Heterogeneous Aligned Fusion for Survival Prediction with Missing Modalities

Zheng Zheng[1] ZXZ7934@MAVS.UTA.EDU
Yuzhi Guo[1] YUZHI.GUO@MAVS.UTA.EDU
Xiao Hu[1] XXH3416@MAVS.UTA.EDU
Yuwei Miao[1] YXM9326@MAVS.UTA.EDU
Hehuan Ma[1] HEHUAN.MA@MAVS.UTA.EDU
Jean Gao[1] GAO@UTA.EDU
Junzhou Huang[1] JZHUANG@UTA.EDU
[1] *University of Texas at Arlington, Arlington, TX, USA*

Editors: Under Review for MIDL 2026

Abstract

Accurate survival prediction is essential for guiding personalized treatment in head and neck cancer. Heterogeneous biomedical data, from histopathology to clinical and laboratory measurements, offer complementary prognostic value but differ in dimensionality, reside in incompatible feature spaces, and are frequently missing, making robust multimodal learning challenging. To address this, we propose **HAF (Heterogeneous Aligned Fusion)**, a three-stage framework for survival prediction under heterogeneous and incomplete multimodal inputs. HAF (i) uses detached prognostic supervision to obtain stable representations, (ii) performs lightweight global alignment that projects all modalities into a shared latent space while preserving patient-level discriminability, and (iii) enforces monotonic robust fusion that encourages performance to remain stable or improve when modalities are added. To the best of our knowledge, HAF is the first approach that jointly leverages all seven modalities in the HANCOCK cohort. Extensive comparisons against representative late-, early-, attention-based, and bilinear-interaction fusion methods demonstrate that HAF consistently improves both accuracy and robustness under heterogeneous and partially missing modalities.

Keywords: Heterogeneous Aligned Fusion (HAF), multimodal learning, head and neck cancer, survival prediction, pathology imaging, MIL

1. Introduction

Accurate survival prediction is central to precision oncology, enabling risk-adaptive decision making for patients with head and neck squamous cell carcinoma (HNSCC) (Tian et al., 2025). Modern cohorts combine high-dimensional pathology imaging with heterogeneous clinical and laboratory measurements, which differ in dimensionality, reside in incompatible feature spaces, and are often missing (Wissel et al., 2023; Li et al., 2024; Khagi and Kwon, 2020; Li and Tang, 2024a; Pan et al., 2020; Aly et al., 2023; Wu et al., 2024; Reza et al., 2024). This makes robust multimodal learning particularly challenging.

A natural starting point for multimodal survival prediction is to build strong unimodal predictors. Pathology-only systems such as CLAM-style MIL models (Tian et al., 2025)

provide reliable prognostic signals, yet treat each modality independently and therefore cannot exploit complementary clinical or laboratory information. Fusion-based multimodal designs address this limitation but introduce new challenges. Early fusion (Li and Tang, 2024b) concatenates heterogeneous features, enabling interaction while exposing the model to disparities in noise, scale, and dimensionality. Late fusion (Tian et al., 2025) mitigates these issues via modality-specific predictors, but loses fine-grained cross-modal complementarity. More expressive approaches, including attention-based fusion (Raza et al., 2025; Dang et al., 2024) and bilinear interaction frameworks (Benkirane et al., 2025), attempt to capture richer dependencies. However, they implicitly assume well-aligned modality embeddings. In reality, heterogeneous modalities remain misaligned (Li et al., 2024; Khagi and Kwon, 2020; Li and Tang, 2024a), allowing low-quality channels to propagate misleading signals that degrade the fused representation.

Existing robustness strategies, such as random-modality training (Li et al., 2025), help tolerate missing clinical modalities (Pan et al., 2020; Aly et al., 2023; Wu et al., 2024; Reza et al., 2024) but operate on unaligned representations, offering no principle for how modalities should substitute for one another. Consequently, performance may still deteriorate when noisy modalities are present, and full-modality predictions are not guaranteed to outperform reduced-modality ones. These limitations reveal three requirements for reliable multimodal survival prediction, especially since real-world clinical pipelines rarely obtain all modalities even though HANCOCK provides complete observations: (i) stable pathology representations before fusion, (ii) aligned latent geometry across heterogeneous modalities, and (iii) monotonic behavior such that adding modalities never degrades performance.

To address these challenges, we propose HAF (Heterogeneous Aligned Fusion), which decouples representation learning, establishes a shared latent geometry, and enforces reliable fusion under partial modality settings. (i) *Detached prognostic supervision* stabilizes pathology encoders by decoupling morphological representation learning from downstream fusion objectives. (ii) *Lightweight global alignment* (Kamboj and Do, 2025) projects all modalities into a shared patient-level latent space with low-rank consensus, mitigating incompatibilities among heterogeneous feature domains. (iii) *Monotonic robust fusion* (Li et al., 2025) penalizes cases where full-modality performance falls below that of partial-modality configurations, encouraging the model to rely more heavily on reliable modalities. All stages incorporate gradient detachment to prevent destructive cross-modality interference or dominance by any single modality.

