RAM-Missing: Retrieval-Augmented Missing-Aware Fusion for Robust Lung Cancer Subtyping

Fulin Li[1], Qing Cai[2⋆], Shugang Zhang[1], and Jinxing Li[3]

[1] Ocean University of China, Qingdao, 266100, China
lifulin@stu.ouc.edu.cn, zsg@ouc.edu.cn
[2] Hefei University of Technology, Hefei, 230009, China
caiqing@hfut.edu.cn
[3] Harbin Institute of Technology, Harbin, 150001, China
lijinxing158@hit.edu.cn

Abstract. In real-world clinical settings, multimodal medical data are often characterized by severe modality imbalance and structured missingness. In our tri-modal lung cancer cohort (WSI, CT, RNA), 75.4% of patients lack CT scans. Most existing multimodal learning methods implicitly assume complete modality availability or rely on naive imputation, leading to catastrophic performance collapse when applied to highly incomplete deployment data. To address this challenge, we propose RAM-Missing, a Retrieval-Augmented Missing-aware fusion framework designed for robust lung cancer subtyping under severe missingness. Specifically, our framework introduces three key components: (1) We design a Retrieval-Augmented Proxy Generator that constructs proxy embeddings for missing CTs by querying a historical memory bank. (2) We introduce a Retrieval Uncertainty Quantifier that mathematically estimates retrieval entropy to gauge the reliability of these proxies. (3) We propose an Uncertainty-Guided Conditional Router that dynamically allocates fusion experts based on the modality mask and uncertainty signal, effectively preventing the model from over-trusting low-quality imputed features. Extensive experiments using patient-level 5-fold cross-validation demonstrate that RAM-Missing consistently mitigates catastrophic degradation under extreme missing ratios, achieving superior prevalence-weighted AUROC and cross-fold stability compared to conventional fusion strategies. The source code is publicly available at https://github.com/Flliin/RAMMissing.

Keywords: Incomplete multimodal learning · Modality missingness · Retrieval-augmented fusion · Uncertainty routing · Lung cancer subtyping.

Multimodal learning is increasingly reshaping lung cancer diagnostic modeling from isolated single-source prediction to collaborative decision-making across imaging–pathology–molecular measurements. Large-scale cancer cohorts such

⋆ Corresponding author.

as The Cancer Genome Atlas (TCGA) have systematically characterized key molecular alterations of lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC), providing a principled foundation for cross-modal joint modeling [3–10].

Recent advances demonstrate that integrating CT imaging, WSI pathology, and molecular profiles significantly improves diagnostic and prognostic performance compared to unimodal systems [8–10,12,13]. Radiogenomic and radiopathomic studies further reveal strong cross-scale correlations between imaging phenotypes and genomic alterations in LUAD [11,12,28]. These multimodal oncology systems align with broader trends in multimodal biomedical AI, disease diagnosis, visual representation learning, and vision-language modeling [1,2,14–16,29–31], highlighting the potential of joint representation learning across morphological and molecular domains to improve LUAD/LUSC discrimination.

Despite these advances, real-world clinical deployment profoundly violates the complete-modality assumption. High-cost imaging modalities such as CT are often available only for a subset of patients, whereas pathology and transcriptomic data are more routinely collected. In our tri-modal LUAD/LUSC cohort, 75.4% of patients lack CT scans, creating a severe modality imbalance that introduces a deployment gap between research and practice.

Existing missing-modality learning frameworks, including HeMIS [17], U-HVED [18], SMIL [19], M3AE [20], ShaSpec [21], and recent missing-representation learning approaches [22], attempt to project available modalities into shared latent spaces or learn modality-agnostic representations. However, when a highly informative modality such as CT is systematically scarce, naive zero-imputation or static aggregation often leads to under-utilization of the scarce modality during training or catastrophic performance degradation under realistic missing patterns at inference. A robust lung cancer diagnostic system must therefore not only approximate the missing information, but also explicitly quantify its reliability before fusion.

