Toward Resilient Medical Multimodal AI: A Framework for Missing Modality Recovery Under Clinical Constraints

Jiahe Hou[1,2], **John Moraros**[1,2], **Guangliang Cheng**[2], **Shuihua Wang**[1,2*]

[1]School of Science, Xi'an Jiaotong-Liverpool University, Suzhou 215000, China
[2]Department of Computer Science, University of Liverpool, Liverpool L69 3BX, UK
Jiahe.Hou24@student.xjtlu.edu.cn, John.Moraros@xjtlu.edu.cn, Guangliang.Cheng@liverpool.ac.uk,
Shuihua.Wang@xjtlu.edu.cn

Abstract

Real-world medical AI systems rarely operate under ideal conditions: multimodal clinical data are frequently incomplete due to resource constraints, equipment limitations, and workflow disruptions—challenges amplified in under-resourced hospitals and community clinics where infrastructure gaps make missing modalities the norm. Existing multimodal learning methods typically assume data completeness or treat missing modalities as a narrow technical problem, limiting their reliability in practice.

We posit that missing-modality learning should be reframed as a core challenge of **AI resilience in medicine**, where systems must maintain reliable performance under real-world constraints while transparently communicating uncertainty. To this end, we propose a **two-tier framework aligned with clinical workflows**. Tier 1 enables efficient, uncertainty-aware inference under missing modalities using prompt-enhanced modality encoding and learnable pseudo-embeddings. Cases with elevated uncertainty are routed to Tier 2, which performs high-fidelity generative recovery using conditional diffusion models for precision-critical decisions.

Component-level validation on a multi-center colorectal cancer cohort (1,679 patients, four centers) demonstrates that each design element contributes measurably to downstream prognostic performance: learnable pseudo-embeddings improve the concordance index (C-index) by **10.1 points** over zero-filling in feature space, while missing-aware prompts and modality-aware prompts contribute **4.4** and **3.7** points, respectively. Full evaluation of uncertainty-guided routing and generative recovery across brain tumor, cardiac, and CT–MRI benchmarks is ongoing.

Introduction

Multimodal clinical data are rarely complete in practice: imaging may be unavailable due to equipment limitations or contraindications, records are fragmented across institutions, and data quality varies with acquisition protocols (Zhang et al. 2022; Lipkova et al. 2022). In under-resourced settings—rural hospitals, community clinics, and low-income regions—missing modalities are the norm, making reliable AI under such constraints inseparable from the goal of resilient healthcare communities. This challenge extends beyond imaging: multimodal data fusion across sensors, electronic medical records, and biosignals faces similar completeness assumptions that frequently fail in practice (Rehman et al. 2025; Shah et al. 2025).

Existing approaches frame missing modalities as a *robustness* problem, seeking to preserve performance through imputation, modality dropout, or learning-based recovery (Zhou et al. 2026; Liu et al. 2025; Qian et al. 2025). We argue this is insufficient: in clinical settings, the challenge is *resilience*—systems must degrade gracefully, adapt to available resources, and communicate uncertainty transparently rather than producing confident predictions from silently inferred data (Chen et al. 2022). The consequences of silent failure are concrete: Lipkova et al. (2022) report that standard multimodal survival models produce confidently incorrect risk stratifications when a modality is absent without notification, potentially altering treatment decisions for patients whose true risk category would differ under complete data.

This distinction has direct design implications. High-fidelity generative models offer detailed reconstruction but impose costs unsuitable for time-critical triage (Boehm et al. 2022), while lightweight inference supports rapid screening but lacks reliability for high-stakes decisions. Treating missing-modality learning as a uniform problem obscures these trade-offs.

We propose a two-tier framework structured around clinical workflows. Tier 1 prioritizes efficiency and uncertainty awareness for resource-constrained settings; Tier 2 provides high-fidelity generative recovery for precision-critical decisions. Uncertainty estimates govern routing between tiers, ensuring inferred information is contextualized rather than silently trusted.