We evaluate HAF on the HANCOCK dataset (Dörrich et al., 2025), which provides an unprecedented multimodal setting combining whole-slide imaging (WSI), tissue microarrays (TMA), and five structured clinical descriptors, capturing complementary aspects of tumor biology. In evaluation, HAF demonstrates stable optimization, improved cross-modality compatibility, and strong robustness under missing or noisy inputs, outperforming representative late-, early-, and attention-based baselines. A comprehensive review of multimodal fusion, alignment, and robustness methods is provided in the Appendix. In conclusion, HAF offers a principled view of multimodal interaction: unimodal semantics are stabilized, modalities are aligned into a substitutable shared geometry, and fusion remains reliable even under realistic missing-modality conditions.

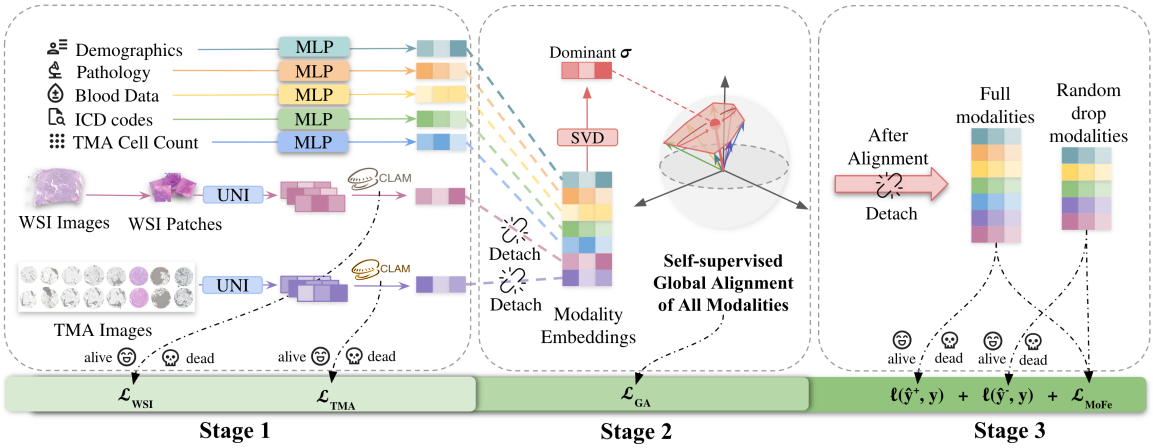

Figure 1: **Overview of the HAF framework.** Stage 1 learns stable unimodal embeddings under detached supervision. Stage 2 globally aligns modalities into a shared, substitutable geometry. Stage 3 fuses aligned features with random-modality training and a monotonic constraint, promoting robustness under missing modalities

2. Methods

2.1. Multimodal Inputs and Task Definition

We study binary survival prediction on the HANCOCK cohort (Dörrich et al., 2025), where the task is to predict whether a patient survives within a fixed follow-up window ($y \in \{0, 1\}$). Each patient provides up to seven heterogeneous modalities: WSI and TMA histopathology, TMA-derived cell-density maps, clinical metadata, pathological staging, blood biomarkers, and ICD diagnostic codes. In practice, not all patients are fully observed; imaging and laboratory assays may be unavailable for logistical or cost-related reasons, leading to patient-specific subsets of observed modalities. Formally, for each patient i we denote by $\mathcal{M}_i \subseteq \mathcal{M}$ the set of available modalities and by $\{x_i^{(m)}\}_{m \in \mathcal{M}_i}$ the corresponding inputs, and the goal is to learn a predictor f that maps these heterogeneous, partially missing observations to a binary event prediction: $\hat{y}_i = f(\{x_i^{(m)}\}_{m \in \mathcal{M}_i}), \hat{y}_i \in \{0, 1\}$.

2.2. Modality Representations

For WSI, we directly use the 1024-dimensional patch embeddings released by HANCOCK, which were extracted using the publicly available UNI (Chen et al., 2024) foundation model for computational pathology. For TMA, since raw core images are provided, we independently apply the same UNI model to extract 1024-dimensional embeddings per core, followed by bag-level aggregation in Stage 1. The remaining five modalities are compact tabular descriptors: $X^{(m)} \in \mathbb{R}^{d_m}$, $m \in \mathcal{M}_{\text{tab}}$, $d_m \leq 51$, where $\mathcal{M}_{\text{tab}} = \{\text{Cell, Clin, Path, Blood, ICD}\}$. All tabular features are preprocessed using min–max normalization, one-hot encoding, and imputation with the most frequent value, following the preprocessing protocol in the HANCOCK dataset. This yields a unified representation where high-dimensional imaging bags and low-dimensional tabular vectors coexist, but still differ substantially in scale, noise level, and semantic content, motivating the staged alignment and fusion strategy introduced in the following sections.

