PARACHUTE: Pathology-Radiology Cross-Modal Fusion for Missing-Modality-Robust Survival Prediction

Pietro Caforio*, Isabella Poles*, Marco D. Santambrogio
Politecnico di Milano
pietro.caforio@mail.polimi.it, {isabella.poles, marco.santambrogio}@polimi.it

Abstract

Survival prediction from medical imaging is a critical challenge in computational oncology, with high clinical relevance for patient stratification and treatment planning. However, current Deep Learning methods suffer from three core limitations: they assume complete modality availability, overlook local-to-global cross-modal interactions, and disregard modality-specific signal reliability during optimization. To address these issues, we introduce PARACHUTE[1]*, a novel Deep Learning framework for robust multimodal survival prediction from heterogeneous and partially missing imaging data.* PARACHUTE *integrates modality-specific pretrained encoders with adapter networks that align radiology and histopathology features into a shared latent space. A Dynamic Contextual Embedding mechanism captures biologically grounded local correlations between pathology and radiology and channels them through a multi-head cross-attention fusion module to guide global survival prediction, while adaptively handling missing modality scenarios. Furthermore, a Gradient Curvature Steering module improves convergence in incomplete data regimes by regularizing gradients via local curvature alignment. Experiments on three CPTAC and TCGA derived cancer cohorts show that* PARACHUTE *achieves a C-index of 0.8367 with full modality input, and it retains strong performance under missing modality conditions (0.7488) while producing clinically meaningful risk stratifications, as confirmed by Kaplan–Meier analysis.*

1. Introduction

Survival analysis predicts patient survival times supporting risk stratification and clinical decision-making. Among the diverse biomarkers, imaging [27] is the most *routinely* employed due to its accessibility and rich phenotypic content [28]. Indeed, clinicians rely primarily on two complementary imaging modalities in everyday practice: radiology to provide organ-scale views for screening and staging [2], and histopathology to offer cellular-level insights for diagnostic confirmation and fine-grained prognosis [14].

An increasing number of studies have explored combining radiological and pathological images to improve cancer prognosis assessment [29, 32, 40]. Notably, An *et al.* [1] present a two-stage cross-modal fusion with cross-attention and low-rank interactions for robust survival prediction across cancers. Building on this, Song *et al.* [31] adapt cross-modality and cross-region self-attention for survival and tumor grading in oropharyngeal carcinoma.

While multimodal cancer analysis is gaining traction, most methods assume full modality availability, an assumption often broken by clinical realities like missing histopathology from invasive biopsies or absent radiology acquired outside workflows [36]. Compounding this, modeling multiscale radiology–histopathology dependencies remains challenging. Locally, histological details shape regional radiographic patterns [17]. Globally, where broader contextual signals are integrated, local cross-modal interactions propagate to influence the direction of the disease assessment [11, 45]. Such dependencies are nonlinear, sparse, and asymmetric, and only specific regions carry cross-modal relevance. To exemplify, in Pancreatic Ductal Adenocarcinoma (PDA), strong local cues, such as radiology ductal cutoff and histopathology stromal desmoplasia, can reinforce an otherwise weak global associations between poor tumor demarcation and widespread stromal involvement, ultimately guiding diagnosis [15, 23]. Common strategies handle missing modalities by requiring auxiliary branches and extra losses via conditional or gated cross-attention [21, 33] or reconstructing one from the other [33, 37], using memory banks [34] or disentangled spaces [30], often combined with self-attention pooling over radiology and histopathology embeddings to model cross-modal interactions. However, three key challenges hinder their effectiveness. *First*, additional objectives in conditional or gated cross-attention add non-trivial training overhead, increase model complexity and destabilize

*Equal contribution
[1] https://github.com/PietroCaforio/parachute

model optimization. *Second*, the sheer scale of Whole Slide Images (WSIs) and radiology volumes (e.g., $150K^2$ pixels and 512^3 voxels) with sparse diagnostic regions [44] renders reconstruction or dictionary learning computationally prohibitive. *Third*, standard attention and pooling lack biologically grounded inductive biases to capture the causal, local-to-global dependencies inherent in cross-modal interactions [18], further amplifying sensitivity to sharp minima and gradient conflicts, especially under missing modalities, where naive optimization easily overfits or collapses.

To address these challenges, we introduce PARACHUTE(Pathology-Radiology Cross-modal Chute), a multimodal fusion framework that captures biologically grounded local correlations between pathology and radiology and channels them through a cross-modal chute to guide global survival prediction, while maintaining robustness to missing modalities. Specifically, PARACHUTE employs a dynamic contextual embedding to model local cross-modal correspondences with awareness of missing data, injecting spatial alignment cues into the global representation. Additionally, a gradient steering mechanism prioritizes informative regions while reducing the influence of incomplete or noisy data. Experiments across three cancer survival cohorts show consistent gains under both complete and missing modality settings. Our main contributions are:

- We introduce a Dynamic Contextual Embedding (DCE) that captures local cross-modal feature correlations while adaptively handling missing modality scenarios.
- We propose a local-to-global relational modeling strategy that injects biologically grounded inductive biases by transferring local cross-modal correlations and missing data awareness into the global fusion process.
- We incorporate a Gradient Curvature Steering (GCS) mechanism that dynamically steers gradient amplitudes and directions along low-curvature paths of the local loss surface geometry when modalities are noisy/incomplete.

2. Related work

2.1. Survival Analysis in Multimodal Learning

Integrating radiology and histopathology captures complementary multiscale tumor features critical for accurate survival prediction. Recent approaches such as M4Survive [18] leverage Mamba-based adapters to fuse pretrained foundation models embeddings learning a shared latent space and a token-based fusion strategy to preserve fine-grained, modality-specific details for survival risk estimation. In HPV-associated oropharyngeal cancer, SMuRF [31] uses a Swin Transformer to fuse radiology features from tumors and lymph nodes with WSIs, modeling cross-modality and cross-region interactions via multi-head self-attention. Further multimodal fusion strategies have emerged from integrating histopathology and genomics biomarkers. MCAT [3] employs a co-attention transformer for WSIs with genomic features dense mapping, while Xu *et al.* use optimal transport to match instances across modalities [39]. Zhang *et al.* address redundancy in multimodal survival modeling with PIBD, and CA-MLIF [1] uses contrastive attention to preserve modality-specific discriminative power during fusion. However, these methods generally assume full modality availability during training and inference, limiting their robustness in real-world clinical settings where missing data is common.

2.2. Missing Modality in Multimodal Learning

To address missing modality in multimodal learning, some methods aim to reconstruct missing data: MMD [6] uses mean fusion with modality reconstruction losses, while LD-CVAE [43] generates missing embeddings from the available modality using a bottleneck transformer and a latent differentiation VAE, fusing them via a product-of-experts strategy. Other methods avoid explicit reconstruction by retrieving representative features from modality-specific memory banks, as in MGIML [34], which augments missing inputs using cross-modal memory reads. To enhance representation learning, auxiliary-task-based methods promote disentanglement of shared and modality-specific features [33], while DRIM [30] explicitly models each embedding as a combination of modality-unique and patient-shared components, improving robustness to missing data. Contrastive learning has also been extended to capture inter-patient similarity [25]: Farooq *et al.* [8] incorporate a cross-patient module to align embeddings across semantically related cases even with partial input. However, recovering missing high-resolution image profiles remains challenging. Existing methods retain modality- and patient-specific signals but rely on uniform fusion, lacking adaptive mechanisms to transfer local cross-modal cues to global prediction based on modality availability.

