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

Supplementary Material

1. Implementation Details and Hyperparameter Configuration

Experimental Setup. The codebase is implemented in PyTorch (`2.5.1v`), training and evaluations are executed on a AMD Ryzen 7 5800X @3.8 GHz with a 24 GB NVIDIA RTX A5000 GPU, while foundation models feature extraction on a 40 GB NVIDIA A100 GPU.

Feature Extraction and Benchmarking Protocols. Feature extraction for radiology was conducted using the official codebases from the MedImageInsight[4] and MedSAM2[5] repositories. For histopathology, we employed Trident[6], a comprehensive package for WSI processing that supports state-of-the-art patch-level and slide-level foundation models, and was used for both tissue segmentation and feature extraction. To standardize foundation model benchmarking, we utilized Patho-Bench[7][42], a library designed for evaluating foundation models under multiple strategies, along with its canonical data splits, which include labels for 42 clinically relevant pathology tasks. Specifically, we adopted the standard train-test splits provided by Patho-Bench for the CPTAC-PDA and CPTAC-UCEC datasets, and we performed 5-fold cross-validation on the training set and evaluated on the independent test set provided released with the MMIST-ccRCC[8] dataset. Thus, all evaluations rely on publicly released splits whenever available, ensuring fair benchmarking and reproducibility. We also report the number of patients per dataset in Table 4. Note that these counts reflect a post-cleaning step, retaining only patients with at least one image case consistently matching to both survival times and censoring information.

Data Masking Strategy. To evaluate model performance under both full and partial modality availability, a comprehensive training protocol was designed with controlled masking scenarios. This approach enables consistent comparison across training conditions and robust estimation of model generalization under realistic missing modality settings. In the full-modality setting, no runtime masking is applied. To assess the model's ability to handle incomplete data, a masked training regime is adopted in which modalities are independently masked at runtime with probability $p_{\text{miss}} \in \{0.05, 0.15, 0.30\}$. The adopted masking strategies are:

- **Radiology Image-missing:** only the histopathology modality is retained;
- **WSI-missing:** only the radiology modality is retained;
- **Mixed-missing:** both modalities are randomly masked with the same probability applied in alternation to simulate missing modality scenarios.

In particular, each model is trained under four configurations:

1. Full-modality training ($p_{\text{miss}} = 0$) with full-modality evaluation;
2. Mixed-modality training with $p_{\text{miss}} = 0.05$, evaluated under Radiology Image-missing, WSI-missing, and mixed missing at 5%;
3. Mixed-modality training with $p_{\text{miss}} = 0.15$, evaluated under Radiology Image-missing, WSI-missing, and mixed missing at 15%;
4. Mixed-modality training with $p_{\text{miss}} = 0.30$, evaluated under Radiology Image-missing, WSI-missing, and mixed missing at 30%.

To reduce the stochastic variance introduced by runtime masking, each evaluation is repeated $n = 30$ times with independently sampled masks, and the resulting metrics are averaged. This strategy yields a more statistically robust estimate of model performance under each missingness condition.

Hyperparameter Setting. The proposed network operates on fixed-size modality tokens extracted via adapter networks and contextual embedding mechanisms. These are then fused using the MHCA fusion transformer, followed by a multilayer perceptron (hazard net) for risk prediction. The dimensionality of the latent tokens is set to $d = 256$, while intermediate representations in adapter modules are projected to $d_{\text{int}} = 256$. All MLP blocks use ReLU activation, followed by dropout layers with $p = 0.1$. Each MHCA block uses $h = 4$ attention heads, and LayerNorm is applied before attention and MLP modules, fol-

Table 4. Number of total images and patients per image modality and per dataset resulted from dataset cleaning.

Dataset	Nr. Images		Nr. Patients	
	Histo	Rad	Histo	Rad
CPTAC-PDA	553	232	76	34
CPTAC-UCEC	883	192	34	33
MMIST-ccRCC	438	379	601	118

