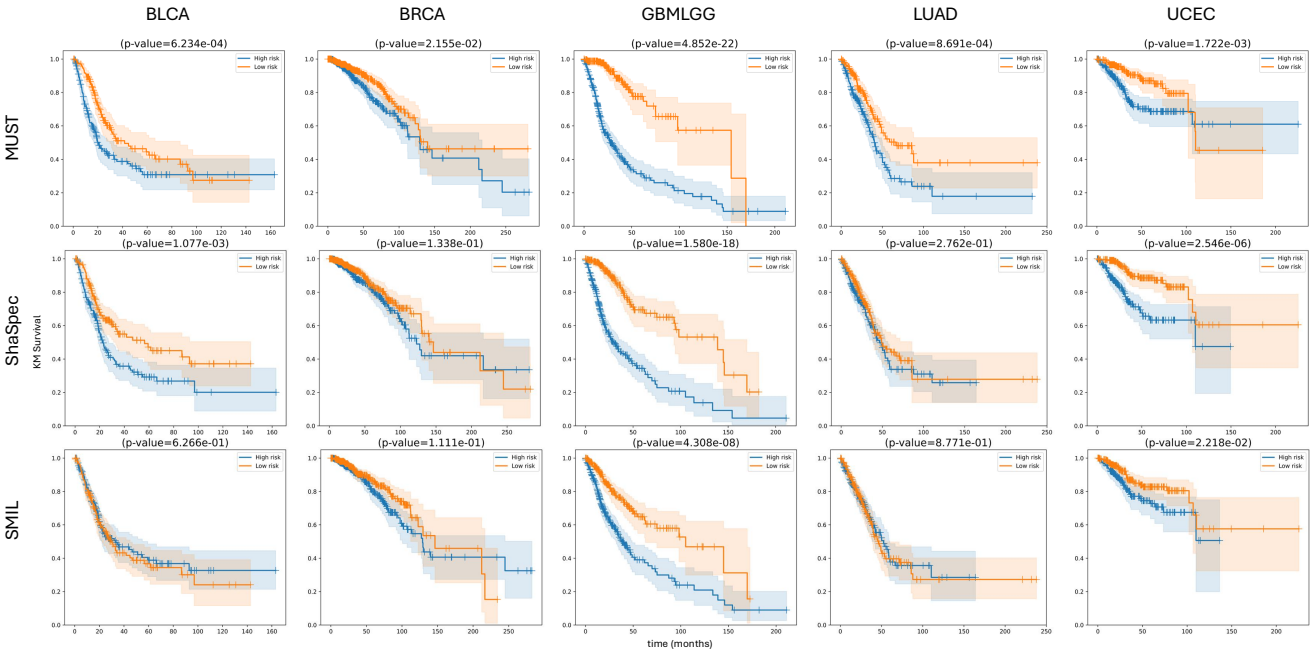

Figure 3. Kaplan-Meier survival curves on five different datasets comparing high-risk and low-risk groups across missing genomics (G) scenario. p-values are computed using the log-rank test. Patients are stratified into high-risk (blue) and low-risk (orange) groups based on predicted risk scores at the median threshold. Shaded regions represent 95% confidence intervals, and censored observations are indicated by vertical tick marks on the curves.

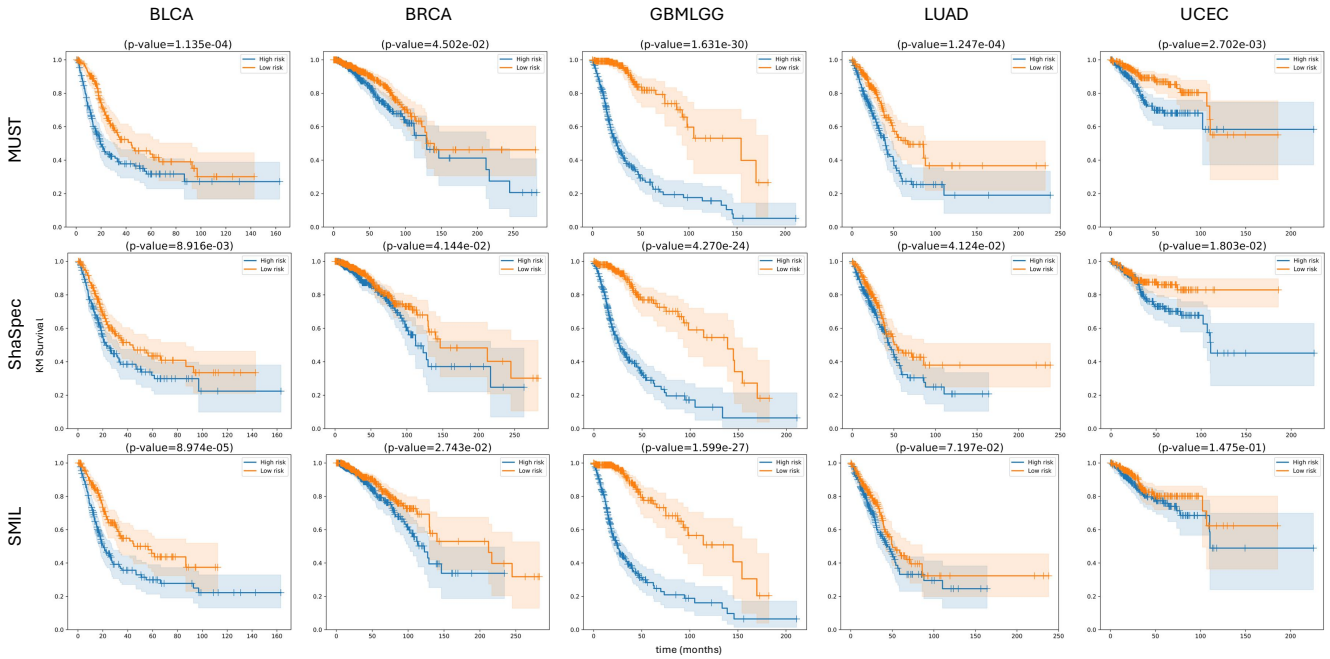

Figure 4. Kaplan-Meier survival curves on five different datasets comparing high-risk and low-risk groups across missing pathology (P) scenario. p-values are computed using the log-rank test. Patients are stratified into high-risk (blue) and low-risk (orange) groups based on predicted risk scores at the median threshold. Shaded regions represent 95% confidence intervals, and censored observations are indicated by vertical tick marks on the curves.

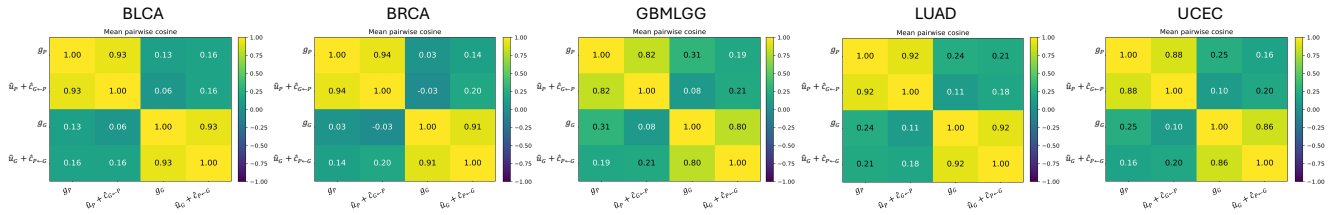

Figure 5. Cosine similarity map between g and corresponding composition of $\hat{u}$ and $\hat{c}$ across all five TCGA datasets.