2.3. Stage 1: Prognostic Pathology Representation Learning

Each pathology modality (WSI or TMA) is represented as a variable-length bag of patch embeddings $H \in \mathbb{R}^{N\times 1024}$, where N depends on sampled tissue content. We adopt the gated-attention MIL encoder from CLAM (Lu et al., 2021).
To obtain compact and expressive instance features, each 1024-dimensional patch embedding is first projected using a linear mapping $w_l \in \mathbb{R}^{1024\times d_1}$. Two learnable projection matrices $U, V \in \mathbb{R}^{d_1\times d_2}$ generate gated-attention features, and a learnable query vector $w_a \in \mathbb{R}^{d_2\times 1}$ computes instance importance:

$$a = \mathrm{softmax}\Big(\big[\tanh\big((Hw_l)V\big) \odot \sigma\big((Hw_l)U\big)\big]w_a\Big). \tag{1}$$

The slide-level representation for modality m is obtained by weighted aggregation:

$$z^{(m)} = (Hw_l)^\top a, \tag{2}$$

where $m \in \{\mathrm{WSI}, \mathrm{TMA}\}$ and d_2 denotes the attention dimension. Each modality-specific embedding is supervised using a cross-entropy loss combined with instance-level regularization:

$$\mathcal{L}^{(m)}_{\mathrm{img}} = \mathcal{L}_{\mathrm{CE}}(y, \hat{y}^{(m)}) + \mathcal{L}^{(m)}_{\mathrm{inst}}. \tag{3}$$

We use $d_1 = 64$ and $d_2 = 32$. The resulting pathology embeddings are detached before Stage 2 to preserve their prognostic semantics and avoid distortion during multimodal training.

2.4. Stage 2: Global Aligned Shared Latent Projection

This stage aligns modalities into a shared latent space, ensuring that downstream robustness operates on substitutable rather than incompatible modality embeddings. To achieve this, we first project each modality into a common latent space using small MLPs, then apply global alignment losses to encourage their convergence onto a unified consensus signal. This yields a coherent multimodal representation that amplifies high-quality modalities while suppressing noisy or weak ones.

2.4.1. Latent Projection into a Shared Space

To place all modalities on a comparable footing, each modality-specific embedding $z^{(m)}$ is mapped into a shared latent space through a lightweight two-layer MLP:

$$u^{(m)} = \phi^{(m)}(z^{(m)}) \in \mathbb{R}^{d_{\mathrm{out}}}, \quad m \in \mathcal{M}, \tag{4}$$

where $\mathcal{M} = \{\mathrm{WSI}, \mathrm{TMA}, \mathrm{Clin}, \mathrm{Path}, \mathrm{Blood}, \mathrm{ICD}, \mathrm{Cell}\}$ and $d_{\mathrm{out}} = 128$. The alignment treats all modalities symmetrically, avoiding vision-centric or tabular-centric bias.

2.4.2. Global Alignment via Singular Value Decomposition

Singular value decomposition (SVD) decomposes a set of vectors into orthogonal directions ordered by how much signal the data concentrate along each direction. The largest singular value σ_1 corresponds to the dominant component shared across vectors, while smaller singular values capture weaker signals. In our setting, if modalities project toward a common

disease-related direction, most information accumulates in this dominant component (large σ_1), and the remaining components become comparatively weak (small $\sigma_2, \ldots, \sigma_k$).
Inspired by principled alignment (Liu et al., 2025), we leverage this property by encouraging all projected modality embeddings to concentrate their variation along the dominant component of the patient-specific matrix $U = [u^{(1)}, \ldots, u^{(M)}] \in \mathbb{R}^{d_{\text{out}} \times M}$. Formally, SVD decomposes U into $U = Q\Sigma R^\top$, where $Q = [q_1, q_2, \ldots, q_k] \in \mathbb{R}^{d_{\text{out}} \times k}$ contains orthonormal left singular vectors, and $\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_k), k = \min(d_{\text{out}}, M)$ stores the singular values in descending order. Here, σ_1 quantifies the strength of the dominant shared component across modalities, and q_1 is the corresponding direction in the latent space. When $\sigma_1 \gg \sigma_2, \ldots, \sigma_k$, the decomposition becomes approximately rank-1, meaning that all modality embeddings align along q_1 and express a common latent signal. To enforce this behavior, we introduce two complementary losses. First, the *singular emphasis loss* increases the dominance of σ_1, promoting alignment of all modalities toward the shared component:

$$\mathcal{L}_{\text{SV}} = -\log \frac{\exp(\sigma_1/\tau)}{\sum_{j=1}^{k} \exp(\sigma_j/\tau)}. \tag{5}$$

Second, during this alignment process, different patients could collapse to the same direction. To retain patient-level distinction, we introduce a *dominant-direction discriminability loss* that separates patients based on their dominant singular vectors:

$$\mathcal{L}_{\text{PD}} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp((q_1^{(i)})^\top q_1^{(i)}/\tau)}{\sum_{j=1}^{B} \exp((q_1^{(i)})^\top q_1^{(j)}/\tau)}. \tag{6}$$

The total alignment objective is

$$\mathcal{L}_{\text{GA}} = \mathcal{L}_{\text{SV}} + \lambda_{\text{PD}} \mathcal{L}_{\text{PD}}. \tag{7}$$

In our experiments, we set the patient-discriminability weight to $\lambda_{PD} = 0.01$, which balances alignment strength and inter-patient separation.