2.3. Gradient Optimization

Gradient-based strategies have been widely employed to mitigate optimization conflicts in multi-task and multimodal learning. GradNorm [4] reweights task-specific gradients to balance magnitudes, while PCGrad [41] and GradVac [35] address direction conflicts through orthogonal projection techniques. RotoGrad [13] and MGIML [34] extend this by learning rotation matrices in feature space to align gradients, with MGIML explicitly modeling modality imbalance via instance-specific gradient homogenization. In parallel, Sharpness-Aware Minimization (SAM) [10] improves generalization by minimizing the worst-case loss via one-step adversarial perturbation, encouraging convergence to flat minima. However, SAM's approximation becomes less reliable in highly non-linear landscapes. To address

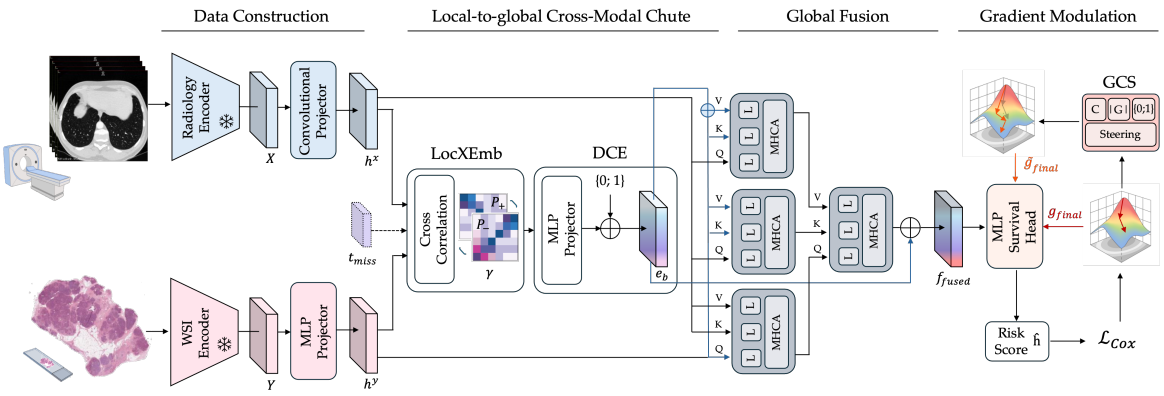

Figure 1. **Overview of PARACHUTE.** Histology and radiology are encoded into modality-specific features. The Dynamic Contextual Embedding (DCE) captures local cross-modal correlations while handling missing modalities. Cues guide global fusion for survival prediction. Gradient Curvature Steering (GCS) refines learning by prioritizing informative regions and mitigating noisy-incomplete signals.

this, CR-SAM [38] introduces a normalized Hessian trace as a regularizer, enabling explicit curvature-aware control of the optimization dynamics. While these methods focus on resolving gradient conflicts or enforcing smoothness post hoc, they neglect the local loss landscape geometry during training, especially in incomplete modality conditions.

3. Methods

3.1. Overview

The overall PARACHUTE architecture (Figure 1) processes radiology and pathology inputs via foundation models to extract modality-specific features. Biologically grounded local cross-modal correlations are captured by the Local Cross-Correlation Embedding (LocXEmb) module through Top-k token similarity, and refined by the Dynamic Contextual Embedding (DCE) module, which introduces a learnable missing modality token and interaction-specific local context for guided cross-attention. A Multi-Head Cross-Attention (MHCA) fusion module integrates the features, and Gradient Curvature Steering (GCS) stabilizes training by favoring low-curvature regions under missing modality conditions. The fused representation informs survival risk prediction via complementary local and global cues.

3.2. Problem Formulation

We address the task of time-to-event prediction in the presence of right-censored data, a common scenario in clinical survival analysis. The objective is to estimate the probability that a k-th patient will experience an event (e.g., death or recurrence) beyond a given time point. We model the survival function as $S^{(k)}(t) = P(T \geq t|\mathbf{I}(k))$ and the hazard function as $h^{(k)}(t) = P(T = t|T \geq t, \mathbf{I}^{(k)})$ where the patient-specific input tuple is defined as $\mathbf{I}^{(k)} = (R^{(k)}, H^{(k)}, c^{(k)}, t^{(k)})$. Here, $R^{(k)}$ and $H^{(k)}$ represent radiology and histopathology-derived features, respectively; $c^{(k)} \in \{0, 1\}$ is the censorship indicator (1 = event observed, 0 = censored); and $t^{(k)} \in \mathbb{R}^+$ denotes the observed survival time. We learn a representation $f(\mathbf{I}^{(k)})$ to estimate the survival outcome and optimize the survival loss $\mathscr{L}_{\text{surv}} = \mathscr{L}_{\text{surv}}\left(\left\{f(\mathbf{I}^{(k)}), t^{(k)}, c^{(k)}\right\}_{k=1}^{N_D}\right)$, where N_D is the number of training samples.

3.3. Feature Construction

3.3.1. Radiological Features Preparation

Radiological data are provided as 3D volumetric scans composed of sequential 2D slices. We denote the radiological volume as $R^k = \{r_i^{(k)}\}_{i=1}^N$ for the k-th patient, where N is the number of axial slices and each $r_i^{(k)} \in \mathbb{R}^{\mathbb{H}\times\mathbb{W}}$ represents a 2D image. Each slice is independently processed by a pretrained radiology encoder $\phi_r(\cdot)$, implemented via MedImageInsight [5], to obtain a set of per-slice feature vectors:

$$X^{(k)} = \{\phi_r(r_i^{(k)})\}_{i=1}^N = \{x_1^k, x_2^k, \cdots, x_N^k\}, \quad x_i^k \in \mathbb{R}^{d_r}, \tag{1}$$

where d_r denotes each slice-level embedding dimension. This set of features retains spatial context along the superior-inferior axis of the body and serves as the radiological input to the cross-modal survival model.

3.3.2. Pathological Features Preparation

Given the WSI $P^{(k)}$ for the k-th patient, we extract tissue-containing image patches $\{p_j^{(k)}\}_{j=1}^M$ at 20× magnification. These patches are processed using TITAN [7], implemented with the CONCHv1.5 [19] architecture, which acts as a pretrained encoder $\phi_p(\cdot)$. TITAN embeds patch-level features into a slide representation. The resulting histopathological embedding is denoted:

$$Y^{(k)} = \phi_p(\{p_j^{(k)}\}_{j=1}^M) \in \mathbb{R}^{d_p}, \tag{2}$$

where d_p is final slide-level feature dimension. This representation captures both local cellular morphology and spatially distributed tissue context, providing a strong pathological signal for survival modeling.

3.4. Local-to-global Cross-Modal Chute

3.4.1. Local Cross-Correlation Embedding

To capture fine-grained interactions between radiological and histopathological patterns, we introduce a Local Cross-Correlation Embedding (LocXEmb) module that projects features from both modalities into a shared space and explicitly models localized cross-modal semantic alignment.