[4] https://huggingface.co/lion-ai/MedImageInsights

[5] https://huggingface.co/wanglab/MedSAM2

[6] https://github.com/mahmoodlab/trident

[7] https://github.com/mahmoodlab/Patho-Bench

[8] https://multi-modal-ist.github.io/datasets/ccRCC/

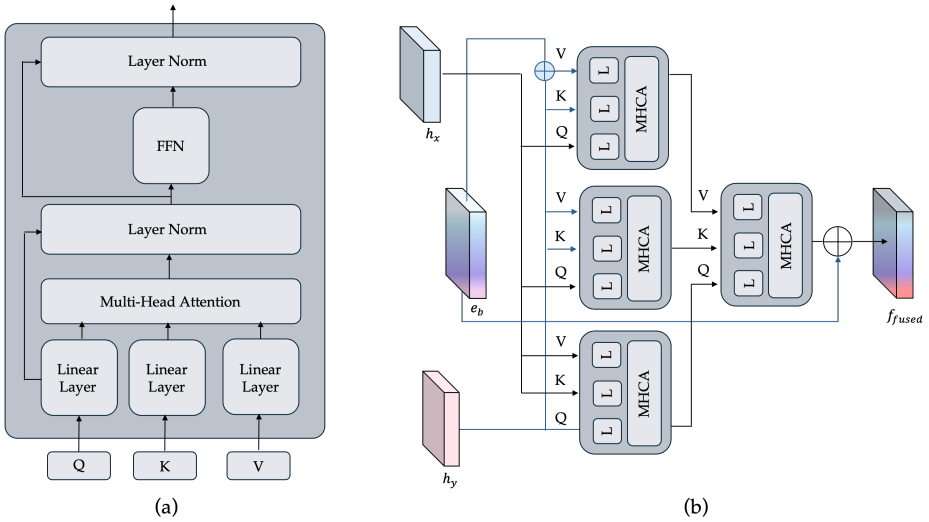

Figure 4. (a) Architecture of the Multi-Head Cross-Attention (MHCA) block. Input features are projected into Query (Q), Key (K), and Value (V) representations, processed through multi-head attention, and refined via residual connections and a Feed-Forward Network (FFN). (b) Hierarchical fusion pipeline comprising four MHCA modules. The first three blocks capture asymmetric cross-modal attention among radiology, histopathology, and contextual embeddings. Their outputs are then aggregated through a final MHCA block, followed by a contextual skip connection for enhanced spatial alignment.

lowing the standard transformer architecture. All models are trained using the AdamW optimizer with a fixed learning rate $\eta_{\max} = 5 \cdot 10^{-3}$ and a weight decay coefficient of 0.01. Learning rate scheduling is implemented via the Cosine Annealing schedule:

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min}) \left(1 + \cos\left(\frac{T_{\text{cur}}}{T_{\max}}\pi\right)\right), \quad (11)$$

where $\eta_{\min} = 9 \cdot 10^{-5}$ and $T_{\max} = 15$. Training is conducted for 800 epochs with a batch size of 16 samples. Each model contains ∼3.2M trainable parameters.

2. Multi-Head Cross Attention Details Visualization

Figure 4 details the architectural design of the Multi-Head Cross-Attention (MHCA) block and its integration within the hierarchical fusion pipeline used for cross-modal learning.

3. Gradient Curvature Steering

Optimization Details. To prevent overfitting and improve generalization, a *Gradient Curvature Steering* (GCS) optimization procedure is introduced, building upon the Sharpness-Aware Minimization (SAM) framework [10]. SAM aims to improve generalization by avoiding sharp minima in the loss landscape, which have been empirically associated with poor test performance. Rather than minimizing only the training loss, SAM performs a min–max optimization, seeking parameters θ that minimize the worst-case loss in a neighborhood of radius ρ:

$$\min_{\theta} \max_{\|\epsilon\|_2 \leq \rho} \mathcal{L}(\theta + \epsilon), \quad (12)$$

where the inner maximization estimates the local sharpness of the loss. The resulting optimization discourages convergence to solutions with high curvature, thus promoting flatter optima and better generalization. However, SAM's standard one-step approximation of the inner maximization can lose effectiveness in regions with high curvature. To address this, GCS augments SAM with a per-sample curvature estimate, enabling the model to dynamically modulate the gradient flow based on local loss geometry and missing modality patterns. The full pseudo-code of GCS optimization procedure is offered in Algorithm 2.

Generalization results. We present full GCS results across datasets against selected baselines to demonstrate the generalization of our method. As shown in Figure 5, our approach remains robust under missing modalities, consistently outperforming all baselines not only on CPTAC-UCEC but also on CPTAC-PDA and MMIST-ccRCC. Notably, at the highest missing-modality ratio, GCS delivers a 4.21% and 4.23% gain over the best-performing CR-SAM and Grad-Vac methods on CPTAC-PDA (Figure 5(b)) and MMIST-ccRCC (Figure 5(c)), respectively, underscoring its strong generalization.