2.5. Stage 3: Fusion with Robust Modality Collaboration

We perform fused multimodal prediction using the aligned features from Stage 2. For each patient, let $\{u^{(m)}\}_{m \in \mathcal{M}}$ denote the aligned modality embeddings and $\mathcal{M}_{\text{obs}} \subseteq \mathcal{M}$ the set of modalities observed for that patient. We construct a fused representation by concatenating the available modalities:

$$h = \text{concat}\left(\{u^{(m)}\}_{m \in \mathcal{M}_{\text{obs}}}\right), \tag{8}$$

and obtain a scalar prediction using a shared fusion MLP,

$$s = \phi(h), \qquad \hat{y} = \sigma(s), \tag{9}$$

where $\sigma(\cdot)$ is the sigmoid function.

Random-modality training. To ensure robustness under missing or unreliable modalities, we adopt random-modality (Li et al., 2025) masking on top of the aligned geometry, enabling the model to rely on substitutable modalities rather than raw heterogeneous features. At each iteration, we randomly mask a subset of modalities to obtain a reduced-modality representation h^- and its prediction $\hat{y}^-$, alongside the full-modality prediction $\hat{y}^+$. This encourages the model to rely on informative modalities while remaining stable when others are absent.

Monotonic collaboration constraint. To guarantee that incorporating more modalities never degrades performance, we impose a monotonicity loss:

$$\mathcal{L}_{\text{MoFe}} = \max\big(\ell(\hat{y}^{+}, y) - \ell(\hat{y}^{-}, y), 0\big), \tag{10}$$

which penalizes cases where the full-modality prediction is worse than that of a reduced subset.

Fusion objective. The final fusion loss combines full-modality supervision, reduced-modality supervision, and the monotonicity constraint:

$$\mathcal{L}_{\text{fusion}} = \ell(\hat{y}^{+}, y) + \ell(\hat{y}^{-}, y) + \lambda_{\text{MoFe}}\, \mathcal{L}_{\text{MoFe}}, \tag{11}$$

where we set $\lambda_{\text{MoFe}} = 0.1$ in our experiments. The full objective therefore enforces consistency across complete and reduced-modality inputs, while the monotonicity term prevents prediction reversals under modality removal. We found this combination to yield stable optimization and improved robustness in practice.

2.6. Overall Training Objective

The full HAF objective integrates the three stages through four loss blocks:

$$\mathcal{L}_{\text{total}} = \underbrace{\mathcal{L}_{\text{WSI}} + \mathcal{L}_{\text{TMA}}}_{\text{Stage 1}} + \underbrace{\mathcal{L}_{\text{GA}}}_{\text{Stage 2}} + \underbrace{\mathcal{L}_{\text{fusion}}}_{\text{Stage 3}}. \tag{12}$$

Here, $\mathcal{L}_{\text{WSI}}$ and $\mathcal{L}_{\text{TMA}}$ include $\mathcal{L}_{\text{CE}}$ and instance-level regularization terms as defined in Eq. (3). The global alignment loss $\mathcal{L}_{\text{GA}}$ follows Eq. (7). The fusion objective $\mathcal{L}_{\text{fusion}}$ follows Eq. (11). Crucially, the outputs of each stage are detached before being passed to the next stage. This stage-wise decoupling prevents conflicting gradients from dominating weaker modalities and ensures that each learning objective, including (i) pathology semantics, (ii) cross-modality compatibility, and (iii) robustness to missing data, is optimized without being overridden by others.

3. Experiments

3.1. Experimental Setup

We evaluate HAF on the HANCOCK head and neck cancer cohort (Dörrich et al., 2025) under a binary survival prediction setting, where the task is to predict whether a patient experiences an event within a fixed follow-up window ($y \in \{0, 1\}$). We follow the official preprocessing for tabular modalities, including min–max normalization, one-hot encoding, and imputation with the most frequent value. For WSI histopathology, we use the 1024-dimensional UNI (Chen et al., 2024) patch embeddings released by HANCOCK, and for TMA we apply the same UNI encoder to the raw cores to obtain 1024-dimensional embeddings.

We adopt stratified 10-fold cross-validation at the patient level and report mean Accuracy and AUC over test folds. All models are trained with Adam (learning rate 10^{-4}, weight decay 10^{-5}) for up to 200 epochs with batch size 64, using a `ReduceLROnPlateau` scheduler

(patience 15, factor 0.5) and early stopping based on validation performance. Unless otherwise stated, all experiments are conducted on an NVIDIA RTX A6000 GPU; full 10-fold training for HAF requires roughly 4 hours in total.