Let $X \in \mathbb{R}^{B\times T_x\times C_r}$ and $Y \in \mathbb{R}^{B\times T_y\times C_y}$ denote tokenized radiology and histopathology features for a batch of B patients. Here, T_x and T_y represent the number of spatially localized tokens (e.g., radiology volume or histological slices), and C indicates channel dimensionality. We first standardize the representations via modality-specific semantic adapters that map raw features to a common intermediate embedding space of dimension D_{inter}:

$$\begin{aligned} h^{\mathrm{x}} &= \phi_{\mathrm{x}}(X) \in \mathbb{R}^{B\times T_{\mathrm{x}}\times D_{\text{inter}}}, \\ h^{\mathrm{y}} &= \phi_{\mathrm{y}}(Y) \in \mathbb{R}^{B\times T_{\mathrm{y}}\times D_{\text{inter}}}. \end{aligned} \tag{3}$$

In particular, the adapter ϕ_{x} comprises two convolutional blocks (kernel size 3, stride 1) with ReLU, dropout, and batch normalization, followed by adaptive average pooling and flattening to produce a fixed-size representation. In contrast, ϕ_{y} is a residual Multi Layer Perceptron (MLP) with stacked linear layers, GELU activations, dropout, and layer normalization. Given the adapted features, we define the local cross-modal similarity matrix for each patient b as: $S^{(b)} = \hat{h}^{\mathrm{x},(b)} \cdot \left(\hat{h}^{\mathrm{y},(b)}\right)^{\top}$ where $S_{x,y}(b)$ quantifies the similarity between the x-th radiological and the y-th histological token for patient b, with all h tokens l_2-normalized to ensure valid similarity computation. The similarity matrix $S^{(b)}$ captures the degree of affinity of localized semantic *contexts* that usually reflect biologically plausible interactions, such as tumor-stroma boundaries in histopathology corresponding to peritumoral edema or vascular encasement in radiology. To distill this matrix into an interpretable descriptor, we compute statistics over the top-bottom similarities as:

$$\begin{aligned} \mathrm{P}_{+}{}^{(b)} &= \frac{1}{k_+}\sum_{i=1}^{k_+} \text{Top-}k_+ \left(\mathrm{S}^{(b)}\right)_i, \\ \mathrm{P}_{\text{-}}{}^{(b)} &= \frac{1}{k_-}\sum_{j=1}^{k_-} \text{Top-}k_- \left(\mathrm{S}^{(b)}\right)_j, \end{aligned} \tag{4}$$

where Top-k_+ and Top-k_- operate on the flattened matrix $S^{(b)} \in \mathbb{R}^{T_x\cdot T_y}$. These statistics form the $\gamma^{(b)} = [\mathrm{P}_{+}{}^{(b)}, \mathrm{P}_{\text{-}}{}^{(b)}]$ 2D signal which quantifies the local cross-modal agreement and divergence for each patient. The formulation handles missing modalities by marginalizing similarity scores with learned modality priors.

3.4.2. Dynamic Contextual Embedding

To inject relational inductive bias into the fusion process, we propose a Dynamic Contextual Embedding (DCE) module that encodes local cross-modal similarity and adapts to missing modality conditions. In particular, the $\gamma^{(b)}$ 2D correlation signal, concatenation between the Top-k_+ and Top-k_- local similarities, is first projected via an MLP to obtain a dynamic embedding $e_{dyn}(b) = MLP(\gamma^{(b)})$, where $e_{\text{dyn}} \in \mathbb{R}^{B\times 1\times f_{\text{dim}}}$ with $f_{\text{dim}} = 64$. Afterwards, a missingness indicator $m(b) = \neg(m_x(b) \wedge m_y(b))$, where $m_x(b)$ and $m_y(b)$ are the availability flags for radiology and histopathology, is defined to adapt e_{dyn} to the presence of missing modalities. The resulting embedding $e(b) \in \mathbb{R}^d$ serves as a dynamic contextual signal that captures sample-specific local alignment and encodes missing-modality-aware adjustment via a learned token $t_{miss} \in \mathbb{R}^d$:

$$\begin{aligned} e(b) &= e_{\text{dyn}}(b) + m(b)\cdot t_{\text{miss}} \\ &= \mathrm{MLP}([\mathrm{P}_{+}{}^{(b)}, \mathrm{P}_{\text{-}}{}^{(b)}]) + m(b)\cdot t_{\text{miss}}. \end{aligned} \tag{5}$$

The final dynamic embedding $e(b) \in \mathbb{R}^{B\times 1\times f_{\text{dim}}}$ with $f_{\text{dim}} = 64$ can be injected into any attention-based fusion mechanism (i.e. self-attention, cross-attention, or co-attention) by modifying the value tensor as $\tilde{V} = V + e(i)$. To exemplify, given standard attention: $Attn(Q,K,V) = softmax\left(\frac{QK^{\intercal}}{\sqrt{d}}\right)V$ the value stream can be biased with our DCE as:

$$Attn_{bias}(Q,K,\tilde{V}) = softmax\left(\frac{QK^{\intercal}}{\sqrt{d}}\right)(V+e(b)). \tag{6}$$

This formulation enables local-to-global bias transfer, where local relational cues across modalities learned from cross-modal contextual similarity guide the global fusion process. Crucially, it preserves semantic guidance under missing modalities, as $e(b)$ adapts to available inputs.

3.4.3. Global Cross Attention Module

Building on the formulation above, we integrate the dynamic contextual bias $e(b)$ into a hierarchical Multi-Head Cross Attention (MHCA) fusion module to model interdependencies between radiology and histopathology at a global semantic level.

Each attention block follows the standard Transformer formulation in Equation (6), with $Q = XW^Q, K = YW^K, V = ZW^V, W^{Q,K,V} \in \mathbb{R}^{d\times d}$. The three MHCA blocks thus arise directly from the inherent Q/K/V decomposition of the Transformer, not ad-hoc design choices, while the attention mechanism value V biased with $e(b)$ serves as a learned, sample-specific local alignment signal $Attn_{bias}(Q,K,\tilde{V})$. The attention mechanism result

is then passed through residual and normalization layers $H = LayerNorm(X + Attn_{bias}(Q, K, \tilde{V})$ and afterwards, through a position-wise Feed-Forward Network (FFN) with skip connections and normalization to obtain the final features: $h^{out} = LayerNorm(H, \text{FFN}(H))$. In our scenario, we instantiate three parallel attention blocks to capture different *asymmetric* interactions between modalities and a fourth to fuse and refine them incorporating the contextual bias via residual connection:

$$\begin{aligned} f_V &= \text{MHCA}_1(h^x, h^y, h^x + h^y), \\ f_K &= \text{MHCA}_2(h^x, h^y, h^y), \\ f_Q &= \text{MHCA}_3(h^y, h^x, h^x), \\ f_{fused} &= \text{MHCA}_4(f_Q, f_K, f_V + e(b)). \end{aligned} \tag{7}$$

This hierarchical strategy allows the model to gradually synthesize *asymmetric local cues* into a *unified global representation*, with the *DCE guiding attention across both complete and incomplete modality settings*. If a modality is missing, modality-specific learned tokens are substituted, ensuring stable fusion behavior regardless of input completeness. MHCA internals are detailed in *Supplemental*.

3.5. Survival Prediction and Gradient Steering Mechanism

The final fused representation $f_{fused} \in \mathbb{R}^d$ is passed to a three-layer MLP, denoted as the hazard network g_ϑ, to estimate the log-risk score $\hat{h}_b = g_\vartheta(f_{fused})$ for patient b. Training minimizes the Cox partial negative log-likelihood:

$$\mathscr{L}_{\text{Cox}} = -\frac{1}{N_D} \sum_{b=1}^{N_D} \delta_b \left(\hat{h}_b - \log \sum_{j:t_j \geq t_b} e^{\hat{h}_j} \right), \tag{8}$$

where each patient is described by a tuple (t_b, δ_b, x_b) with t_b the observed time, $\delta_b \in \{0, 1\}$ the event indicator (1 if event, 0 if right-censored), and x_b the multimodal features. Right-censoring is thus explicitly handled via δ_b: censored samples do not contribute to the numerator but remain in the denominator risk sets, ensuring robust learning from both censored and observed data while modeling relative risk ordering across the cohort.