Our contributions are as follows: (1) we reframe missing-modality learning as an AI resilience challenge, shifting the design focus from algorithmic robustness to clinically grounded, workflow-aware adaptation; (2) we propose a two-tier architecture that integrates prompt-enhanced modality encoding with uncertainty-guided routing, enabling resource-adaptive deployment across heterogeneous clinical environments; and (3) we outline a research agenda with concrete hypotheses and evaluation protocols toward trustworthy medical multimodal AI under data imperfection.

*Corresponding author.

Related Work

Missing-modality learning has been addressed through several paradigms. *Feature-level methods* learn modality-agnostic representations, with recent work targeting modality imbalance (Zhou et al. 2026) and decoupling shared from modality-specific features for arbitrary missing patterns (Liu et al. 2025). *Attention-based fusion* dynamically weights available modalities through cross-modal mechanisms, including co-attention over histopathology–genomic features (Chen et al. 2022) and transformers unifying heterogeneous inputs with learned modality embeddings (Zhou et al. 2023); however, these methods do not explicitly model uncertainty over inferred representations. *Generative recovery* synthesizes missing modalities using VAEs or diffusion models, with recent work combining generation with risk stratification (Qian et al. 2025), though late fusion alternatives sacrifice deep inter-modal interactions for modularity (Boehm et al. 2022).

Despite this progress, existing approaches can be mapped onto our two-tier framing to clarify the gap each leaves. Feature-level methods (Zhou et al. 2026; Liu et al. 2025) and attention-based fusion (Chen et al. 2022; Zhou et al. 2023) achieve efficient inference suitable for Tier 1, but lack uncertainty quantification over inferred representations and cannot flag cases that exceed their reconstruction capacity. Generative recovery methods (Qian et al. 2025) provide Tier 2-level fidelity but impose uniform computational cost regardless of case difficulty, while late fusion alternatives (Boehm et al. 2022) sacrifice deep inter-modal interactions for modularity. Beyond imaging, analogous completeness challenges arise in EEG-based biometric systems requiring robust fusion across heterogeneous sensor streams (Rehman et al. 2025), natural language understanding under resource-constrained settings where data scarcity limits model reliability (Ali et al. 2023), and data-mining frameworks that must reason over historically incomplete patient records to identify clinical risk factors (Halim et al. 2023)—reinforcing that missing-modality resilience is a cross-cutting concern, not an imaging-specific problem. Critically, no prior framework jointly provides efficient lightweight recovery, uncertainty-aware escalation, and high-fidelity generation within a single clinically grounded architecture. Our work organizes these capabilities within a resilience-oriented architecture defined by clinical workflows rather than algorithmic convenience.

Methodology

We propose a two-tier architecture (Figure 1) operationalizing resilience through explicit modality awareness, uncertainty-guided routing, and adaptive recovery. The two tiers are trained sequentially: Tier 1 is optimized first, then Tier 2 is trained with frozen Tier 1 estimates as conditioning input.

Tier 1: Prompt-Enhanced Lightweight Recovery

Let $\mathcal{M} = \{m_1, \ldots, m_K\}$ denote the set of K modalities. For patient i, we observe a subset $\mathcal{M}_i^{\text{obs}} \subseteq \mathcal{M}$ with features