3.2. Baselines and Variants

We evaluate HAF against a broad set of pathology-only, multimodal, and alignment–robustness variants to isolate the contributions of each component. The official HANCOCK benchmarks consist of **WSI-CLAM**, **TMA-CLAM**, and **WSI+TMA-CLAM**, which serve as unimodal and dual-modality pathology references. We also reproduce WSI+TMA-CLAM using the released UNI features to ensure evaluation consistency. Moving beyond pathology, we include two naive fusion models: a **WSI+TMA Fusion MLP** and an **All-Modality Fusion MLP**, both of which directly concatenate modality embeddings without any alignment or robustness mechanisms. These baselines quantify the limitations of simple multimodal aggregation in heterogeneous feature spaces.

To examine the effects of HAF's core components, we evaluate **Global Alignment (GA)**, the SVD-based latent alignment from Stage 2; **CLIP Alignment**, a vision-centric contrastive baseline anchoring non-visual modalities to WSI as a comparable alignment method; and **Random-Modality Drop**, the robustness mechanism from Stage 3. We further test their combination to assess whether alignment and robustness act synergistically. Finally, we compare against representative multimodal approaches such as **PS3**, **MDLM**, **MFMF**, and a **Simple Feature Interaction** model to situate HAF relative to existing fusion strategies. All model variants share the same detached pathology encoders from Stage 1, ensuring that differences stem solely from their alignment and fusion strategies.

3.3. Overall Quantitative Results

We first examine pathology-only and naive fusion models. The official HANCOCK baselines (**WSI-CLAM**, **TMA-CLAM**, **WSI+TMA-CLAM**) provide a unimodal reference point, and our reproduction of **WSI+TMA-CLAM** closely matches the reported results (Accuracy 0.712, AUC 0.679). The **WSI+TMA Fusion MLP** improves accuracy (0.739) but yields a lower AUC (0.668). Extending fusion to all seven modalities with the **All-Modality Fusion MLP** gives moderate improvements (AUC 0.694), although performance varies considerably across folds. Among the single-component variants, **Global Alignment (GA)** achieves the highest accuracy (0.752), while **CLIP Alignment** shows a similar but slightly weaker trend. **Random-Modality Drop** primarily improves AUC (0.715) relative to the All-Modality Fusion MLP. Combining Drop with alignment further increases discrimination: **CLIP + Drop** reaches an AUC of 0.735, and the full **HAF (GA + Drop)** achieves the best overall performance (AUC 0.739). For comparison, representative multimodal frameworks such as **PS3** (0.626 / 0.718), **MDLM** (0.557 / 0.626), **MFMF** (0.675 / 0.732), and **Simple Feature Interaction** (0.705 / 0.677) perform notably worse across accuracy and AUC. These results highlight the challenge posed by heterogeneous modality quality and demonstrate the benefit of combining alignment and robustness for stable and discriminative multimodal fusion.

Table 1: **Performance of multimodal fusion, alignment, and robustness variants, along with pathology-only and comparable multimodal baselines.**

Method	Accuracy	AUC
Pathology-only baselines		
WSI-CLAM	–	0.65
TMA-CLAM	–	0.52
WSI+TMA-CLAM	–	0.69
WSI+TMA-CLAM (reproduced)	0.712±0.087	0.679±0.119
Fusion models and HAF variants		
WSI+TMA Fusion MLP	0.739±0.041	0.668±0.133
All-Modality Fusion MLP	0.748±0.046	0.694±0.113
Global Alignment (GA)	**0.752±0.047**	0.698±0.127
CLIP Alignment	0.741±0.074	0.697±0.103
Random-Modality Drop	0.748±0.074	0.715±0.099
CLIP + Random Drop	0.741±0.073	0.735±0.097
HAF (GA + Random Drop)	0.745±0.065	**0.739±0.092**
Comparable multimodal methods		
PS3	0.626±0.123	0.718±0.117
MDLM	0.557±0.145	0.626±0.122
MFMF	0.675±0.089	0.732±0.127
Simple Feature Interaction	0.705±0.083	0.677±0.098

3.4. Ablations on Stage-wise Detachment

Training without detachment corresponds to end-to-end optimization across all stages, which often entangles heterogeneous objectives and distorts the geometry needed for modality substitutability. When heterogeneous objectives are trained jointly without isolation, cross-stage gradients can conflict and distort earlier representations. Detachment prevents such interference by allowing each stage to converge independently before passing non-trainable features forward.

A small trade-off appears in the alignment-only setting: without detachment, global alignment is less complete and modality-specific biases leak into the shared space. These residual biases can sometimes boost accuracy via correlated but non-generalizable cues, yet they blur patient-level separability and reduce AUC. In contrast, detachment removes such interference and yields more consistent, semantically grounded representations, improving the robustness of the full HAF framework.