To improve generalization and avoid sharp minima, we introduce Gradient Curvature Steering (GCS), inspired by SAM but distinct from SAM/CR-SAM and gradient-conflict methods. While SAM/CR-SAM perturb parameters to penalize sharp optima, and RotoGrad/PCGrad orthogonalize or rotate task gradients, GCS steers *feature-level* gradients using per-sample curvature, gradient magnitude, and modality completeness. For each patient b, the feature-level gradient and curvature estimate are given by:

$$g_b = \nabla_{f_{\text{final}_b}} \mathscr{L}_{\text{Cox}}, \quad c_b \approx \tfrac{g_b^\top r_b - g_b^\top(-r_b)}{2\varepsilon}, \quad r_b \sim \text{Rademacher}, \tag{9}$$

Algorithm 1 Gradient Curvature Steering (GCS)

Require: Mini-batch $\mathcal{B} = \{(f_b^{(x)}, f_b^{(y)}, t_b, \delta_b, m_b)\}_{b=1}^B$; parameters θ (fusion encoder), ϑ (hazard net)
Ensure: Updated θ, ϑ

1: **Forward:** $f_b \leftarrow \text{FUSION}(f_b^{(x)}, f_b^{(y)}, m_b)$; $\hat{h}_b \leftarrow \text{HAZARDNET}(f_b)$
2: Compute Cox loss: $\mathcal{L}_{\text{Cox}} = -\frac{1}{B}\sum_b \delta_b \left(\hat{h}_b - \log \sum_{j:t_j \geq t_i} e^{\hat{h}_j} \right)$
3: **Curvature steering:** $g_b \leftarrow \nabla_{f_b} \mathcal{L}_{\text{Cox}}$; $c_b \approx (g_b^\top r_b - g_b^\top(-r_b))/(2\varepsilon)$ $\omega_b \leftarrow \sigma(a_1 c_b + a_2 \|g_b\| + a_3 m_b + c)$ $\tilde{g}_b \leftarrow (1 - \omega_b) g_b + \omega_b g_b / (1 + \gamma c_b)$
4: **Backward:** Update θ with $\tilde{g}_b$, and ϑ with $\mathcal{L}_{\text{Cox}}$
5: **Step:** Apply optimizer update on θ, ϑ

with c_b approximating local curvature via Hutchinson's method. A controller combines curvature, gradient norm, and missing-modality flag into a gating weight $\omega_b = \sigma(a_1 c_b + a_2 \|g_b\|_2 + a_3 m_b + c)$ yielding the steered gradient

$$\tilde{g}_b = (1 - \omega_b) g_b + \omega_b \cdot \tfrac{g_b}{1 + \gamma c_b}. \tag{10}$$

Algorithm 1 describes how the optimization is performed via two-step backward pass ensuring curvature-regularized updates for the fusion encoder while keeping risk estimation unbiased. Full pseudocode in the *Supplemental*.

3.6. Datasets and Settings

Datasets and Evaluation. We evaluate our method on three publicly available cancer datasets from the Clinical Proteomic Tumor Analysis Consortium (CPTAC)[2] and The Cancer Genome Atlas (TCGA)[3]: CPTAC-PDA (Pancreatic Ductal Adenocarcinoma), CPTAC-UCEC (Uterine Corpus Endometrial Carcinoma), and the combined CPTAC-CCRCC and TCGA-KIRC (via MMIST-ccRCC [24]) for Clear Cell Renal Cell Carcinoma. Each dataset undergoes 5-fold cross-validation. Performance is measured using the concordance index (C-index), with patient stratification assessed via Kaplan–Meier survival curves [16] and the Log-rank test [22].

Implementation. For each WSI, we use the Transformer-based TITAN [7] image encoder as the pathology backbone, and the MedImageInsight [5] model for radiology. Each MHCA block employs four attention heads with pre-LayerNorm applied to both attention and MLP modules, following standard Transformer design. Models are trained with AdamW and a cosine annealing learning rate schedule. Further implementation and hyperparameter details are provided in the *Supplemental*.

[2] https://proteomic.datacommons.cancer.gov/pdc/
[3] https://portal.gdc.cancer.gov/

Table 1. C-index comparison ($\mu \pm \sigma$) with state-of-the-art methods across three cancer datasets. The multimodal panels show results for models using individual and combined modalities, respectively. (**) and (*) indicate 0.001 and 0.05 p-values significant difference.

Method	Multimodal	CPTAC-PDA		CPTAC-UCEC		MMIST-ccRCC	
		Rad.	Path.	Rad.	Path.	Rad.	Path.
CoxPH	×	0.6898±0.0237**	**0.7743±0.0539****	0.6251±0.0569**	0.6962±0.0558**	0.4677±0.0843**	0.6443±0.0948**
RankSVMs	×	0.7026±0.0226**	0.7238±0.0618*	0.7517±0.0237**	0.7128±0.0435**	0.5927±0.0721**	**0.6636±0.0853****
RSF	×	0.7523±0.0294**	0.7242±0.0717*	0.7640±0.0348**	0.7183±0.0416**	0.6137±0.0865**	0.6564±0.0628**
GBoost	×	0.7192±0.0526**	0.7310±0.0825**	0.7936±0.0365**	0.7312±0.0337**	0.6296±0.0752*	0.6644±0.0861**
PARACHUTE	×	**0.7813±0.1268**	0.7001±0.0623	**0.8182±0.0643**	**0.7773±0.0495**	**0.6367±0.1858**	0.5979±0.1536
Method	**Multimodal**	Full		Full		Full	
CoxPH	✓	0.6114±0.0674**		0.4975±0.0231**		0.5653±0.1898**	
M4Survive [18]	✓	0.5851±0.1188**		0.6831±0.1517**		0.6863±0.1021**	
Pathomic Fusion [3]	✓	0.5712±0.0812**		0.6812±0.1027**		0.6611±0.0742**	
CA-MLIF [1]	✓	0.7242±0.0931**		0.7006±0.0981**		0.5549±0.0461**	
SMuRF [31]	✓	0.7225±0.0627**		0.7213±0.0712**		0.6767±0.1094**	
PARACHUTE	✓	**0.7848±0.1041**		**0.8367±0.0805**		**0.7323±0.1106**	

Table 2. C-index comparison ($\mu \pm \sigma$) with state-of-the-art methods under missing modality conditions across three cancer datasets. The missing ratio panels indicate different missing modality percentages. (**) and (*) indicate 0.001 and 0.05 p-values significant difference.