To this end, we propose **RAM-Missing**, a Retrieval-Augmented Missing-aware fusion framework tailored for robust lung cancer subtyping under severe structured missingness. Unlike existing approaches that passively impute missing data, our core innovation lies in actively quantifying and circumventing imputation risks through a self-preservation architectural design.

To summarize, our contributions are:

- We propose a novel Retrieval-Augmented Missing-aware fusion framework (RAM-Missing) to proactively tackle severe structured missingness and bridge the real-world clinical deployment gap.
- We design a **Retrieval-Augmented Proxy Generator** coupled with an explicit **Uncertainty Quantifier** to autonomously construct robust proxy CT embeddings by querying a historical memory bank and mathematically gauging their reliability via retrieval entropy.
- We design a bespoke **Uncertainty-Guided Conditional Router** to actively bypass noisy proxies and prevent catastrophic performance collapse by embed-

ding the modality mask and retrieval uncertainty into a Mixture-of-Experts (MoE) gating mechanism [24].
– Extensive experiments on a tri-modal TCGA cohort demonstrate that RAM-Missing exhibits superior damage-mitigation capability, cross-fold stability, and prevalence-weighted performance compared to standard fusion baselines.

1 Methods

Problem Setup and Overall Architecture. We consider a clinically realistic tri-modal learning setting for lung cancer subtyping (LUAD vs. LUSC) involving whole-slide images (WSI), CT scans, and RNA expression profiles. For each patient i, modality-specific encoders produce dense embeddings $\mathbf{z}_i^w$, $\mathbf{z}_i^c$, and $\mathbf{z}_i^r$. Since CT is frequently missing in practice, we denote modality availability using a binary mask $\mathbf{m}_i = [m_i^w, m_i^c, m_i^r] \in \{0,1\}^3$. Our objective is to learn a fusion mapping that is robust to structural missingness ($m_i^c = 0$).

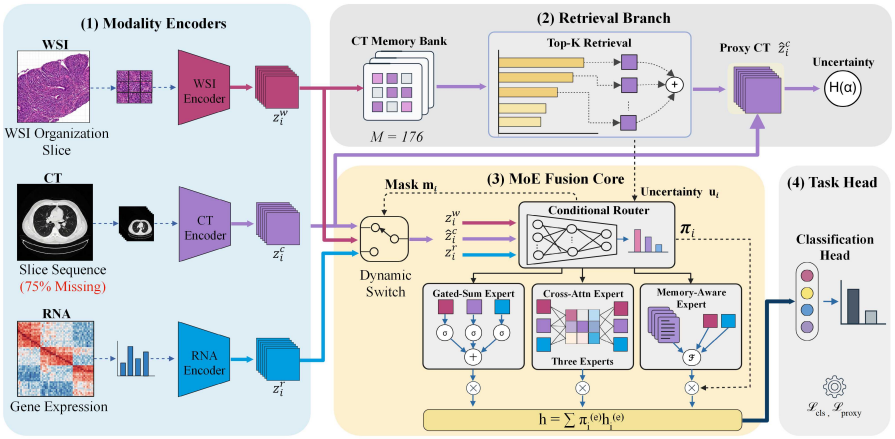

Fig. 1. Overall architecture of the proposed RAM-Missing framework. The pipeline consists of four main modules: (1) **Modality Encoders** extract base features from WSI, CT, and RNA. (2) When CT is missing, the **Retrieval Branch** constructs a proxy CT embedding ($\tilde{\mathbf{z}}_i^c$) from a historical CT Memory Bank and estimates the retrieval uncertainty (u_i). (3) The **MoE Fusion Core** utilizes a Dynamic Switch and a Conditional Router—guided by the modality mask ($\mathbf{m}_i$) and uncertainty (u_i)—to dynamically assign sparse weights to three complementary experts. (4) The fused representation ($\mathbf{h}_i$) is fed into the **Task Heads** for robust classification.