3.5. Robustness to Modality Drop

Fig. 2 summarizes model robustness under varying drop probabilities during testing. AUC and accuracy remain nearly constant up to $p = 0.4$, indicating that the model compensates for missing modalities by relying on reliable inputs, and degrade only when most modalities

Table 2: **Ablations on stage-wise detachment.** "w" denotes training *with* detachment, "w/o" indicates no detachment.

Setting	Metric	w (with detach)	w/o (no detach)
All modalities	Accuracy	**0.748±0.046**	0.738±0.063
	AUC	0.694±0.113	**0.698±0.110**
+ Global Alignment	Accuracy	**0.752±0.047**	0.752±0.050
	AUC	0.698±0.127	**0.727±0.108**
+ Random Drop	Accuracy	**0.748±0.074**	0.748±0.081
	AUC	**0.715±0.099**	0.714±0.101
HAF	Accuracy	0.745±0.065	**0.748±0.052**
	AUC	**0.739±0.092**	0.721±0.098

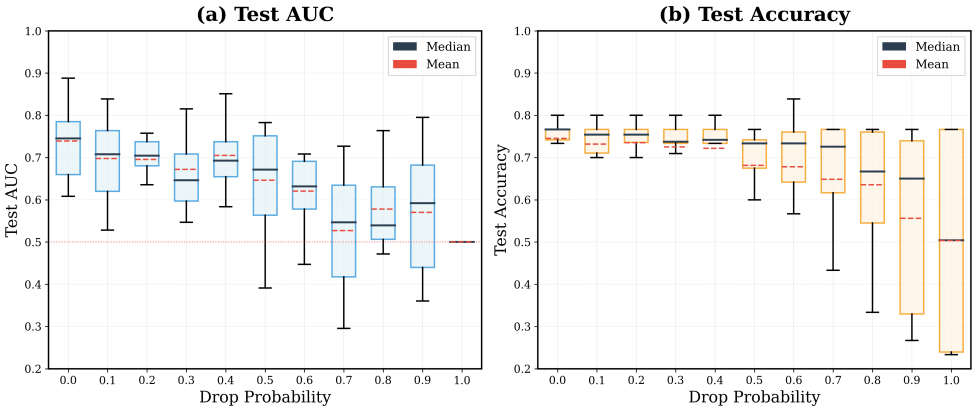

Figure 2: **Model robustness under modality dropout.**

are absent ($p > 0.4$), demonstrating improved resilience to real-world incompleteness and noise.

3.6. Representation Analysis

To illustrate the alignment effect, Fig. 3 visualizes multimodal embeddings before and after global alignment. Each color represents one modality, and each point corresponds to a patient from the test set.

Before alignment, modalities form well-separated clusters with large inter-modality gaps, while patient embeddings within the same modality are highly overlapping, indicating strong modality bias but limited patient discriminability. After alignment, modalities become more coherent along shared axes and different patients are pulled further apart, simultaneously reducing inter-modality discrepancies and enhancing inter-patient separability.

The heatmap in Fig. 4 shows the same trend. Before alignment, feature intensity sequences across modalities are largely uncorrelated, whereas after alignment they become synchronized for the same patient, indicating that representations are projected onto a coherent semantic basis.

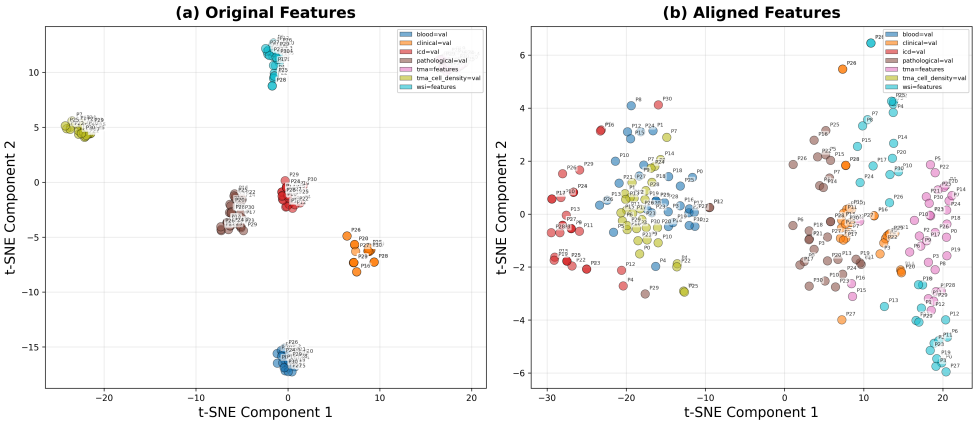

Figure 3: **t-SNE projection of multimodal embeddings before and after global alignment.**

Figure 4: **Heatmap of aligned modality representations for a representative patient.**

4. Conclusion

We presented **HAF**, a staged and detached multimodal fusion framework that mitigates cross-objective interference, stabilizes pathology representations, establishes a shared cross-modality geometry, and improves robustness under missing or noisy modalities. Beyond its empirical gains on HANCOCK, HAF also provides a general recipe for designing principled fusion pipelines in highly heterogeneous clinical settings. Future work will focus on exploring stronger task-aware alignment, incorporating finer spatial reasoning, and validating HAF on broader clinical datasets. In addition, extending HAF to prospective cohorts and evaluating its utility in real clinical decision support will further illuminate its translational potential.