Missing Ratio	0%	5%			15%			30%		
Modalities	Full	Rad.	Path.	Mixed	Rad.	Path.	Mixed	Rad.	Path.	Mixed
CPTAC-PDA Dataset										
CoxPH	0.6114±0.0674**	0.5511±0.0835**	0.5965±0.0920**	0.5892±0.0950**	0.5777±0.1122**	0.5910±0.0732**	0.6137±0.0664**	0.5820±0.1384**	0.5786±0.1336**	0.6081±0.1110*
DRIM [30]	0.6720±0.0840**	0.6727±0.0948**	0.6696±0.0881**	0.6742±0.0898**	0.6904±0.0877**	0.7081±0.1037**	0.6983±0.0525	0.6201±0.0716*	0.6072±0.0760**	0.6967±0.0644**
AttentionMOI [26]	0.6923±0.0816**	0.6612±0.0817**	0.6812±0.0718**	0.6928±0.0812**	0.7012±0.0682**	0.7822±0.0927**	0.6721±0.0626*	0.6221±0.0639*	0.6426±0.0726**	0.6028±0.0352*
MMD [6]	0.7112±0.0812**	0.7219±0.0561**	0.7317±0.0517**	0.7113±0.8131**	0.6912±0.0316**	0.6816±0.0552**	0.6713±0.0651**	0.6112±0.0416**	0.6317±0.0612**	0.5818±0.0581**
ShaSpec [33]	0.7315±0.0332**	0.7418±0.0233**	0.7319±0.0247**	0.7419±0.0518**	0.7229±0.0319*	0.6929±0.0423**	0.6806±0.0723*	0.6228±0.0213**	0.6528±0.0428*	0.6139±0.0212*
PARACHUTE	**0.7848±0.1041**	**0.8006±0.0949**	**0.7828±0.1237**	**0.7676±0.1315**	**0.7335±0.1398**	**0.7381±0.1146**	**0.6871±0.1074**	**0.6324±0.0585**	**0.6630±0.0622**	**0.6196±0.0485**
CPTAC-UCEC Dataset										
CoxPH	0.4975±0.0231**	0.5397±0.1585**	0.5226±0.1356**	0.5347±0.1410**	0.5860±0.1706**	0.6084±0.1505**	0.5750±0.1710**	0.4968±0.0866**	0.5142±0.0969**	0.4793±0.0551**
DRIM [30]	0.6972±0.0253**	0.6695±0.0414**	0.6832±0.0290**	0.6966±0.0242**	0.6748±0.0504**	0.6836±0.0432**	0.6861±0.0483**	0.6920±0.0281**	0.7115±0.0278**	0.7123±0.0253**
AttentionMOI [26]	0.7033±0.0137**	0.6852±0.0526**	0.6912±0.0317**	0.7052±0.0313**	0.6712±0.0416**	0.6917±0.0516**	0.6812±0.0312**	0.7012±0.0311**	0.7012±0.0452**	0.7056±0.0359**
MMD [6]	0.7232±0.0518**	0.7013±0.0734**	0.6937±0.0719**	0.7134±0.0328**	0.6842±0.0713**	0.7013±0.0617**	0.6712±0.0513**	0.6832±0.0563**	0.7019±0.0631**	0.6937±0.0613**
ShaSpec [33]	0.7536±0.0417**	0.7312±0.0429**	0.7422±0.0518**	0.7512±0.0431**	0.6987±0.0617**	0.7312±0.0512**	0.7128±0.0311**	0.7313±0.0313**	0.7611±0.0612**	0.7012±0.0756**
PARACHUTE	**0.8367±0.0805**	**0.7865±0.0986**	**0.8020±0.0952**	**0.8273±0.0842**	**0.7546±0.0366**	**0.7974±0.0481**	**0.7588±0.0492**	**0.7647±0.1168**	**0.7947±0.0770**	**0.7488±0.0839**
MMIST-ccRCC Dataset										
CoxPH	0.5653±0.1898**	0.4907±0.1158**	0.4845±0.1907**	0.4651±0.1515**	0.4767±0.1250**	0.5307±0.1770**	0.4103±0.1371**	0.4894±0.1725**	0.5436±0.1821**	0.5805±0.1641**
DRIM [30]	0.6985±0.0451**	0.6113±0.0451**	0.6524±0.0760**	0.6244±0.0515**	0.7113±0.0451**	0.6746±0.0941**	0.6985±0.0451*	0.6144±0.0563**	0.6078±0.0636**	0.6149±0.0622**
AttentionMOI [26]	0.6243±0.0626**	0.6132±0.0861**	0.5867±0.0710**	0.6001±0.0751**	0.5777±0.0757**	0.6129±0.0600**	0.5927±0.0733**	0.6001±0.0743**	0.5610±0.0826**	0.5987±0.0712**
MMD [6]	0.7072±0.0528**	0.6801±0.0791*	0.6701±0.0749*	0.6820±0.0728	0.6903±0.0623**	0.6711±0.0801**	0.6763±0.0688**	0.6255±0.0729*	0.7163±0.0891**	0.6302±0.0744**
ShaSpec [33]	0.7234±0.0626**	0.6822±0.0518	0.6707±0.0720*	0.6734±0.0810*	0.6969±0.0719**	0.6826±0.0836**	0.7044±0.0791*	0.6137±0.0908**	0.6471±0.0822**	0.6399±0.0854*
PARACHUTE	**0.7436±0.0721**	**0.6830±0.0840**	**0.6862±0.0901**	**0.6855±0.0934**	**0.7131±0.0957**	**0.7093±0.1179**	**0.7160±0.0886**	**0.6370±0.0706**	**0.7454±0.1043**	**0.6485±0.0909**

4. Experiments

4.1. Comparison with State-Of-The-Art

To evaluate our model under both complete and incomplete modality settings, we compare against state-of-the-art unimodal and multimodal survival prediction methods. Unimodal baselines include linear Cox Proportional Hazards model (CoxPH), Support Vector Machine for Ranking (RankSVM), Random Survival Forests (RSF), and Gradient Boosted Models (GBoost), applied separately to radiology and histopathology features from the same pretrained foundation models. Multimodal baselines include SMuRF [31], M4Survive [18], Pathomic Fusion [3], and CA-MLIF [1]. For missing modality scenarios, we benchmark against: (i) reconstruction-based methods like MMD [6] (mean fusion with reconstruction loss); (ii) disentanglement-based methods such as DRIM [30] and ShaSpec [33], which decompose embeddings into modality-specific and shared components; and (iii) static fusion methods like AttentionMOI [26], which applies attention-based integration on denoised features. All models use identical pretrained features to ensure fair comparison across fusion and modality-missing settings and a paired t-test for statistical analysis.

Compared with Unimodal Models. Table 1 highlights the consistent benefit of integrating radiology and histology, as our method notably achieves superior performance compared to models using only radiology or histopathology data on all the datasets. In particular, we demonstrate multimodal uplift over the 5.38% strongest pathology-only model for PDA, and over the best radiology-based approaches by 4.31% and 10.27% in UCEC and CCRCC respectively, emphasizing the strength of cross-modal fusion.

Compared with Multimodal Models. As shown in Table 1, PARACHUTE consistently outperforms all multimodal baselines across the selected cancer datasets, achieving gains up to 11.54%–13.61% in UCEC over the strongest SMuRF and CA-MLIF multimodal competitors. These results highlight the model's ability to translate fine-grained cross-modal cues, such as spatially localized patterns in histopathology and radiology, into globally predictive representations. In PDA, PARACHUTE leverages local features like abrupt ductal cutoff and stromal desmoplasia to resolve ambiguous global tumor boundaries, resulting in a 6.06% improvement over CA-MLIF leading method. For UCEC, it effectively fuses heterogeneous histological textures with organ-level radiological features. In CCRCC, where multimodal signatures are highly variable, the dynamic embedding adapts to cross-scale inconsistencies, yielding a 4.60% gain over M4Survive best baseline. In addition, unlike naïve concatenation or other static fusion strategies,

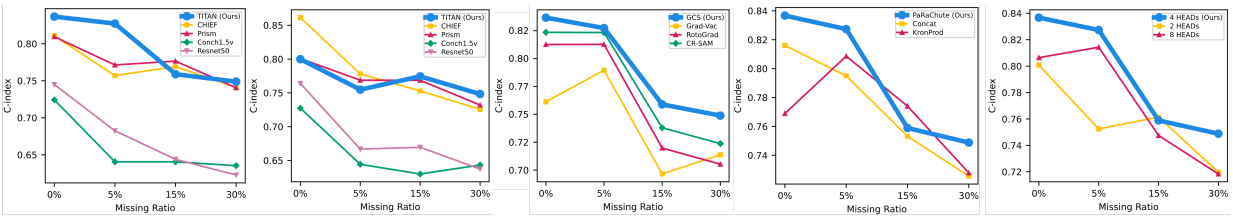

Figure 2. C-index robustness ablations under varying missing modality ratios by: (a) changing WSI encoders while keeping MedImageInsight and (b) MedSAM2 [20] radiology encoders fixed; and by varying (c) gradient modulation, (d) fusion, and (e) MHCA schemes.