As explicitly illustrated in Fig. 1, the proposed RAM-Missing framework consists of four sequential stages: Modality Encoders, a Retrieval Branch, an MoE Fusion Core, and Task Heads.

(1) Modality Encoders. Given a patient i, we first extract dense feature representations from the available modalities.

- WSI Encoder: We utilize a ResNet-50 backbone pre-trained on ImageNet. The slide is tiled into non-overlapping patches, and features are aggregated via Global Average Pooling to produce a slide-level embedding $\mathbf{z}_i^w \in \mathbb{R}^{1024}$.
- CT Encoder: Similarly, for CT scans, we employ a 3D-ResNet backbone to extract volumetric features, resulting in a compact embedding $\mathbf{z}_i^c \in \mathbb{R}^{1024}$.
- RNA Encoder: High-dimensional gene expression data is projected into the same latent dimension using a 3-layer Multi-Layer Perceptron (MLP) with ReLU activation and BatchNorm, yielding $\mathbf{z}_i^r \in \mathbb{R}^{1024}$.

These aligned unimodal representations serve as the foundation for the subsequent retrieval and dynamic fusion stages.

(2) Retrieval Branch: Proxy Construction & Uncertainty. This module is explicitly activated when CT is unavailable ($m_i^c = 0$) to prevent the catastrophic degradation of the latent space caused by naive zero-masking. Proxy Construction: We construct a proxy embedding from a memory bank $\mathcal{B} = \{(\mathbf{k}_j, \mathbf{v}_j)\}_{j=1}^M$ built strictly from the training subset with complete modalities (M is the number of complete cases). The value $\mathbf{v}_j = \mathbf{z}_j^c$ stores the historical CT features, and the key $\mathbf{k}_j$ is a concatenated representation of the available WSI and RNA queries ($\mathbf{k}_j = [\mathbf{z}_j^w \| \mathbf{z}_j^r]$). For an incomplete patient i, we compute cosine similarities s_{ij} between their available query features and the bank keys. The top-K neighbors are retrieved and aggregated via a softmax-weighted sum to form the proxy CT:

$$\tilde{\mathbf{z}}_i^c = \sum_{j \in \mathrm{TopK}(i)} \alpha_{ij} \mathbf{v}_j, \quad \alpha_{ij} = \frac{\exp(s_{ij}/\tau_r)}{\sum_{k \in \mathrm{TopK}(i)} \exp(s_{ik}/\tau_r)}. \tag{1}$$

The effective CT embedding is dynamically routed by a physical switch, defined as $\hat{\mathbf{z}}_i^c = \mathbf{z}_i^c$ if $m_i^c = 1$, otherwise $\tilde{\mathbf{z}}_i^c$. The retrieved proxy is used as phenotype-level evidence rather than an exact CT reconstruction. By aggregating top-K neighbors instead of relying on a single nearest case, the retrieval branch provides a conservative approximation that is further controlled by the downstream uncertainty-guided router.

Uncertainty Estimation: Not all retrieved proxies are clinically reliable. If the patient's WSI/RNA profile is an outlier, the retrieval process may return irrelevant neighbors. We quantify this retrieval ambiguity using the Shannon entropy of the top-K similarity weights:

$$u_i = -\sum_{j \in \mathrm{TopK}(i)} \alpha_{ij} \log(\alpha_{ij} + \epsilon). \tag{2}$$

A high entropy u_i indicates a flat distribution (i.e., the query matches many distinct keys weakly), signaling high uncertainty. We emphasize that this entropy term is used as a retrieval reliability signal for routing, rather than as a fully calibrated probabilistic uncertainty estimate. This scalar u_i is a critical signal that warns the downstream MoE core to distrust the proxy.