References

Farhannah Aly, Christian Rønn Hansen, Daniel Al Mouiee, Purnima Sundaresan, Ali Haidar, Shalini Vinod, and Lois Holloway. Outcome prediction models incorporating clinical variables for head and neck squamous cell carcinoma: A systematic review of methodological conduct and risk of bias. *Radiotherapy and Oncology*, 183:109629, 2023.

Hakim Benkirane, Maria Vakalopoulou, David Planchard, Julien Adam, Ken Olaussen, Stefan Michiels, and Paul-Henry Cournède. Multimodal customics: A unified and interpretable multi-task deep learning framework for multimodal integrative data analysis in oncology. *PLOS Computational Biology*, 21(6):e1013012, 2025.

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.

Giordano Cicchetti, Eleonora Grassucci, Luigi Sigillo, and Danilo Comminiello. Gramian multimodal representation learning and alignment. *arXiv preprint arXiv:2412.11959*, 2024.

Thao M Dang, Yuzhi Guo, Hehuan Ma, Qifeng Zhou, Saiyang Na, Jean Gao, and Junzhou Huang. Mfmf: multiple foundation model fusion networks for whole slide image classification. In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–8, 2024.

Marion Dörrich, Matthias Balk, Tatjana Heusinger, Sandra Beyer, Hamed Mirbagheri, David J Fischer, Hassan Kanso, Christian Matek, Arndt Hartmann, Heinrich Iro, et al. A multimodal dataset for precision oncology in head and neck cancer. *Nature Communications*, 16(1):7163, 2025.

Shuang Hao, Wenfeng Han, Tao Jiang, Yiping Li, Haonan Wu, Chunlin Zhong, Zhangjun Zhou, and He Tang. Synthetic data in ai: Challenges, applications, and ethical implications. *arXiv preprint arXiv:2401.01629*, 2024.

Abhi Kamboj and Minh N Do. Towards achieving perfect multimodal alignment. *arXiv preprint arXiv:2503.15352*, 2025.

Bijen Khagi and Goo-Rak Kwon. 3d cnn design for the classification of alzheimer's disease using brain mri and pet. *IEEE Access*, 8:217830–217847, 2020. doi: 10.1109/ACCESS.2020.3040486.

Sijie Li, Chen Chen, and Jungong Han. Simmlm: A simple framework for multi-modal learning with missing modality. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24068–24077, 2025.

Songtao Li and Hao Tang. Multimodal alignment and fusion: A survey. *arXiv preprint arXiv:2411.17040*, 2024a.

Songtao Li and Hao Tang. Multimodal alignment and fusion: A survey. *arXiv preprint arXiv:2411.17040*, 2024b.

Yihao Li, Mostafa El Habib Daho, Pierre-Henri Conze, Rachid Zeghlache, Hugo Le Boité, Ramin Tadayoni, Béatrice Cochener, Mathieu Lamard, and Gwenolé Quellec. A review of deep learning-based information fusion techniques for multimodal medical image classification. *Computers in Biology and Medicine*, 177:108635, 2024.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

Xiaohao Liu, Xiaobo Xia, See-Kiong Ng, and Tat-Seng Chua. Principled multimodal representation learning. *arXiv preprint arXiv:2507.17343*, 2025.

Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.

Yongsheng Pan, Mingxia Liu, Chunfeng Lian, Yong Xia, and Dinggang Shen. Spatially-constrained fisher representation for brain disease identification with incomplete multi-modal neuroimages. *IEEE Transactions on Medical Imaging*, 39(9):2965–2975, 2020. doi: 10.1109/TMI.2020.2983085.

Yang Qin, Yuan Sun, Dezhong Peng, Joey Tianyi Zhou, Xi Peng, and Peng Hu. Cross-modal active complementary learning with self-refining correspondence. *Advances in neural information processing systems*, 36:24829–24840, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

Abshishek Rajora, Shubham Gupta, and Suman Kundu. Cross-aligned fusion for multimodal understanding. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5730–5740. IEEE, 2025.

Manahil Raza, Ayesha Azam, Talha Qaiser, and Nasir Rajpoot. Ps3: A multimodal transformer integrating pathology reports with histology images and biological pathways for cancer survival prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22175–22186, 2025.

