

# Heterogeneous Aligned Fusion for Survival Prediction with Missing Modalities

Zheng Zheng<sup>1</sup>

Yuzhi Guo<sup>1</sup>

Xiao Hu<sup>1</sup>

Yuwei Miao<sup>1</sup>

Hehuan Ma<sup>1</sup>

Jean Gao<sup>1</sup>

Junzhou Huang<sup>1</sup>

<sup>1</sup> *University of Texas at Arlington, Arlington, TX, USA*

ABC@SAMPLE.EDU

XYZ@SAMPLE.EDU

ALPHABETA@EXAMPLE.EDU

UVW@FOO.AC.UK

FGH@BAR.COM

UVW@FOO.AC.UK

HUANG@EXAMPLE.EDU

**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 our knowledge, HAF is the first approach that jointly leverages all seven modalities in the HANCOCK cohort. Extensive comparisons against representative late-fusion, early-fusion, and attention-based 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). Recent advances allow the integration of multimodal data—ranging from high-dimensional pathology imaging to structured clinical records. However, effectively leveraging these heterogeneous modalities is fundamentally challenging due to (i) drastic differences in dimensionality and noise levels (Wissel et al., 2023), (ii) incompatible representation spaces that hinder cross-modality integration (Li et al., 2024; Khagi and Kwon, 2020;

Li and Tang, 2024a), and (iii) inconsistent modality availability in clinical workflows (Pan et al., 2020; Aly et al., 2023; Wu et al., 2024; Reza et al., 2024). Together, these issues make it difficult to reliably benefit from informative modalities while preventing noisy or missing ones from degrading prediction quality.

Existing multimodal systems primarily follow three fusion paradigms. (1) *Late fusion* (Tian et al., 2025) aggregates modality-specific risks after independent encoding, limiting the ability to model cross-modal complementarity. (2) *Early fusion* (Li and Tang, 2024b) concatenates heterogeneous features within a single predictor, causing the joint space to inherit modality-specific noise and scale disparities. (3) *Attention-based fusion* (Raza et al., 2025; Dang et al., 2024) learns cross-modality interactions through token-level attention, yet low-quality modalities can propagate misleading keys or values, thereby degrading the representations of other modalities. Ultimately, none of these paradigms provide a mechanism to selectively exploit high-quality information while suppressing harmful signals.

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) enforces the performance under partial-modality configurations does not exceed that of the full-modality setting, 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 validate HAF on the HANCOCK dataset (Dörrieh 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.

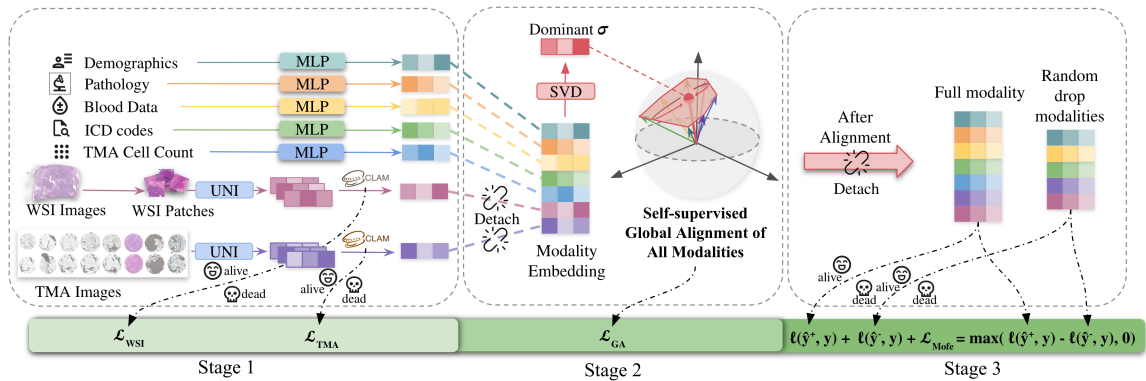


Figure 1: Pipeline Overview

## 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 ( $y \in \{0, 1\}$ ). 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.

### 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. 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 methods in the HANCOCK dataset.

### 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:

$$a = \text{softmax}\left(\left[\tanh((Hw_l)V) \odot \sigma((Hw_l)U)\right]w_a\right). \quad (1)$$

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

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

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

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

In our implementation,  $d_1 = 64$  and  $d_2 = 32$ . This stage ensures that high-dimensional pathology features encode clinically meaningful prognostic cues before interacting with other modalities. To prevent subsequent multimodal objectives from distorting pathology-specific semantics, both  $z^{\text{WSI}}$  and  $z^{\text{TMA}}$  are detached from gradient flow before Stage 2.