(3) MoE Fusion Core: Missingness-Conditional Routing. To adaptively process inputs, we employ a Mixture-of-Experts (MoE) fusion core [24] with three complementary experts ($e \in \{1, 2, 3\}$):

- Gated-Sum Expert: Performs a simple weighted summation, providing a stable baseline fusion.
- Cross-Attention Expert: Models complex non-linear interactions between modalities using a Transformer block.
- Memory-Aware Expert: Specifically designed to align the retrieved proxy with the original WSI/RNA context.

The core innovation is the Conditional Router. Unlike standard MoE routers that depend only on input content, our router predicts sample-specific gating weights $\boldsymbol{\pi}_i$ conditioned on three factors: the feature content, the missingness mask $\mathbf{m}_i$, and the critical uncertainty signal u_i:

$$\boldsymbol{\pi}_i = \text{Softmax}\left(\frac{\text{MLP}([\mathbf{z}_i^w \| \hat{\mathbf{z}}_i^c \| \mathbf{z}_i^r \| \mathbf{m}_i \| u_i])}{\tau_g}\right), \quad \mathbf{h}_i = \sum_{e=1}^{3} \pi_i^{(e)} \mathbf{h}_i^{(e)}. \tag{3}$$

This design establishes a self-preservation mechanism: when $m_i^c = 0$ and u_i is high, the router learns to suppress experts that heavily rely on the (likely noisy) CT proxy, diverting the decision flow towards experts that focus on the reliable WSI and RNA features.

(4) Task Heads and Optimization. The universally fused patient representation $\mathbf{h}_i$ is fed into the Classification Head (a 2-layer MLP) for LUAD vs. LUSC prediction. The model is optimized end-to-end via Cross-Entropy loss ($\mathcal{L}_{cls}$). To explicitly enforce the quality of the proxy generation during training, we apply a proxy consistency loss $\mathcal{L}_{proxy} = \|\tilde{\mathbf{z}}_i^c - \mathbf{z}_i^c\|_2^2$ exclusively on patients with complete modalities. The total objective is $\mathcal{L} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{proxy}$.

2 Experiments

Dataset and Realistic Missingness. We evaluate our framework on a tri-modal TCGA lung cancer cohort (LUAD and LUSC) comprising WSI, CT, and RNA expression profiles [6, 7]. Crucially, only 24.6% of patients possess all three modalities, while 75.4% lack CT scans, reflecting structural missingness in the patient-level aligned cohort rather than an artificially simulated missing-modality setting. To ensure rigorous evaluation, we employ patient-level 5-fold cross-validation using StratifiedGroupKFold. This ensures that samples from the same patient (if any) are kept within the same fold, preventing subject leakage. The CT memory bank is constructed strictly within the training folds for each split; retrieval for validation/test patients is performed only against this isolated bank, guaranteeing zero cross-fold information leakage.

Task and Metrics. The objective is LUAD vs. LUSC subtyping. We evaluate models using both condition-specific AUROC and a prevalence-weighted AUROC ($\text{wAUC} = p_{\text{full}} \cdot \text{AUC}_{\text{full}} + p_{\text{miss}} \cdot \text{AUC}_{\text{miss}}$), where $p_{\text{full}} = 0.246$ and $p_{\text{miss}} = 0.754$. This weighted metric better reflects the expected performance in a real clinical deployment.

Baselines and Implementation Details. We benchmark against unimodal models and widely adopted fusion baselines, including Early Fusion (concatenation) [23], Cross-Attention [26], Vanilla MoE [24], and Simple Fusion [25, 27].

Table 1. Main comparison on LUAD vs. LUSC classification. **wAUC** follows the real-world modality prevalence (Full: 24.6%, Missing-CT: 75.4%). Best results are in bold.