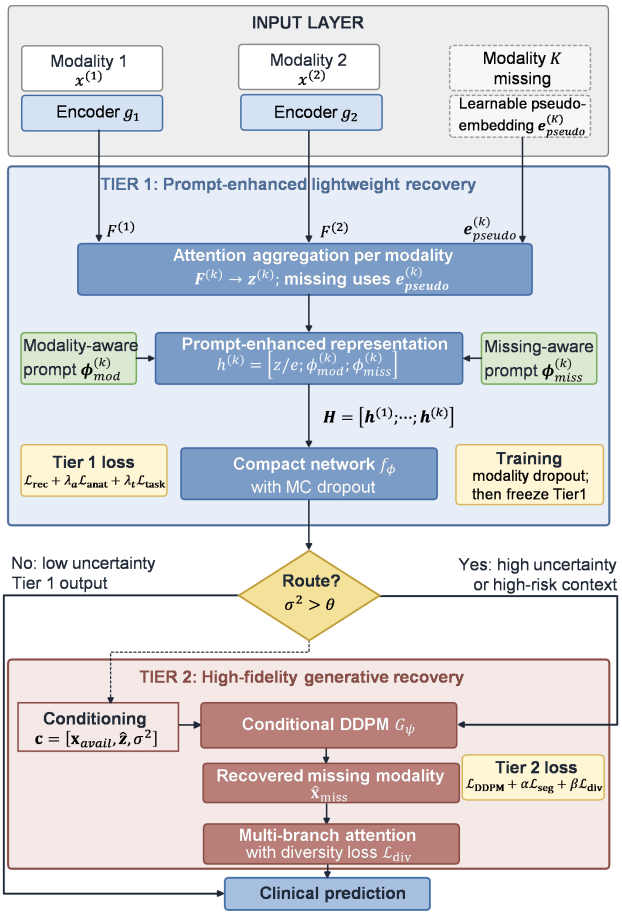

Figure 1. Two-tier framework with prompt-enhanced missing modality handling and uncertainty-based routing.

$\{\mathbf{x}_i^{(k)}\}_{k \in \mathcal{M}_i^{\text{obs}}}$. Each modality is encoded via a modality-specific encoder g_k, producing feature set $\mathbf{F}_i^{(k)} \in \mathbb{R}^{N_i^{(k)} \times d}$ where $N_i^{(k)}$ varies by instance (e.g., number of image patches or report sections). We aggregate each modality's variable-length features through a learnable attention mechanism:

$$\mathbf{z}_i^{(k)} = \sum_{j=1}^{N_i^{(k)}} a_{i,j}^{(k)} \mathbf{F}_{i,j}^{(k)},$$
$$a_{i,j}^{(k)} = \frac{\exp\{\mathbf{w}^\top \tanh(\mathbf{V}\mathbf{F}_{i,j}^{(k)\top})\}}{\sum_{j'} \exp\{\mathbf{w}^\top \tanh(\mathbf{V}\mathbf{F}_{i,j'}^{(k)\top})\}} \quad (1)$$

where $\mathbf{w} \in \mathbb{R}^{d'}$ and $\mathbf{V} \in \mathbb{R}^{d' \times d}$ are learnable parameters.

To explicitly encode modality status, we augment each aggregated representation with *modality-aware prompts* $\boldsymbol{\phi}_{\text{mod}}^{(k)} \in \mathbb{R}^{d_p}$ (learnable embeddings identifying modality type) and *missing-aware prompts* $\boldsymbol{\phi}_{\text{miss}}^{(k)} \in \{0, 1\}$ (binary presence/absence indicators). When modality k is absent, a learnable *pseudo-embedding* $\mathbf{e}_{\text{pseudo}}^{(k)} \in \mathbb{R}^d$ substitutes for the

encoded feature. The prompt-enhanced representation is:

$$\mathbf{h}_i^{(k)} = \begin{cases} [\mathbf{z}_i^{(k)};\ \boldsymbol{\phi}_{\text{mod}}^{(k)};\ 1] & \text{if } k \in \mathcal{M}_i^{\text{obs}} \\ [\mathbf{e}_{\text{pseudo}}^{(k)};\ \boldsymbol{\phi}_{\text{mod}}^{(k)};\ 0] & \text{otherwise} \end{cases} \tag{2}$$

yielding $\mathbf{h}_i^{(k)} \in \mathbb{R}^{d+d_p+1}$.

Tier 1 then maps the concatenated prompt-enhanced representations $\mathbf{H}_i = [\mathbf{h}_i^{(1)}; \ldots; \mathbf{h}_i^{(K)}]$ through a compact network f_ϕ to estimate missing features $\hat{\mathbf{z}}_i^{(k)}$ for $k \notin \mathcal{M}_i^{\text{obs}}$. Predictive uncertainty is estimated via Monte Carlo (MC) dropout—chosen for its negligible computational overhead, consistent with Tier 1's efficiency constraints:

$$\sigma_i^2 = \frac{1}{T}\sum_{t=1}^{T} \|\hat{\mathbf{z}}_i^{(k,t)} - \bar{\mathbf{z}}_i^{(k)}\|^2, \quad \bar{\mathbf{z}}_i^{(k)} = \frac{1}{T}\sum_{t=1}^{T} \hat{\mathbf{z}}_i^{(k,t)} \tag{3}$$