Md Kaykobad Reza, Ashley Prater-Bennette, and M Salman Asif. Robust multimodal learning with missing modalities via parameter-efficient adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Yujie Sun, Dongfang Sheng, Zihan Zhou, and Yifei Wu. Ai hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content. *Humanities and Social Sciences Communications*, 11(1):1–14, 2024.

Ruxian Tian, Feng Hou, Haicheng Zhang, Guohua Yu, Ping Yang, Jiaxuan Li, Ting Yuan, Xi Chen, Ying Chen, Yan Hao, et al. Multimodal fusion model for prognostic prediction and radiotherapy response assessment in head and neck squamous cell carcinoma. *npj Digital Medicine*, 8(1):302, 2025.

Shicai Wei, Chunbo Luo, and Yang Luo. Boosting multimodal learning via disentangled gradient learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22879–22888, 2025.

David Wissel, Daniel Rowson, and Valentina Boeva. Systematic comparison of multi-omics survival models reveals a widespread lack of noise resistance. *Cell Reports Methods*, 3(4), 2023.

Zhenbang Wu, Anant Dadu, Nicholas Tustison, Brian Avants, Mike Nalls, Jimeng Sun, and Faraz Faghri. Multimodal patient representation learning with missing modalities and labels. In *The Twelfth International Conference on Learning Representations*, 2024.

Zihao Zhao, Yuxiao Liu, Han Wu, Mei Wang, Yonghao Li, Sheng Wang, Lin Teng, Disheng Liu, Zhiming Cui, Qian Wang, et al. Clip in medical imaging: A survey. *Medical Image Analysis*, page 103551, 2025.

Appendix A. Related Work

A.1. Multimodal survival prediction in oncology

Recent multimodal survival models can be grouped into four mainstream fusion paradigms: (A) *Early fusion*, which concatenates heterogeneous modality embeddings prior to prediction, but often suffers from modality mismatch and large quality disparities across inputs; (B) *Late fusion*, as in MDLM (Tian et al., 2025), which computes modality-specific risk scores and aggregates them at the prediction level, thereby limiting deeper cross-modal interaction; (C) *Attention-based fusion*, where methods such as PS3 (Raza et al., 2025) perform token-level cross-modal attention but still lack explicit mechanisms for aligning heterogeneous representations; and (D) *Bilinear interaction frameworks*, exemplified by CustOmics (Benkirane et al., 2025), which model pairwise multiplicative relationships across modalities but remain constrained when modalities occupy incompatible or poorly aligned feature spaces. Within the attention-based family, MFMF (Dang et al., 2024) adopts a distinct asymmetric strategy: it uses auxiliary modalities to compute attention over a single target modality, effectively maximizing information extraction from that modality. Our reproductions reveal that this design often induces single-modality dominance, with performance driven largely by the modality chosen as the value provider.

Overall, current methods lack a principled treatment of modality-quality differences and the inherent variability across modalities. This omission makes it difficult to design mechanisms that allow modalities to substitute for each other or to bias the model appropriately toward higher-quality modalities. In addition, existing approaches typically assume that all modalities are fully available, with little discussion of how fusion should remain effective under realistic conditions where one or more modalities may be missing.

A.2. Cross-modality alignment

A common strategy for multimodal fusion is to directly concatenate encoded features (Li et al., 2024; Wissel et al., 2023), but this straightforward approach provides no guarantee that heterogeneous modalities occupy compatible representation spaces. Recent work has shown that aligning modalities before fusion—rather than fusing them directly—yields stronger and more semantically coherent representations, as demonstrated in CLIP-style contrastive models and cross-aligned fusion frameworks (Radford et al., 2021; Rajora et al., 2025; Zhao et al., 2025). Contrastive alignment, however, requires curated modality pairs and becomes inefficient as the number of modalities increases. Facing this limitation, approaches shift toward *geometry-driven alignment*, which aligns modalities by enforcing shared latent structure leveraging volume minimization, or SVD formulations (Cicchetti et al., 2024; Kamboj and Do, 2025; Liu et al., 2025). However, they have rarely been applied in clinical survival prediction, and, more importantly, existing alignment studies typically consider only two or three modalities, leaving their behavior under large-scale heterogeneous modality alignment essentially unexplored.

A.3. Missing-modality robustness

Missing data is pervasive in clinical workflows where imaging, assays, or laboratory tests may be absent for logistical or cost-related reasons (Pan et al., 2020; Aly et al., 2023; Wu et al., 2024; Reza et al., 2024). Generative imputation strategies (Hao et al., 2024; Liang et al., 2022; Qin et al., 2023) attempt to synthesize absent modalities but can introduce hallucinations or low-fidelity artifacts (Sun et al., 2024), making reliability difficult to guarantee. Random-modality training and stochastic gating (Li et al., 2025; Wei et al., 2025) provide an alternative by encouraging models to remain predictive even when some modalities are dropped during training. However, these robustness strategies are typically applied without explicit geometric alignment, leaving cross-modality interactions loosely constrained and susceptible to representation collapse.