Model	AUC	Acc	Full AUC	Missing-CT AUC	wAUC
WSI-only	0.680	0.690	–	–	–
RNA-only	0.640	0.650	–	–	–
CT-only	0.710	0.700	–	–	–
Early Fusion [23]	0.732	0.742	0.735	0.492	0.546
Cross-Attention [26]	0.735	0.744	0.739	0.498	0.552
Vanilla MoE [24]	0.543	0.508	0.543	0.510	0.518
Simple Fusion [27]	0.739	0.750	0.739	0.486	0.545
RAM-Missing (Ours)	**0.743**	**0.742**	**0.743**	**0.509**	**0.564**

For fair comparison, all baselines operate on the same pre-extracted modality embeddings and utilize zero-filling for the missing CT branch when necessary. RAM-Missing is implemented in PyTorch using the AdamW optimizer with an initial learning rate of 1×10^{-4} and weight decay of 1×10^{-3}. Models are trained for 100 epochs with early stopping (patience=10). We empirically set $K = 5$, $\tau_r = 0.1$, $\tau_g = 1.0$, and $\lambda = 0.5$. All experiments are conducted on a single NVIDIA RTX 3090 GPU.

2.1 Main Comparison and Damage Mitigation

Analysis of Main Results. Table 1 demonstrates the vulnerability of conventional multimodal architectures. CT serves as the most discriminative unimodal signal (AUC = 0.710). When all modalities are present, standard methods like Simple Fusion perform well (Full AUC = 0.739). However, under the realistic condition where CT is missing, naive fusion mechanisms experience catastrophic collapse. Specifically, Simple Fusion degrades to an AUC of 0.486 on Missing-CT cases—worse than random guessing (0.500) and far inferior to using WSI alone (0.680). This degradation indicates that the network's latent space is poisoned by the naive zero-imputed missing branch. In stark contrast, RAM-Missing successfully mitigates this collapse, pulling the Missing-CT AUC back up to 0.509 and achieving the highest prevalence-weighted wAUC (0.564). By explicitly quantifying retrieval uncertainty, our dynamic router successfully prevents the network from over-trusting unreliable CT proxies.

2.2 Ablation Study

Analysis of Ablation Variants. To isolate the contribution of each proposed module, we perform detailed ablation studies (Table 2).

(1) w/o Retrieval: Removing the memory bank and relying solely on zero-filling causes the Missing-CT AUC to plummet to 0.482, proving that explicitly constructing a proxy template is crucial. (2) w/o Router: Disabling the conditional

Table 2. Ablation study of RAM-Missing components under patient-level 5-fold cross-validation. Top performance is bolded.

Ablated Module	Full AUC	Missing-CT AUC	wAUC
w/o Retrieval	0.728	0.482	0.546
w/o Router	0.721	0.471	0.538
w/o Uncertainty	0.734	0.496	0.553
Full RAM-Missing	**0.743**	**0.509**	**0.564**

MoE routing forces the network into a static, uniform fusion path. This inflexible strategy yields the worst Missing-CT performance (0.471), highlighting the necessity of sample-wise adaptivity. (3) w/o Uncertainty: Crucially, removing the uncertainty signal (u_i) from the router's condition forces the network to blindly trust all retrieved proxies regardless of their match quality. This degrades the Missing-CT AUC from 0.509 to 0.496, empirically confirming that calibrating the fusion path with retrieval confidence is essential to filter out noisy proxies.

2.3 Robustness to Increasing CT Missingness

Analysis of Stress Testing. Beyond the empirical 75.4% prevalence, we comprehensively stress-test the models by synthetically elevating the CT missing ratio from 0% to 90% (Table 3).

Table 3. Robustness under increasing CT missing ratios (AUROC). All values are mean AUC across 5 folds. Best result per row is bolded.