Visual Results Representations. To further interpret the GCS module contribution to the optimization's stability and convergence, Figure 6 presents a visual comparison

Algorithm 2 Gradient Curvature Steering (GCS) Optimization Procedure

Require: Mini-batch $\mathcal{B} = \{(f_b^{(x)}, f_b^{(y)}, t_b, \delta_b, m_b)\}_{b=1}^{B}$ where $f_b^{(x)}$, $f_b^{(y)}$: radiology and histopathology features, t_b: survival time, δ_b: censorship indicator, m_b: missing modality flag

Ensure: Updated model parameters θ (fusion encoder), ϑ (hazard net)

▷ **Step 1: Forward pass**

for $i = 1$ to B **do**

  Compute fused feature: $f_{\text{final}_b} \leftarrow \textsc{Fusion}(f_b^{(x)}, f_b^{(y)}, m_b)$

  Predict risk score: $\hat{h}_b \leftarrow \textsc{HazardNet}(f_{\text{final}_b})$

end for

Compute Cox loss: $\mathcal{L}_{\text{Cox}} \leftarrow -\frac{1}{B}\sum_{b=1}^{B} \delta_b \left(\hat{h}_b - \log \sum_{j:t_j \geq t_b} e^{\hat{h}_j}\right)$

▷ **Step 2: Compute loss gradient w.r.t. fused features**

for $b = 1$ to B **do**

  $g_b \leftarrow \nabla_{f_{\text{final}_b}} \mathcal{L}_{\text{Cox}}$ ▷ Gradient of loss

  $r_b \sim \text{Rademacher}(\pm 1)$

  $c_b \leftarrow \frac{g_b^\top r_b - g_b^\top (-r_b)}{2\varepsilon}$ ▷ Hutchinson's curvature estimate

  $\omega_b \leftarrow \textsc{Controller}(c_b, \|g_b\|_2, m_b)$ ▷ Gating weight

  $\tilde{g}_b \leftarrow (1-\omega_b) g_b + \omega_b \cdot \frac{g_b}{1+\gamma c_b}$ ▷ Steered gradient

end for

▷ **Step 3: First backward pass – steer encoder**

for $i = 1$ to B **do**

  Backpropagate $\tilde{g}_b$ through fusion encoder (keep hazard net fixed)

end for

▷ **Step 4: Second backward pass – update hazard net**

Backpropagate $\mathcal{L}_{\text{Cox}}$ to update ϑ

▷ **Step 5: Optimizer step**

Update parameters θ, ϑ using optimizer

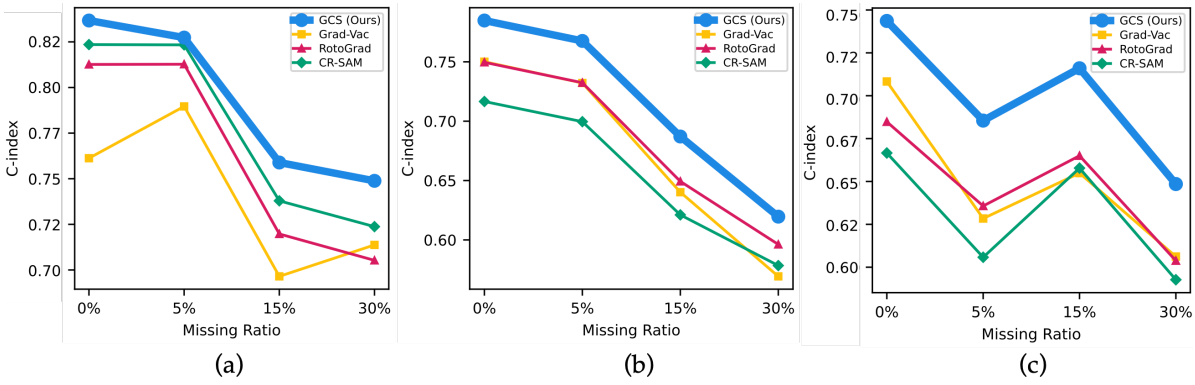

Figure 5. GCS generalization across datasets, showing C-index robustness under increasing missing-modality ratios for (a) CPTAC-UCEC, (b) CPTAC-PDA, and (c) MMIST-ccRCC.

of the loss landscapes obtained from training PARACHUTE with and without the GCS module. In order to produce the loss landscape plot, first the Hessian of the Cox Loss $\mathcal{L}$, is estimated $\nabla_\theta^2 \mathcal{L}$ with respect to the model's learned parameter vector θ, and its two dominant eigenvectors v_1, v_2 are extracted through block-Lanczos algorithm [9]. These two eigen vectors are used to sample perturbations in the directions that most influence the loss value, obtaining $\theta + \alpha v_1 + \beta v_2$, and hence evaluating the loss $\mathcal{L}(\theta + \alpha v_1 + \beta v_2)$ over a grid of values $\alpha, \beta \in [-\epsilon, \epsilon]$ to visualize the landscape around the optimum.