## 2.4. Stage 2: Global Aligned Shared Latent Projection

This stage aims to resolve this mismatch by aligning all modalities toward a shared disease-relevant direction. 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:  $u^{(m)} = \phi^{(m)}(z^{(m)}) \in \mathbb{R}^{d_{\text{out}}}$ ,  $m \in \mathcal{M}$ , where  $\mathcal{M} = \{\text{WSI, TMA, Clin, Path, Blood, ICD, Cell}\}$  and  $d_{\text{out}} = 128$ .

### 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  $\sigma_1$  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  $\sigma_1$ ), and the remaining components become comparatively weak (small  $\sigma_2, \dots, \sigma_k$ ).

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  $U = [u^{(1)}, \dots, u^{(M)}] \in \mathbb{R}^{d_{\text{out}} \times M}$ . Formally, SVD decomposes  $U$  into  $U = Q\Sigma R^\top$ , where  $Q = [q_1, q_2, \dots, q_k] \in \mathbb{R}^{d_{\text{out}} \times k}$  contains orthonormal left singular vectors, and  $\Sigma = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_k)$ ,  $k = \min(d_{\text{out}}, M)$  stores the singular values in descending order. Here,  $\sigma_1$  quantifies the strength of the dominant shared component across modalities, and  $q_1$  is the corresponding direction in the latent space. When  $\sigma_1 \gg \sigma_2, \dots, \sigma_k$ , 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  $\sigma_1$ , promoting alignment of all modalities toward the shared component:

$$\mathcal{L}_{\text{SV}} = -\log \frac{\exp(\sigma_1/\tau)}{\sum_{j=1}^k \exp(\sigma_j/\tau)} \quad (4)$$

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:

$$\mathcal{L}_{\text{PD}} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp((q_1^{(i)})^\top q_1^{(i)}/\tau)}{\sum_{j=1}^B \exp((q_1^{(i)})^\top q_1^{(j)}/\tau)} \quad (5)$$

The total alignment objective is

$$\mathcal{L}_{\text{GA}} = \mathcal{L}_{\text{SV}} + \lambda_{\text{PD}} \mathcal{L}_{\text{PD}} \quad (6)$$

In our experiments, we set the patient-discriminability weight to  $\lambda_{\text{PD}} = 0.01$ , 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:

$$h = \text{concat}(\{u^{(m)}\}_{m \in \mathcal{M}_{\text{obs}}}), \quad (7)$$

and obtain a scalar prediction using a shared fusion MLP,

$$s = \phi(h), \quad \hat{y} = \sigma(s), \quad (8)$$

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

**Random-modality training.** To ensure robustness under missing or unreliable modalities, we adopt random-modality masking (Li et al., 2025). 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:

$$\mathcal{L}_{\text{MoFe}} = \max(\ell(\hat{y}^+, y) - \ell(\hat{y}^-, y), 0), \quad (9)$$

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:

$$\mathcal{L}_{\text{fusion}} = \ell(\hat{y}^+, y) + \ell(\hat{y}^-, y) + \lambda_{\text{MoFe}} \mathcal{L}_{\text{MoFe}}. \quad (10)$$

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:

$$\mathcal{L}_{\text{total}} = \underbrace{\mathcal{L}_{\text{WSI}} + \mathcal{L}_{\text{TMA}}}_{\text{Stage 1}} + \underbrace{\mathcal{L}_{\text{GA}}}_{\text{Stage 2}} + \underbrace{\mathcal{L}_{\text{fusion}}}_{\text{Stage 3}}. \quad (11)$$

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  $\mathcal{L}_{\text{GA}}$  applied Eq. (6). The fusion objective  $\mathcal{L}_{\text{fusion}}$  follows Eq. (10). 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, which are (i) pathology semantics, (ii) cross-modality compatibility, and (iii) robustness to missing data.

## 3. Experiments

### 3.1. Experimental Setup

We evaluate HAF on the HANCOCK head and neck cancer cohort (Dörrieh 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 and TMA histopathology, we use the 1024-dimensional UNI (Chen et al., 2024) patch embeddings released by HANCOCK and apply the same UNI encoder to raw TMA cores, respectively.

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—**WSI-CLAM**, **TMA-CLAM**, and **WSI+TMA-CLAM**—serve as unimodal and dual-modality pathology references, and we additionally reproduce WSI+TMA-CLAM using the released UNI features to ensure evaluation consistency. Moving beyond pathology, we include two naïve 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.