The training objective balances reconstruction fidelity, anatomical plausibility, and downstream task utility:

$$\mathcal{L}_{\text{Tier1}} = \mathcal{L}_{\text{rec}} + \lambda_a \mathcal{L}_{\text{anat}} + \lambda_t \mathcal{L}_{\text{task}} \tag{4}$$

where $\mathcal{L}_{\text{anat}} = \sum_{r\in\mathcal{R}} \max(0,\ \|\hat{\mathbf{z}}_r - \boldsymbol{\mu}_r\| - \kappa\boldsymbol{\sigma}_r)$ penalizes reconstructions deviating beyond κ standard deviations from region-wise statistics $(\boldsymbol{\mu}_r, \boldsymbol{\sigma}_r)$ estimated from complete-data training samples. Here $\mathcal{R}$ indexes anatomically meaningful feature partitions (e.g., corresponding to tissue regions or organ substructures), κ controls the tolerance margin, and the hinge formulation ensures the penalty activates only when reconstructions fall outside the plausible range—acting as a learned anatomical prior that prevents hallucination of features inconsistent with observed population statistics. During training, modality-level augmentation randomly drops each modality with probability p (ensuring $|\mathcal{M}_i^{\text{train}}| \geq 1$), simulating real-world missingness patterns.

Uncertainty-Guided Routing

Cases are routed to Tier 2 when uncertainty exceeds threshold θ or clinical context demands precision:

$$\text{Route}(i) = \begin{cases} \text{Tier 2} & \text{if } \sigma_i^2 > \theta \text{ or } c_i \in \mathcal{C}_{\text{high-risk}} \\ \text{Tier 1} & \text{otherwise} \end{cases} \tag{5}$$

where θ is calibrated on held-out data by computing the uncertainty distribution over complete-data cases and setting θ at the 95th percentile of this distribution, targeting a referral rate that balances Tier 2 utilization against Tier 1 coverage. $\mathcal{C}_{\text{high-risk}}$ denotes predefined high-stakes clinical contexts (e.g., oncology staging, surgical planning) where Tier 2 is invoked regardless of uncertainty.

Tier 2: High-Fidelity Generative Recovery

Tier 2 employs a conditional diffusion model G_ψ synthesizing missing modality data $\hat{\mathbf{x}}_{\text{miss}}$ conditioned on available data and Tier 1 estimates:

$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\psi(\mathbf{x}_t, t, \mathbf{c})\|^2\right] \tag{6}$$

where $\mathbf{c} = [\mathbf{x}_{\text{avail}}, \hat{\mathbf{z}}, \sigma^2]$ is the conditioning context comprising the raw observed modality data $\mathbf{x}_{\text{avail}}$, Tier 1's feature-space estimates $\hat{\mathbf{z}}$ for missing modalities (frozen during Tier 2 training), and the associated uncertainty map σ^2 from MC dropout. This design allows Tier 2 to leverage both the raw fidelity of available data and the structural guidance provided by Tier 1's imputed features, while the uncertainty signal enables the diffusion model to allocate denoising capacity toward regions where Tier 1 was least confident. To extract diverse prognostically relevant features from synthesized outputs, we employ multi-branch attention with an exclusive diversity loss that maximizes inter-branch feature dissimilarity:

$$\mathcal{L}_{\text{div}} = -\frac{1}{|\mathcal{P}|}\sum_{(m,n)\in\mathcal{P}} \left(1 - \frac{\langle \mathbf{o}^{(m)}, \mathbf{o}^{(n)} \rangle}{\|\mathbf{o}^{(m)}\|\|\mathbf{o}^{(n)}\|}\right) \tag{7}$$

where $\mathbf{o}^{(m)}$ denotes the output of the m-th attention branch and $\mathcal{P}$ is the set of all branch pairs. The full Tier 2 objective combines generation quality, segmentation consistency, and feature diversity:

$$\mathcal{L}_{\text{Tier2}} = \mathcal{L}_{\text{DDPM}} + \alpha\mathcal{L}_{\text{seg}} + \beta\mathcal{L}_{\text{div}} \tag{8}$$