As shown in Figure 6(a), the model trained with GCS converges to a broad and flat region of the loss surface, indicating a more stable and generalizable solution. In contrast, the loss landscape of the GCS-ablated model in Figure 6(b) exhibits sharper curvature and noisier surroundings, sug-

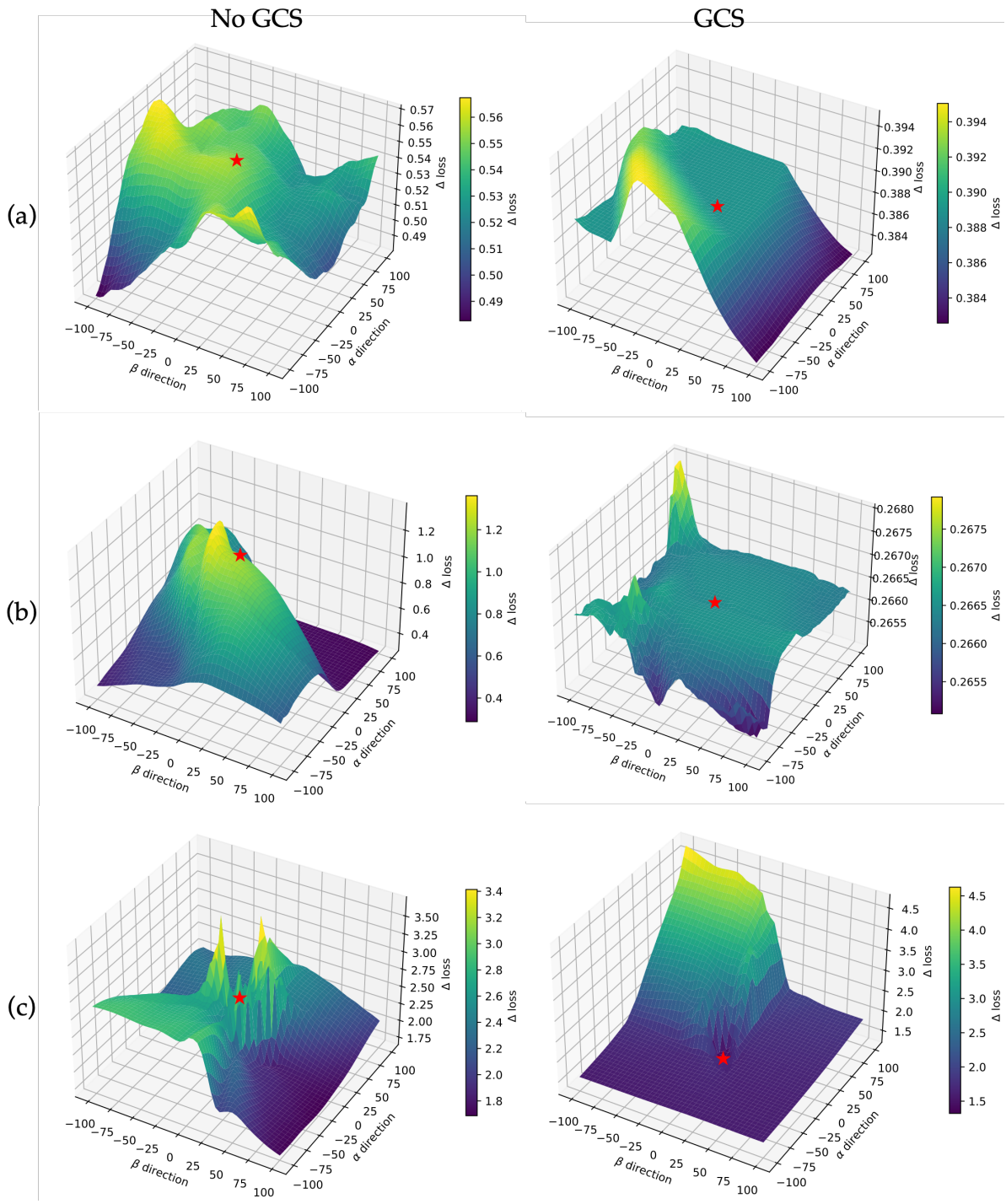

Figure 6. Loss landscape example plots on the CPTAC-PDA (a), CPTAC-UCEC (b) and MMIST-ccRCC (c) datasets. The model enhanced with Gradient Curvature Steering (GCS) converges to flatter minima, whereas its ablated counterpart exhibits sharper curvature and increased noise in the surrounding landscape. (⋆) marks the location of the model's unperturbed optimum.