Table 3. Ablation study under complete and missing modality scenarios across the PDA, UCEC and CCRCC cohorts. (**) and (*) indicate 0.001 and 0.05 p-values significant difference.

Components		Missing Ratio			
DCE	GCS	0%	5%	15%	30%
CPTAC-PDA Dataset					
×	×	0.7799±0.0960*	0.7311±0.0819**	0.6543±0.0859**	0.5721±0.0637**
×	✓	0.7648±0.0872**	0.7419±0.0729**	0.6410±0.0833**	0.5936±0.0623**
✓	×	0.7822±0.0845	0.7453±0.0873**	0.6629±0.0745**	0.6123±0.0562
✓	✓	**0.7848±0.1041**	**0.7676±0.1315**	**0.6871±0.1074**	**0.6196±0.0485**
CPTAC-UCEC Dataset					
×	×	0.7569±0.0288**	0.6765±0.0678**	0.6801±0.0654**	0.6779±0.0426**
×	✓	0.7667±0.0406**	0.6877±0.0575**	0.6890±0.0484**	0.6753±0.0633**
✓	×	0.7788±0.0718**	0.7815±0.0551**	0.7609±0.0492*	0.7556±0.0754**
✓	✓	**0.8367±0.0805**	**0.8053±0.0463**	**0.7703±0.0446**	**0.7694±0.0926**
MMIST-ccRCC Dataset					
×	×	0.6448±0.1721**	0.6512±0.1046**	0.6330±0.1022**	0.6136±0.1089**
×	✓	0.6543±0.1757**	0.6590±0.1093**	0.6261±0.1006**	0.6292±0.0999**
✓	×	0.5883±0.1962**	0.6799±0.1179*	0.6832±0.1026**	**0.7006±0.0847****
✓	✓	**0.7436±0.0721**	**0.6855±0.0934**	**0.7160±0.0886**	0.6485±0.0909

PARACHUTE's local-to-global bias transfer adapts to the joint presence of high-dimensional inputs, mitigating overfitting. DCE limits over-reliance on modality-specific noise by reducing effective capacity, akin to attention dropout, while GCS promotes generalization through loss landscape smoothing, even under complete input conditions.

Compared with Baseline Models for Missing Modality. PARACHUTE achieves state-of-the-art performance under all missingness scenarios, with substantial gains particularly in PDA and UCEC (Table 2). Notably, in PDA, PARACHUTE obtains a C-index of 0.8006 even when radiology is missing (5%), outperforming the full-modality baseline (0.7848) and all competing methods. This counterintuitive improvement highlights modality dominance in PDA, where histopathology conveys most of the prognostic signal (e.g., stromal desmoplasia), while radiology may inject modality-specific noise (e.g., poor tumor demarcation), reducing fusion effectiveness in naive baselines. In contrast, PARACHUTE's dynamic contextual embedding and curvature-aware optimization prevent overfitting to dominant modalities by modulating feature relevance, preserving robustness even when fusing partially informative inputs. In UCEC, where both modalities contribute complementary but heterogeneous signals, PARACHUTE demonstrates balanced resilience: it retains a high C-index under 15% missing data (e.g., 0.7974 with missing pathology, outperforming ShaSpec at 0.7312), and sustains 0.8273 under mild 5% mixed dropouts. Overall performance in UCEC is higher than PDA, reflecting differences in cancer phenotype: UCEC presents clearer spatial correspondence between radiology and histology (e.g., endometrial thickening and glandular architecture), enabling more consistent cross-modal alignment and stronger global survival signals. In CCRCC, where modality contributions vary widely across patients due to heterogeneous tumor architecture and inconsistent radiological presentation, PARACHUTE again proves effective. It achieves a 0% missingness C-index of 0.7436, benefits from multimodal uplift under partial dropout (0.7160 with 30% missing), and remains robust under severe dropout (0.6485 with 30% missing), outperforming DRIM and ShaSpec by up to 3.36%. PARACHUTE adapts to available modalities and fuses only useful cues, unlike static methods prone to misleading inputs. These results validate PARACHUTE's ability to dynamically adjust to modality presence, and fuse informative cues without being misled by spurious or dominant inputs, traits that static and modality-agnostic fusion strategies fail to exhibit under real-world missingness.

4.2. Ablation Study

To evaluate the individual contributions of the proposed components, we conduct a series of ablation experiments by selectively removing the DCE and GCS modules from the PARACHUTE architecture. As shown in Table 3, removing either DCE or GCS results in consistent performance degradation across most ablative levels. For example, on the UCEC dataset with 15% missingness, removing DCE drops the C-index of ∼8.13% (0.7703±0.0446 *v.s.* 0.6890±0.0484). Similarly, excluding GCS reduces performance to 0.7609±0.0492 at the same masking level, again reflecting a drop of ∼0.94%. When both modules are ablated, the impact becomes more pronounced: the C-index at 30% masking falls from 0.7694±0.0926 to 0.6779±0.0426, marking a decline of over 9.15%. An exception is observed in the ccRCC dataset, where DCE alone underperforms at 0% missingness and GCS slightly lowers performance at 30%; however, these effects are dataset-specific and align with known modality dominance in ccRCC, as the

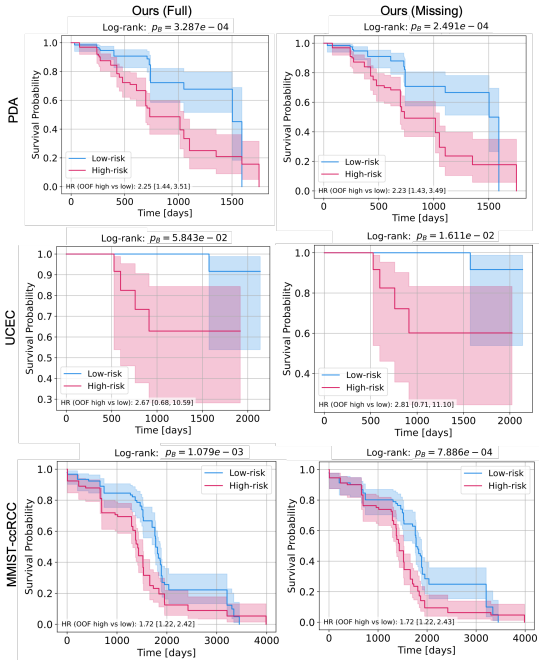

Figure 3. Kaplan–Meier analysis of predicted high-risk and low-risk groups under both complete (*Full*) and missing modality (*Missing*) settings. Shaded areas indicate confidence intervals.

full DCE+GCS configuration still yields the strongest overall results. These results confirm that DCE and GCS offer complementary gains: DCE improves cross-modal alignment via sample-spatial bias, while GCS stabilizes optimization by smoothing gradients under high missingness.