Configuration	C-index
Full model (all components)	0.812
w/o modality augmentation	0.793
w/o modality-aware prompts	0.775
w/o missing-aware prompts	0.768
w/o learnable pseudo-embeddings (zero-fill)	0.711
w/o diversity loss	0.796
single-branch (no multi-branch attention)	0.798

Table 1. Component ablation on the FUSCC cohort (overall survival prediction, C-index). C-index measures downstream prognostic task performance after missing-modality imputation, not reconstruction quality directly. Each row removes one component from the full model. All components are shared with our Tier 1 design.

Preliminary Results

While full evaluation of the two-tier framework is ongoing, we report component-level validation supporting our core design choices. Using a colorectal cancer prognostic prediction task across four independent clinical centers (1,679 patients, three modalities: pathology, radiology, clinical reports) (Qu et al. 2026), we isolate the contribution of each missing-modality component shared with our framework.

As shown in Table 1, each component contributes meaningfully: learnable pseudo-embeddings yield the largest individual gain (+10.1 points over zero-filling in the *feature* space, i.e., replacing $\mathbf{e}_{\text{pseudo}}^{(k)}$ with $\mathbf{0} \in \mathbb{R}^d$ for absent modalities), followed by missing-aware prompts (+4.4) and modality-aware prompts (+3.7). The multi-branch attention with diversity loss provides complementary improvements (+1.4 and +1.6, respectively). These results empirically validate the component design underlying our Tier 1 architecture.

Critically, these results reflect a *single-tier* system without uncertainty-guided routing or generative recovery. We hypothesize that the full two-tier design will yield further

gains: uncertainty routing should reduce silent failures by directing low-confidence cases to Tier 2, while DDPM-based recovery should provide higher-fidelity feature reconstruction for the subset of cases where Tier 1 confidence is insufficient.

Discussion and Future Work

Our component-level results confirm that explicit modality encoding—prompts, pseudo-embeddings, and modality augmentation—provides substantial and additive gains under missing modalities, validating the design foundation of our framework. The key open question is whether the two-tier architecture with uncertainty-guided routing yields benefits beyond what single-tier approaches achieve.

Our ongoing evaluation targets three axes: (1) whether uncertainty routing improves clinical calibration by reliably identifying cases that exceed Tier 1's reconstruction capacity; (2) whether the two-tier design achieves diagnostic accuracy comparable to complete-modality baselines while substantially reducing computational cost for routine cases handled by Tier 1 alone; and (3) whether DDPM-based recovery in Tier 2 produces anatomically valid reconstructions that outperform feature-level imputation for high-stakes decisions. We also anticipate investigating expected failure modes: in particular, we hypothesize that routing to Tier 2 may degrade when *all* or nearly all modalities are absent, leaving insufficient conditioning signal for the diffusion model to produce reliable reconstructions. In such extreme-missingness scenarios, the framework should ideally abstain rather than generate overconfident outputs, and we plan to evaluate whether the uncertainty calibration remains well-behaved under these boundary conditions. Experiments on BraTS (brain tumor segmentation), M&Ms (cardiac imaging), and multi-center CT–MRI datasets are in progress.

More broadly, this work frames the missing-modality problem in medical AI not as algorithmic robustness but as *system-level resilience*—designing for graceful degradation, resource adaptation, and transparent uncertainty communication. This framing aligns naturally with the goals of resilient healthcare communities, where infrastructure heterogeneity is the norm.

References

Ali, N.; Tubaishat, A.; Al-Obeidat, F.; Shabaz, M.; Waqas, M.; Halim, Z.; Rida, I.; and Anwar, S. 2023. Towards Enhanced Identification of Emotion from Resource-Constrained Language Through a Novel Multilingual BERT Approach. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Boehm, K. M.; Aherne, E. A.; Ellenson, L.; Nikolovski, I.; Alghamdi, M.; Vázquez-García, I.; Zamarin, D.; Long Roche, K.; Liu, Y.; Patel, D.; et al. 2022. Multimodal Data Integration Using Machine Learning Improves Risk Stratification of High-Grade Serous Ovarian Cancer. *Nature Cancer*, 3(6): 723–733.