gesting higher sensitivity to parameter perturbations. This behavior highlights the stabilizing role of the GCS module, which modulates gradient updates using local curvature estimates to avoid sharp minima. These results provide further empirical evidence that the proposed optimization strategy helps the model bypass the noise introduced

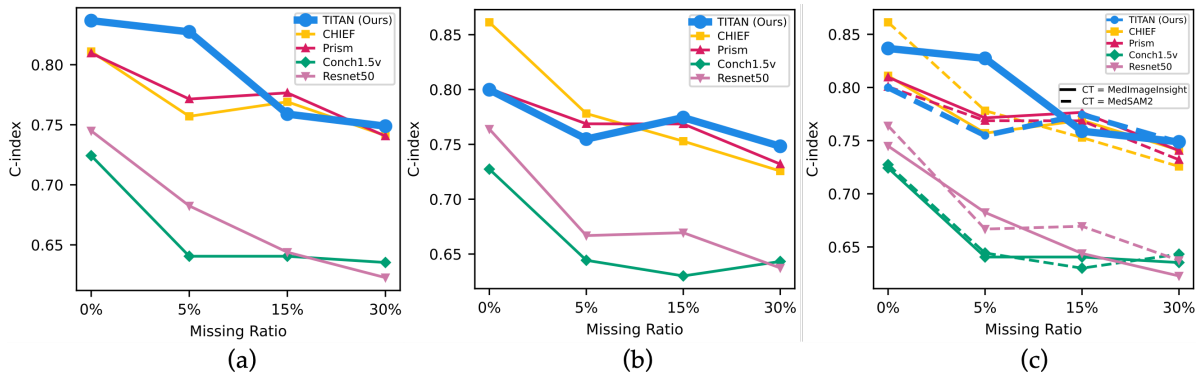

Figure 7. Further ablations assessing C-index robustness under varying missing modality ratios by: (a) changing WSI encoders while keeping MedImageInsight radiology encoder fixed, (b) changing WSI encoders while keeping MedSAM2 radiology encoder fixed and changing both WSI and radiology encoders (c).

by missing modality scenarios during training, ultimately leading to more robust and flatter optima.

4. Naïve Concatenation vs. PARACHUTE

A natural question is whether multimodal survival analysis can be adequately addressed by simply concatenating radiology and histopathology features and fitting a standard Cox model. To test this, we implemented a multimodal CoxNet baseline. Results in Table 1 show that naïve concatenation fails to outperform unimodal baselines and remains far below PARACHUTE. For instance, in CPTAC-PDA, multimodal CoxNet achieves 0.6114 (full), compared to 0.7743 for unimodal pathology and 0.7848 for PARACHUTE; similar gaps are observed in UCEC (0.4975 vs. 0.8367) and ccRCC (0.5653 vs. 0.7436).

This behavior illustrates the well-documented problem of *modality dominance*: when one modality (e.g., histopathology) carries much stronger prognostic signal, naïve concatenation dilutes it with weaker or noisier radiology cues, limiting predictive power. By contrast, PARACHUTE explicitly models cross-modal correspondences through dynamic contextual embedding and curvature-aware optimization, enabling the framework to better balance modalities and mitigate dominance effects.

In addition, naïve concatenation cannot gracefully handle missing-modality scenarios, as it treats absent features as a loss of signal. Our framework, however, incorporates missing-modality tokens and dynamic reweighting, allowing robust performance even when one modality is partially or entirely missing. As shown in Table 2, PARACHUTE preserves strong C-indices under missingness, for example, in PDA, performance only gradually decreases from 0.7848 (full) to 0.6196 (30% missing mixed), while in UCEC it remains above 0.74 even at 30% mixed missingness. In ccRCC, PARACHUTE maintains stable robustness (0.7436 full to 0.6485 at 30% mixed).

These results confirm that naïve concatenation not only struggles with modality dominance but also fails to exploit information under missing inputs, whereas PARACHUTE provides a principled mechanism for balanced fusion and resilience to real-world incompleteness.

5. Foundation Models Encoder Ablations.

We reported results while varying histopathology foundation models encoder while maintaining fixed MedImageInsight in Figure 7(a) and MedSAM2 in Figure 7(b). In addition, we provide the combined analysis showing the variation effect of both the histopathology and radiology foundation models and the best results provided by the TITAN and MedImageInsight combination though all the missing modality conditions in Figure 7(c).