CT missing ratio	Simple Fusion	Cross-Attention	RAM-Missing
0.00	**0.5370**	0.5327	0.5354
0.50	0.5298	0.5315	**0.5332**
0.70	0.5224	0.5231	**0.5485**
0.84	0.5291	0.5295	**0.5369**
0.90	0.5365	**0.5417**	0.5140

The absolute performance in this simulated regime is inherently lower due to dataset sub-sampling, but the relative trends are highly informative. RAM-Missing provides the most consistent gains in the critical moderate-to-high missing regimes (70%–84%), which mirror realistic clinical scenarios. In extreme missing conditions (90%), proxy retrieval lacks sufficient historical templates within the memory bank, which remains a fundamental challenge across all missing-data algorithms. Overall, RAM-Missing provides a significantly more stable trajectory compared to the erratic fluctuations of Simple Fusion.

3 Discussion

Our results expose a pervasive pitfall in clinical multimodal learning: evaluating complex fusion models solely on fully-intersected modalities yields a deceptive assessment of clinical utility. By adopting a prevalence-weighted evaluation protocol, we reveal that conventional fusion strategies (e.g., Early Fusion, Simple Fusion) often fail catastrophically in real-world incomplete deployments. While these methods perform competitively when data is complete, their rigid fusion mechanisms cannot adapt to the noise introduced by zero-filling missing modalities.

A defining observation from our experiments is that the primary clinical value of RAM-Missing lies in its damage-mitigation capability and cross-fold stability. In medical diagnostics, a model that behaves erratically or performs worse than a coin flip (AUC < 0.50) on a subset of patients is unacceptable, regardless of its average performance. RAM-Missing effectively acts as a safety net: when the retrieved CT proxy is potentially unreliable, the uncertainty-guided router can reduce reliance on proxy-dependent fusion paths and shift the decision toward the observed WSI/RNA evidence. This behavior improves the robustness of the fusion process while keeping the model interpretable at the level of routing decisions.

Furthermore, the robustness analysis (Table 3) highlights that our method maintains stability even as the missingness ratio climbs to 84%. This resilience is crucial for institutions with limited resources where high-cost modalities like CT are the exception rather than the norm. While the absolute performance gains may appear modest, the prevention of "catastrophic collapse" represents a significant step towards safe clinical deployment.

4 Conclusion

We propose RAM-Missing, a retrieval-augmented missing-aware fusion framework. By querying a historical CT memory bank to construct proxy embeddings and utilizing retrieval uncertainty to dynamically gate a Mixture-of-Experts fusion core, RAM-Missing actively prevents catastrophic failure on highly incomplete cases. The primary limitations of this study are the modest absolute discrimination performance, the use of retrieval proxies rather than true CT reconstruction, and the limited availability of external patient-level WSI/RNA/CT cohorts for full tri-modal validation. Future work will focus on integrating stronger vision-language foundation models, developing more principled cross-modal retrieval and uncertainty calibration, and validating the framework on external multi-institutional cohorts.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant No. 62471448; in part by the Shandong Provincial Natural Science Foundation under Grant No. ZR2024YQ004; and in part by the Taishan Scholars Youth Expert Program of Shandong Province under Grant No. tsqn202312109.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Waqas, A., Tripathi, A., Ramachandran, R.P., Stewart, P., Rasool, G.: Multimodal data integration for oncology in the era of deep neural networks: A review. Frontiers in Artificial Intelligence **7**, 1408843 (2024)
2. Topol, E.J.: High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine **25**(1), 44–56 (2019)
3. Coudray, N., Ocampo, P.S., Sakellaropoulos, T., et al.: Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nature Medicine **24**, 1559–1567 (2018)
4. Boehm, K.M., Khosravi, P., Vanguri, R., Gao, J., Shah, S.P.: Harnessing multimodal data integration to advance precision oncology. Nature Reviews Cancer **22**(2), 114–126 (2022)
5. Zhou, M., Leung, A., Echegaray, S., et al.: Non-small cell lung cancer radiogenomics map identifies relationships between molecular and imaging phenotypes with prognostic implications. Radiology **286**(1), 307–315 (2018)
6. The Cancer Genome Atlas Research Network: Comprehensive molecular profiling of lung adenocarcinoma. Nature **511**(7511), 543–550 (2014)
7. The Cancer Genome Atlas Research Network: Comprehensive genomic characterization of squamous cell lung cancers. Nature **489**(7417), 519–525 (2012)
8. Chen, R.J., Lu, M.Y., Wang, J., Williamson, D.F.K., Rodig, S.J., Lindeman, N.I., Mahmood, F.: Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis. IEEE Transactions on Medical Imaging **41**(4), 757–770 (2022)
9. Chen, R.J., Lu, M.Y., Weng, W.H., Chen, T.Y., Williamson, D.F.K., Manz, T., Shady, M., Mahmood, F.: Multimodal Co-Attention Transformer for Survival Prediction in Gigapixel Whole Slide Images. In: Proc. IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4015–4025 (2021)
10. 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., Mahmood, F.: Pan-cancer integrative histology-genomic analysis via multimodal deep learning. Cancer Cell **40**(8), 865–878.e6 (2022)
11. Fu, Y., Jung, A.W., Torne, R.V., et al.: Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. Nature Cancer **1**, 800–810 (2020)
12. Perez-Johnston, R., Araujo-Filho, J.A., Connolly, J.G., et al.: CT-based radiogenomic analysis of clinical stage I lung adenocarcinoma with histopathologic features and oncologic outcomes. Radiology **303**(3), 664–672 (2022)
13. Boehm, K.M., Aherne, E.A., Ellenson, L., et al.: Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. Nature Cancer **3**, 723–733 (2022)
14. Acosta, J.N., Falcone, G.J., Rajpurkar, P., Topol, E.J.: Multimodal biomedical AI. Nature Medicine **28**, 1773–1784 (2022)
15. Lipkova, J., Chen, R.J., Chen, B., et al.: Artificial intelligence for multimodal data integration in oncology. Cancer Cell **40**(10), 1095–1110 (2022)