Chen, R. J.; Lu, M. Y.; Williamson, D. F. K.; Chen, T. Y.; Lipkova, J.; Noor, Z.; Shaban, M.; Shady, M.; Williams, M.; Joo, B.; and Mahmood, F. 2022. Pan-Cancer Integrative Histology-Genomic Analysis via Multimodal Deep Learning. *Cancer Cell*, 40(8): 865–878.e6.

Halim, Z.; Khan, G.; Shah, B.; Naseer, R.; Anwar, S.; and Shah, A. 2023. On the Utility of Parents' Historical Data to Investigate the Causes of Autism Spectrum Disorder: A Data Mining-Based Framework. *IRBM*, 44(4): 100780.

Lipkova, J.; Chen, R. J.; Chen, B.; Lu, M. Y.; Barbieri, M.; Shao, D.; Vaidya, A. J.; Chen, C.; Zhuang, L.; Williamson, D. F. K.; and Mahmood, F. 2022. Artificial Intelligence for Multimodal Data Integration in Oncology. *Cancer Cell*, 40(10): 1095–1110.

Liu, H.; Shi, Y.; Xu, Y.; Li, A.; and Wang, M. 2025. Agnostic-Specific Modality Learning for Cancer Survival Prediction from Multiple Data. *IEEE Journal of Biomedical and Health Informatics*, 29(9): 6311–6322.

Qian, X.; Pei, J.; Han, C.; Liang, Z.; Zhang, G.; Chen, N.; Zheng, W.; Meng, F.; Yu, D.; Chen, Y.; Sun, Y.; Zhang, H.; Qian, W.; Wang, X.; Er, Z.; Hu, C.; Zheng, H.; and Shen, D. 2025. A Multimodal Machine Learning Model for the Stratification of Breast Cancer Risk. *Nature Biomedical Engineering*, 9(3): 356–370.

Qu, L.; Zhang, C.; Hou, Y.; Tang, F.; Sheng, W.; Huang, D.; and Song, Z. 2026. Foundation Model-Enabled Multimodal Deep Learning for Prognostic Prediction in Colorectal Cancer with Incomplete Modalities: A Multi-Institutional Retrospective Study. *Advanced Science*, e10931.

Rehman, T. U.; Alruwaili, M.; Siddiqi, M. H.; Alhwaiti, Y.; Anwar, S.; Halim, Z.; and Alam, M. 2025. Advancing EEG-Based Biometric Identification Through Multi-Modal Data Fusion and Deep Learning Techniques. *Complex & Intelligent Systems*, 11(9): 398.

Shah, K. A.; Halim, Z.; Anwar, S.; Hsu, C.-C.; and Rida, I. 2025. Multi-Sensor Data Fusion for Smart Healthcare: Optimizing Specialty-Based Classification of Imbalanced EMRs. *Information Fusion*, 103503.

Zhang, Y.; He, N.; Yang, J.; Li, Y.; Wei, D.; Huang, Y.; Zhang, Y.; He, Z.; and Zheng, Y. 2022. mmFormer: Multimodal Medical Transformer for Incomplete Multimodal Learning of Brain Tumor Segmentation. In *Proc. MICCAI*, 107–117.

Zhou, C.; Wang, M.; Shi, Y.; Zhang, A.; and Li, A. 2026. Understanding and Tackling the Modality Imbalance Problem in Multimodal Survival Prediction. *Pattern Recognition*, 172: 112398.

Zhou, H.-Y.; Yu, Y.; Wang, C.; Zhang, S.; Gao, Y.; Pan, J.; Shao, J.; Lu, G.; Zhang, K.; and Li, W. 2023. A Transformer-Based Representation-Learning Model with Unified Processing of Multimodal Input for Clinical Diagnostics. *Nature Biomedical Engineering*, 7(6): 743–755.