4.2.1. Component-Specific Analysis

To assess each component's impact on robustness under missing data, we report the ablations on the UCEC dataset as an example (Figure 2), varying backbone encoders, fusion schemes, gradient modulation, and hazard heads.

Radiology Encoders. While MedSAM2 (Figure 2(b)) marginally outperforms MedImageInsight (Figure 2(a)) in the full-data setting (0.8577±0.0621 *v.s.* 0.8372±0.0810, it is less resilient under missing data. At 30% masking, MedSAM2 drops to 0.7021±0.0635, whereas MedImageInsight maintains 0.7650±0.1170, emphasizing the importance of stable feature extraction over peak full-data performance.

Histopathology Encoders. Slide-level WSI encoders (TITAN, CHIEF, PRISM) outperform patch-based ones (CONCHv1.5, ResNet50), especially under high missingness. At 30% masking, TITAN reaches 0.7483±0.0515 *v.s.* 0.6433±0.0463 (CONCHv1.5) and 0.6373±0.0399 (ResNet50), underscoring the value of slide context. While CHIEF exceeds TITAN with full data (0.8613±0.0744 *v.s.* 0.7998±0.0790), its performance drops faster with masking, confirming TITAN's robustness as our default encoder.

Gradient Modulation. Our Gradient Curvature Steering (GCS) enhances robustness to missing modalities, outperforming RotoGrad and CR-SAM on CPTAC-UCEC (Figure 2(c)). It achieves the best C-index with full data (0.8367±0.0805) and sustains strong performance at 15% and 30% missingness (0.7974±0.0482, 0.7947±0.0770). By smoothing the loss landscape, GCS stabilizes training as incompleteness rises, proving effective for real-world multimodal survival prediction. Detailed loss morphology analysis is provided in the *Supplemental*.

Fusion Mechanism. Our hierarchical MHCA-based fusion surpasses simpler alternatives (Figure 2(d)). At 15% masking, naive concatenation results in a lower C-index (0.7400±0.0550), while Kronecker-product fusion performs slightly better (0.7490±0.0530) but always remaining worse than MHCA (0.7550±0.0370). This suggests that attention-based fusion better captures complex cross-modal relationships than static operators.

Attention Heads. Varying the number of heads in the MHCA reveals that four heads strike the best trade-off between capacity and generalization (Figure 2(e)). Reducing to two results in slightly lower performance at 15% masking (0.7420±0.0580), while increasing to eight provides negligible benefits under both complete and partial inputs.

4.3. Patient Stratification

In addition to evaluating prognosis with the concordance index (C-index), a key goal is patient stratification, separating individuals into risk groups with significantly different outcomes to guide personalized treatment. In Figure 3, we assess stratification performance under both complete and missing-modality settings from Out-Of-Fold (OOF) risk predictions. Across all three cohorts, our method consistently distinguishes groups with significant survival differences, supported by low log-rank p-values adjusted via Bonferroni correction. We further report statistically significant Hazard Ratios (HR) with 95% confidence intervals, confirming the robustness and clinical relevance.

5. Conclusion

In this paper, we introduce PARACHUTE, a multimodal survival prediction framework designed to robustly fuse radiology and histopathology through Dynamic Contextual Embedding (DCE) and Gradient Curvature Steering (GCS). These components enable the model to align local semantic cues and adapt to incomplete inputs by transferring modality-aware spatial signals into the global prediction process. Extensive experiments across three cancer types, PDA, UCEC, and CCRCC, demonstrate that PARACHUTE consistently outperforms existing approaches under both complete and missing modality settings, highlighting its effectiveness, adaptability, and biological interpretability.

Acknowledgements

We acknowledge NVIDIA Corporation for the Academic Hardware Grant Program, ISCRA for awarding this project access to the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CINECA (Italy), and A. Zeni for valuable suggestions and discussions.

References

[1] Yajun An, Jiale Chen, Huan Lin, Zhenbing Liu, Siyang Feng, Hualong Zhang, Rushi Lan, Zaiyi Liu, and Xipeng Pan. Camlif: Cross-attention and multimodal low-rank interaction fusion framework for tumor prognostic prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1764–1772, 2025. 1, 2, 6

[2] Lars Björnebo, Andrea Discacciati, Ugo Falagario, Hari T Vigneswaran, Fredrik Jäderling, Henrik Grönberg, Martin Eklund, Tobias Nordström, and Anna Lantz. Biomarker vs mri-enhanced strategies for prostate cancer screening: the sthlm3-mri randomized clinical trial. *JAMA Network Open*, 7(4):e247131–e247131, 2024. 1

[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. 2, 6

[4] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pages 794–803. PMLR, 2018. 2

[5] Noel CF Codella, Ying Jin, Shrey Jain, Yu Gu, Ho Hin Lee, Asma Ben Abacha, Alberto Santamaria-Pang, Will Guyman, Naiteek Sangani, Sheng Zhang, et al. Medimageinsight: An open-source embedding model for general domain medical imaging. *arXiv preprint arXiv:2410.06542*, 2024. 3, 5

[6] Can Cui, Han Liu, Quan Liu, Ruining Deng, Zuhayr Asad, Yaohong Wang, Shilin Zhao, Haichun Yang, Bennett A Landman, and Yuankai Huo. Survival prediction of brain cancer with incomplete radiology, pathology, genomic, and demographic data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 626–635. Springer, 2022. 2, 6, 7

[7] Tong Ding, Sophia J Wagner, Andrew H Song, Richard J Chen, Ming Y Lu, Andrew Zhang, Anurag J Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, et al. A multimodal whole-slide foundation model for pathology. *Nature Medicine*, pages 1–13, 2025. 3, 5

[8] Aiman Farooq, Deepak Mishra, and Santanu Chaudhury. Survival prediction in lung cancer through multi-modal representation learning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3907–3915. IEEE, 2025. 2

[9] Bo Feng and Gang Wu. A block lanczos method for large-scale quadratic minimization problems with orthogonality constraints. *SIAM Journal on Scientific Computing*, 46(2): A884–A905, 2024. 3

[10] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020. 2

[11] Karolina Frankowska, Michał Zarobkiewicz, Izabela Dabrowska, and Agnieszka Bojarska-Junak. Tumor infiltrating lymphocytes and radiological picture of the tumor. *Medical Oncology*, 40(6):176, 2023. 1

[12] Fabian Hörst, Moritz Rempe, Lukas Heine, Constantin Seibold, Julius Keyl, Giulia Baldini, Selma Ugurel, Jens Siveke, Barbara Grünwald, Jan Egger, et al. Cellvit: Vision transformers for precise cell segmentation and classification. *Medical Image Analysis*, 94:103143, 2024. 7

[13] Adrian Javaloy and Isabel Valera. Rotograd: Gradient homogenization in multitask learning. In *The Tenth International Conference on Learning Representations (Virtual)*, 2022. 2

[14] Albert Juan Ramon, Chaitanya Parmar, Oscar M Carrasco-Zevallos, Carlos Csiszer, Stephen SF Yip, Patricia Raciti, Nicole L Stone, Spyros Triantos, Michelle M Quiroz, Patrick Crowley, et al. Development and deployment of a histopathology-based deep learning algorithm for patient prescreening in a clinical trial. *Nature Communications*, 15 (1):4690, 2024. 1