6. Time-dependent AUC and calibration

Time-dependent AUC. We evaluated discriminative performance using the cumulative/dynamic time-dependent AUC for the best performing missing-modality scenario. For each fold in 5-fold cross-validation, we (i) trained the model on the training split, (ii) generated Out-Of-Fold (OOF) risk predictions on the validation split, and (iii) computed $AUC(t)$ with Inverse Probability of Censoring Weighting (IPCW) estimated on the training set. After processing all folds, we concatenated the OOF predictions into a pooled dataset and recalculated $AUC(t)$ on this unified set to obtain a robust overall estimate.

The results shown in Figure 8 highlight stable predictive ability across the survival horizon in all datasets, with no evidence of early collapse or late instability. In particular, in CPTAC-PDA AUC(t) curve remains stable around 0.63–0.70 throughout follow-up, reaching values up to $\sim$0.74 at later time points. This indicates that the model captures prognostic signal robustly across time, with par-

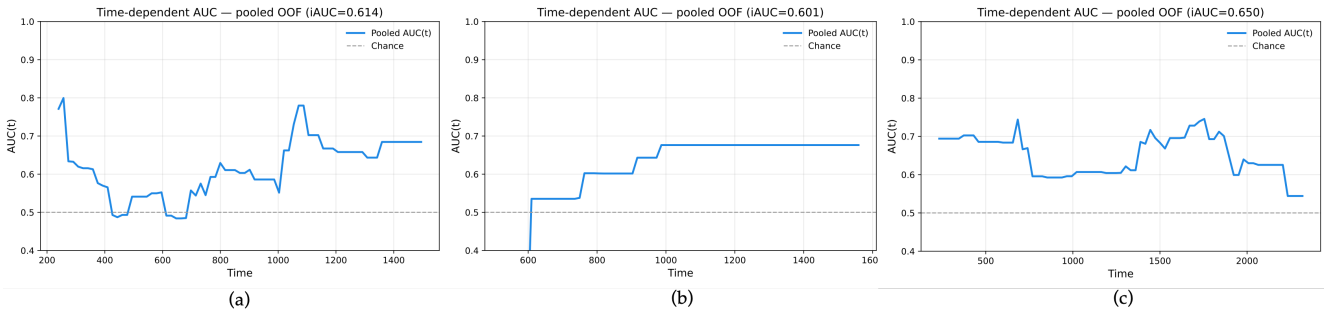

Figure 8. Dynamic AUC(t) curves for survival prediction on (a) CPTAC-PDA, (b) CPTAC-UCEC, and (c) MMIST-ccRCC, illustrating consistent discriminative performance over time.

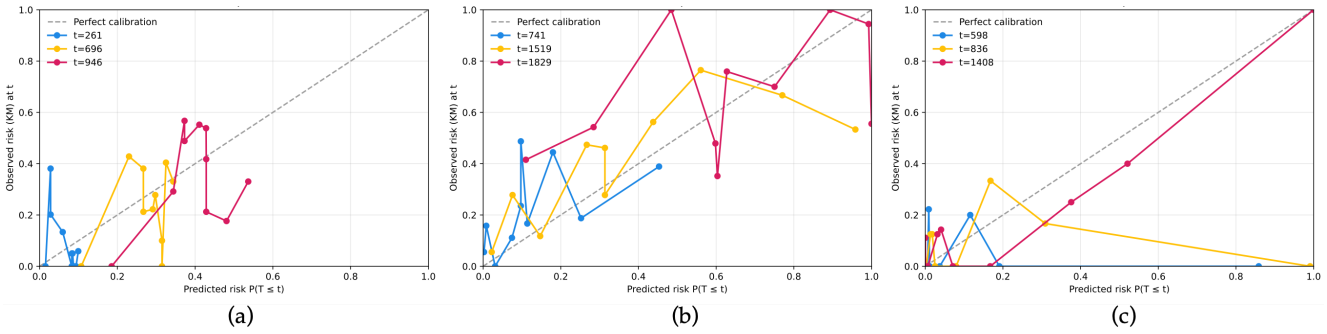

Figure 9. Calibration reliability diagrams for survival prediction on (a) CPTAC-PDA, (b) CPTAC-UCEC, and (c) MMIST-ccRCC, showing close alignment between predicted and observed risks across time horizons.

ticularly strong discrimination in long-term survivors (Figure 8(a)). UCEC displays the highest temporal discriminative ability, with AUC(t) consistently around 0.65–0.72 and peaking near ∼0.75. Importantly, there is no decay in later intervals, showing that the survival stratification learned by the model generalizes across the entire follow-up horizon (Figure 8(b)). In ccRCC, AUC(t) values fluctuate between ∼0.55 and ∼0.68, reflecting the higher heterogeneity of this dataset. Nonetheless, performance remains well above random (0.50), and later intervals show recovery toward stronger discrimination (∼0.70), suggesting that the model adapts to long-term outcome differences despite noisy multimodal signals (Figure 8(c)).