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

Method	Accuracy	AUC
<i>Pathology-only baselines</i>		
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
<i>Fusion models and HAF variants</i>		
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)	<b>0.752±0.047</b>	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
<b>HAF (GA + Random Drop)</b>	0.745±0.065	<b>0.739±0.092</b>
<i>Comparable multimodal methods</i>		
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

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—**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 naïve 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 multi-modal 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.

### 3.4. Ablation on Stage-wise Detachment

Training with gradient detachment consistently yields more stable and effective optimization. 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.

Table 2: **Ablation 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	<b>0.748±0.046</b>	0.738±0.063
	AUC	0.694±0.113	<b>0.698±0.110</b>
+ Global Alignment	Accuracy	<b>0.752±0.047</b>	0.752±0.050
	AUC	0.698±0.127	<b>0.727±0.108</b>
+ Random Drop	Accuracy	<b>0.748±0.074</b>	0.748±0.081
	AUC	<b>0.715±0.099</b>	0.714±0.101
<b>HAF</b>	Accuracy	0.745±0.065	<b>0.748±0.052</b>
	AUC	<b>0.739±0.092</b>	0.721±0.098

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



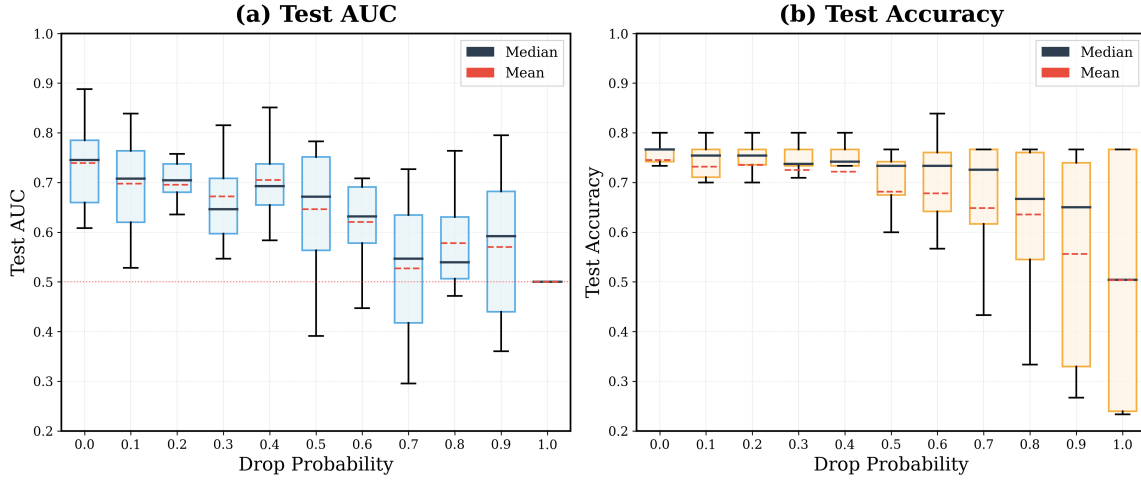


Figure 2: Model robustness under modality dropout.

for missing modalities by relying on reliable inputs, and degrade only when most modalities are absent ( $p > 0.4$ ), demonstrating improved resilience to real-world incompleteness and noise.