[15] Jessie D Kang, Sharon E Clarke, and Andreu F Costa. Factors associated with missed and misinterpreted cases of pancreatic ductal adenocarcinoma. *European Radiology*, 31(4): 2422–2432, 2021. 1

[16] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958. 5

[17] Azar Kianzad, Lilian J Meijboom, Esther J Nossent, Eva Roos, Bernadette Schurink, Peter I Bonta, Inge AH van den Berk, Rieneke Britstra, Jaap Stoker, Anton Vonk Noordegraaf, et al. Covid-19: Histopathological correlates of imaging patterns on chest computed tomography. *Respirology*, 26 (9):869–877, 2021. 1

[18] Ho Hin Lee, Alberto Santamaria-Pang, Jameson Merkov, Matthew Lungren, and Ivan Tarapov. Multi-modal mamba modeling for survival prediction (m4survive): Adapting joint foundation model representations. *arXiv preprint arXiv:2503.10057*, 2025. 2, 6

[19] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30:863–874, 2024. 3

[20] Jun Ma, Zongxin Yang, Sumin Kim, Bihui Chen, Mohammed Baharoon, Adibvafa Fallahpour, Reza Asakereh, Hongwei Lyu, and Bo Wang. Medsam2: Segment anything in 3d medical images and videos. *arXiv preprint arXiv:2504.03600*, 2025. 7

[21] 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. 1

[22] Nathan Mantel et al. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*, 50(3):163–170, 1966. 5

[23] Yoichi Miyata, Naoto Yonamine, Ibuki Fujinuma, Takazumi Tsunenari, Yasuhiro Takihata, Hiroyuki Hakoda, Akiko Nakazawa, Toshimitsu Iwasaki, Takahiro Einama, Junichi Togashi, et al. Impact of preoperative tumor size on prognosis of resectable and borderline resectable pancreatic ductal adenocarcinomas. *Annals of Surgical Oncology*, 30(13): 8621–8630, 2023. 1

[24] Tiago Mota, M Rita Verdelho, Diogo J Araújo, Alceu Bissoto, Carlos Santiago, and Catarina Barata. Mmist-ccrcc: A real world medical dataset for the development of multimodal systems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2395–2403, 2024. 5

[25] Susmita Palmal, Sriparna Saha, Nikhilanad Arya, and Somanath Tripathy. Cagcl: Predicting short-and long-term breast cancer survival with cross-modal attention and graph contrastive learning. *IEEE Journal of Biomedical and Health Informatics*, 2024. 2

[26] Jiali Pang, Bilin Liang, Ruifeng Ding, Qiujuan Yan, Ruiyao Chen, and Jie Xu. A denoised multi-omics integration framework for cancer subtype classification and survival prediction. *Briefings in Bioinformatics*, 24(5), 2023. 6, 7

[27] Hania Paverd, Konstantinos Zormpas-Petridis, Hannah Clayton, Sarah Burge, and Mireia Crispin-Ortuzar. Radiology and multi-scale data integration for precision oncology. *NPJ Precision Oncology*, 8(1):158, 2024. 1

[28] Xuejun Qian, Jing Pei, Chunguang Han, Zhiying Liang, Gaosong Zhang, Na Chen, Weiwei Zheng, Fanlun Meng, Dongsheng Yu, Yixuan Chen, et al. A multimodal machine learning model for the stratification of breast cancer risk. *Nature Biomedical Engineering*, 9(3):356–370, 2025. 1

[29] Saima Rathore, Ahmad Chaddad, Muhammad A Iftikhar, Michel Bilello, and Ahmed Abdulkadir. Combining mri and histologic imaging features for predicting overall survival in patients with glioma. *Radiology: Imaging Cancer*, 3(4): e200108, 2021. 1

[30] Lucas Robinet, Ahmad Berjaoui, Ziad Kheil, and Elizabeth Cohen-Jonathan Moyal. Drim: Learning disentangled representations from incomplete multimodal healthcare data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 163–173. Springer, 2024. 1, 2, 6, 7

[31] Bolin Song, Amaury Leroy, Kailin Yang, Tanmoy Dam, Xiangxue Wang, Himanshu Maurya, Tilak Pathak, Jonathan Lee, Sarah Stock, Xiao T Li, et al. Deep learning informed multimodal fusion of radiology and pathology to predict outcomes in hpv-associated oropharyngeal squamous cell carcinoma. *Ebiomedicine*, 114, 2025. 1, 2, 6

[32] Pranjal Vaidya, Mohammadhadi Khorrami, Kaustav Bera, Pingfu Fu, Lukas Delasos, Amit Gupta, Cristian Barrera, Nathan A Pennell, Vamsidhar Velcheti, and Anant Madabhushi. Computationally integrating radiology and pathology image features for predicting treatment benefit and outcome in lung cancer. *npj Precision Oncology*, 9(1):1–10, 2025. 1

[33] 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. 1, 2, 6, 7

[34] Pengyu Wang, Huaqi Zhang, Meilu Zhu, Xi Jiang, Jing Qin, and Yixuan Yuan. Mgiml: Cancer grading with incomplete radiology-pathology data via memory learning and gradient homogenization. *IEEE Transactions on Medical Imaging*, 43 (6):2113–2124, 2024. 1, 2

[35] Zirui Wang, Yulia Tsvetkov, Orhan Firat, and Yuan Cao. Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models. *arXiv preprint arXiv:2010.05874*, 2020. 2

[36] David Weller, Usha Menon, Alina Zalounina Falborg, Henry Jensen, Andriana Barisic, Anne Kari Knudsen, Rebecca J Bergin, David H Brewster, Victoria Cairnduff, Anna T Gavin, et al. Diagnostic routes and time intervals for patients with colorectal cancer in 10 international jurisdictions; findings from a cross-sectional study from the international cancer benchmarking partnership (icbp). *BMJ open*, 8(11): e023870, 2018. 1

[37] Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. *Advances in neural information processing systems*, 31, 2018. 1

[38] Tao Wu, Tie Luo, and Donald C Wunsch II. Cr-sam: Curvature regularized sharpness-aware minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6144–6152, 2024. 3

[39] Yingxue Xu and Hao Chen. Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 21241–21251, 2023. 2

[40] Xiaochun Yi, Xiaoping Yu, Congrui Li, Junjian Li, Hui Cao, Qiang Lu, Junjun Li, and Jing Hou. Deep learning radiopathomics based on pretreatment mri and whole slide images for predicting over survival in locally advanced nasopharyngeal carcinoma. *Radiotherapy and Oncology*, page 110949, 2025. 1

[41] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33:5824–5836, 2020. 2

[42] Andrew Zhang, Guillaume Jaume, Anurag Vaidya, Tong Ding, and Faisal Mahmood. Accelerating data processing and benchmarking of ai models for pathology. *arXiv preprint arXiv:2502.06750*, 2025. 1

[43] 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

[44] Xinliang Zhu, Jiawen Yao, Feiyun Zhu, and Junzhou Huang. Wsisa: Making survival prediction from whole slide histopathological images. In *Proceedings of the IEEE con-*

ference on computer vision and pattern recognition, pages 7234–7242, 2017. 2

[45] Anoek Zomer, Davide Croci, Joanna Kowal, Leon van Gurp, and Johanna A Joyce. Multimodal imaging of the dynamic brain tumor microenvironment during glioblastoma progression and in response to treatment. *Iscience*, 25(7), 2022. 1