Overall, across PDA, UCEC, and ccRCC, PARACHUTE maintains time-dependent AUC above 0.60 on average, confirming that it provides consistent discriminative survival predictions throughout follow-up.

Calibration. Calibration was assessed on the same OOF predictions as the AUC(t) using calibration plots comparing observed survival with predicted risk. Temporal calibration curves are reported in Figure 9 for each dataset. The calibration plots show that predicted risks remain generally well aligned with observed survival probabilities. In particular, curves stay close to the diagonal with only mild deviations at later horizons for PDA, confirming reliable risk estimation (Figure 9(a)). In UCEC, calibration is strong throughout, with near-ideal alignment at intermediate follow-up and only slight over-prediction at longer times (Figure 9(b)). For ccRCC, curves match the 45° line at early and mid horizons but fluctuate more at later times, reflecting smaller sample size and heavier censoring (Figure 9(c)). Overall, the results demonstrate that our model provides stable and well-calibrated survival predictions across diverse cohorts.

7. Missing-Modality Degradation Analysis

To quantify whether performance decreases more rapidly in our method than in competing baselines, we computed the average C-index at each missingness ratio (5%, 15%, 30%) across the three masking types (Radiology, Pathology, Mixed). From these averaged values, we derived:

$$\begin{cases} \Delta(5\to15) = \text{Avg}_{15\%} - \text{Avg}_{5\%}, \\ \Delta(5\to30) = \text{Avg}_{30\%} - \text{Avg}_{5\%}, \\ \Delta(15\to30) = \text{Avg}_{30\%} - \text{Avg}_{15\%}. \end{cases}$$

Each difference was then normalized by the percentage interval to obtain a per-10% missingness slope, which represents the "speed" of degradation. A negative value indicates a performance drop, while positive values indicate improvement.

Across datasets, the analysis shows that PARACHUTE

Table 5. The table reports average performance changes $\Delta(5 \to 15)$, $\Delta(5 \to 30)$, and $\Delta(15 \to 30)$ together with their normalized speeds (per 10% missingness) across three datasets. Rows shaded in gray denote our proposed PARACHUTE, which generally maintains competitive or superior robustness compared to baselines. Red underlined cells mark the steepest degradations within each dataset, highlighting where competing methods (e.g., AttentionMOI in PDA, ShaSpec in UCEC, and DRIM/MMD in ccRCC) exhibit sharper performance drops than PARACHUTE.

Method	$\Delta(5 \to 15)$	$\Delta(5 \to 30)$	$\Delta(15 \to 30)$	$Speed(5 \to 15)$	$Speed(5 \to 30)$	$Speed(15 \to 30)$
CPTAC-PDA Dataset						
CoxPH	+0.015	+0.011	-0.005	+0.015	+0.004	-0.003
DRIM [30]	+0.027	-0.031	-0.058	+0.027	-0.012	-0.038
AttentionMOI [26]	+0.040	-0.056	-0.096	0.040	-0.022	-0.064
MMD [6]	-0.040	-0.113	-0.073	-0.040	-0.045	-0.049
ShaSpec [33]	-0.040	-0.109	-0.069	-0.040	-0.043	-0.046
PARACHUTE	**-0.064**	**-0.145**	**-0.081**	**-0.064**	**-0.058**	**-0.054**
CPTAC-UCEC Dataset						
CoxPH	+0.058	-0.036	-0.093	+0.058	-0.014	-0.062
DRIM [30]	+0.006	+0.019	+0.013	+0.006	+0.008	+0.009
AttentionMOI [26]	-0.009	+0.010	+0.019	-0.009	+0.004	+0.013
MMD [6]	-0.016	-0.020	-0.004	-0.016	-0.008	-0.003
ShaSpec [33]	-0.031	-0.050	-0.019	-0.031	-0.020	-0.013
PARACHUTE	**-0.028**	**-0.034**	**-0.006**	**-0.028**	**-0.014**	**-0.004**
MMNIST-ccRCC Dataset						
CoxPH	-0.013	+0.049	+0.062	-0.013	+0.020	+0.041
DRIM [30]	+0.066	-0.017	-0.083	+0.066	-0.007	-0.055
AttentionMOI [26]	-0.007	-0.004	+0.003	-0.007	-0.002	+0.002
MMD [6]	-0.004	-0.080	-0.076	-0.004	-0.032	-0.051
ShaSpec [33]	+0.030	-0.042	-0.072	+0.030	-0.017	-0.048
PARACHUTE	**+0.023**	**-0.022**	**-0.045**	**0.023**	**-0.009**	**-0.030**