### 3.6. Representation Analysis

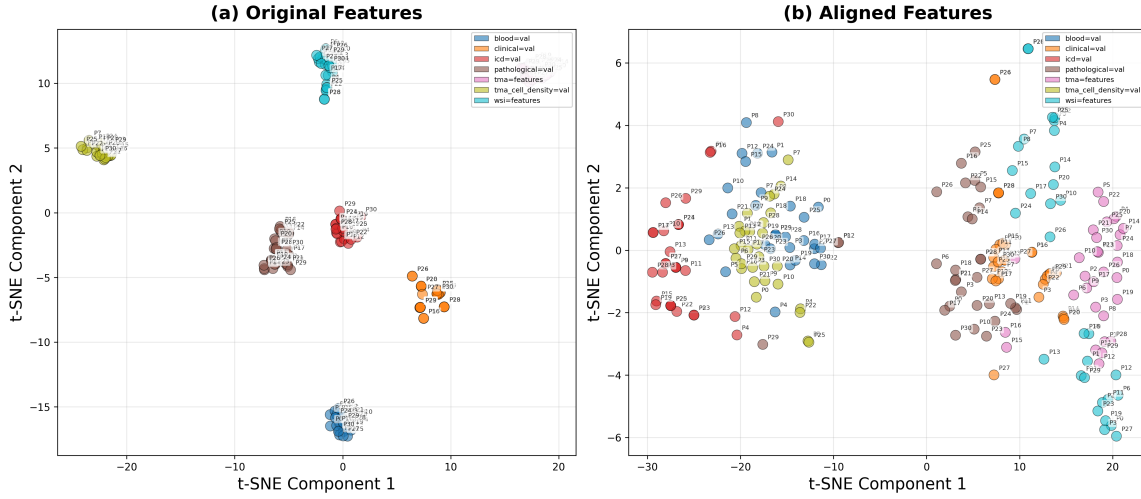


Figure 3: t-SNE projection of multimodal embeddings before and after global alignment.

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.

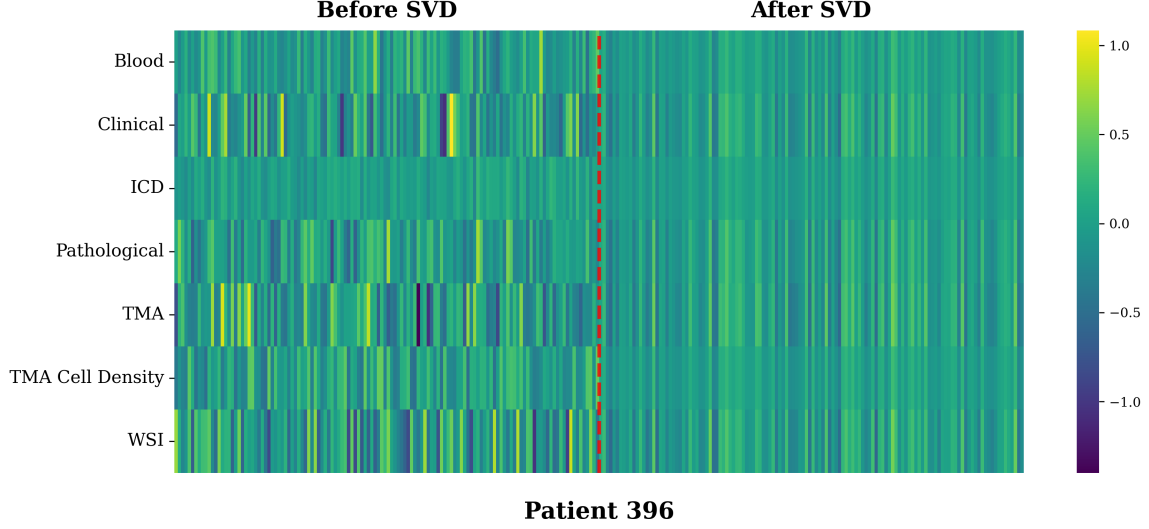


Figure 4: **Heatmap of aligned modality representations for a representative patient.**

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.

#### 4. Conclusion

We presented **HAF (Heterogeneous Aligned Fusion)**, a staged and detached multi-modal fusion framework that mitigates cross-objective interference, effectively exploits complementary prognostic signals, and enhances robustness to missing modalities in survival prediction. Future work will focus on developing stronger, task-aware alignment mechanisms that encourage modalities to compensate for each other rather than collapse toward a single dominant source, as well as incorporating more fine-grained spatial reasoning and prospective clinical validation. In addition, we plan to evaluate the proposed pipeline across a broader range of datasets to further verify its robustness and generalization.

## 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.
- 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 Quéllec. 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 multimodal 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.
- Wei Yuan, Yijiang Chen, Biyue Zhu, Sen Yang, Jiayu Zhang, Ning Mao, Jinxi Xiang, Yuchen Li, Yuanfeng Ji, Xiangde Luo, et al. Pancancer outcome prediction via a unified weakly supervised deep learning model. *Signal transduction and targeted therapy*, 10(1): 285, 2025.
- 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 typically integrate WSI with a small set of additional descriptors. MDLM (Tian et al., 2025) uses co-attention between WSI and clinical features, ProgPath (Yuan et al., 2025) extends multimodal survival prediction to large cohorts using Cox and ranking losses.

However, these systems generally: (1) rely on a narrow subset of modalities (typically two sources such as WSI+clinical or WSI+pathway), which does not reflect the richness and heterogeneity of real-world clinical data; (2) perform late or shallow fusion without explicit cross-modality alignment; and (3) assume fully observed inputs, lacking mechanisms to remain reliable under missing or noisy modalities.

### 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 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.