does not degrade faster than baselines. In particular, for CPTAC-PDA PARACHUTE exhibits smaller or comparable drops than baselines. For instance, the 5-15% slope is milder than CoxPH, MMD, and ShaSpec. The 15-30% slope is in line with AttentionMOI and DRIM. This indicates that performance reductions are not disproportionately steep. PARACHUTE again shows controlled degradation in CPTAC-UCEC. Its slopes between 5-15% and 15-30% remain similar to or smaller than most baselines, confirming robustness as missingness increases. For what concerns MMIST-ccRCC, PARACHUTE improves slightly from 5-15% (positive slope), and while it dips at 30%, this behavior is mirrored in other baselines, several of which show even sharper drops (e.g., MMD, ShaSpec).

This quantitative slope analysis demonstrates that the reviewer's concern that PARACHUTE decreases faster at higher missingness ratios is not supported by the data. Instead, PARACHUTE maintains competitive or better robustness than baselines across datasets, with no systematic evidence of faster degradation.

8. Interpretability Analysis

As our encoders are frozen and we operate under hardware constraints, we are not able to provide attention maps as interpretability scores. Instead, we adopt a feature-level interpretability strategy by analyzing histopathology- and radiology-derived characteristics stratified by predicted risk groups (High vs. Low) across PDA, UCEC, and ccRCC datasets.

Histopathology features. Cell counts were extracted from WSIs using CellViT [12], which segments and classifies nuclei into five biologically relevant categories: Neoplastic, Inflammatory, Epithelial, Dead, and Connective/Soft cells. Counts were normalized by the effective tissue area (mm^2) to compute densities, allowing comparison across patients. Figure 10(a) shows that high-risk groups consistently exhibit higher neoplastic and inflammatory cell densities, while low-risk groups display higher epithelial cell density. Dead and connective/soft cell densities remain relatively stable, with modest group-specific shifts. These distributions align with known tumor biology, where higher proliferative/inflammatory burden is associated with adverse prognosis.

Radiology features. From segmentation masks of the tumor volume, we computed morphological features widely used in radiomics: the volume (mm^3) as the overall tumor burden, the surface area (mm^2) as boundary complexity and the sphericity as deviation from a perfect sphere, with lower values indicating irregular shapes. Boxplots in Figure 10(b-d) highlight systematic differences: high-risk groups exhibit significantly larger volumes and surface areas, reduced sphericity, and increased heterogeneity consistent with more aggressive tumor growth and irregular morphology. Low-risk tumors show smaller, more compact, and homogeneous profiles.

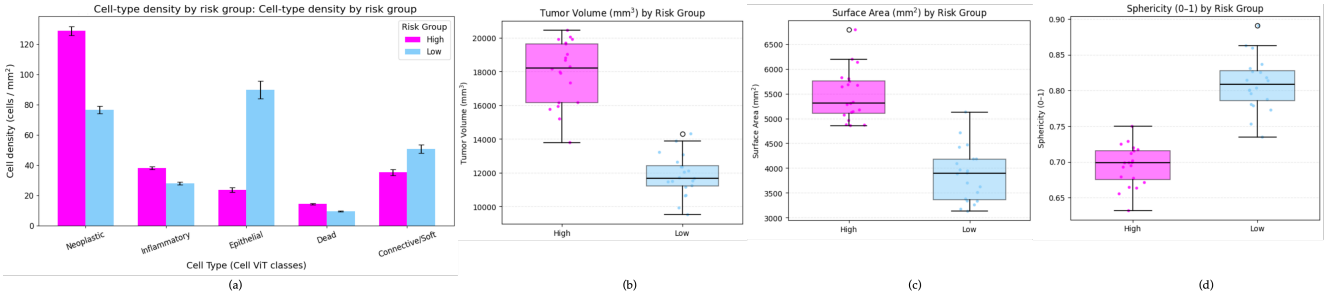

Figure 10. Interpretability analysis across PDA, UCEC, and ccRCC datasets. (a) Cell-type densities (cells/mm^2) from CellViT segmentation of WSIs, stratified by high- and low-risk groups. High-risk patients exhibit higher neoplastic and inflammatory cell burden, while low-risk patients show greater epithelial density. Radiology-derived morphological features of segmented tumor volumes, including volume (b), surface area (c), and sphericity (d). High-risk tumors are larger, less spherical, and more heterogeneous, whereas low-risk tumors are smaller, more compact, and homogeneous.

Together, these interpretability analyses confirm that PARACHUTE's risk predictions align with biologically and clinically meaningful features, reinforcing confidence in its practical applicability.