16. Aerts, H.J.W.L., Velazquez, E.R., Leijenaar, R.T.H., et al.: Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature Communications **5**, 4006 (2014)
17. Havaei, M., Guizard, N., Chapados, N., Bengio, Y.: HeMIS: Hetero-Modal Image Segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2016. LNCS, vol. 9901, pp. 469–477. Springer, Cham (2016)
18. Dorent, R., Joutard, S., Modat, M., et al.: Hetero-Modal Variational Encoder-Decoder for Joint Modality Completion and Segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. LNCS, vol. 11765, pp. 74–82. Springer, Cham (2019)
19. Ma, M., Ren, J., Zhao, L., Tulyakov, S., Wu, C., Peng, X.: SMIL: Multimodal Learning with Severely Missing Modality. In: Proc. AAAI Conference on Artificial Intelligence, pp. 2302–2310 (2021)
20. Liu, H., Wei, D., Lu, D., Sun, J., Wang, L., Zheng, Y.: M3AE: Multimodal Representation Learning for Brain Tumor Segmentation with Missing Modalities. In: Proc. AAAI Conference on Artificial Intelligence **37**(2), 1657–1665 (2023)
21. Wang, H., Chen, Y., Ma, C., Avery, J., Hull, L., Carneiro, G.: Multi-modal Learning with Missing Modality via Shared-Specific Feature Modelling. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15878–15887 (2023)
22. Han, Z., Zhang, C., Fu, H., Zhou, J.T.: Trusted Multi-View Classification. In: International Conference on Learning Representations (ICLR) (2021)
23. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Proc. International Conference on Machine Learning (ICML), pp. 689–696 (2011)
24. Shazeer, N., Mirhoseini, A., Maziarz, K., et al.: Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. arXiv preprint arXiv:1701.06538 (2017)
25. Fedus, W., Zoph, B., Shazeer, N.: Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. Journal of Machine Learning Research **23**(120), 1–39 (2022)
26. Tsai, Y.-H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.-P., Salakhutdinov, R.: Multimodal Transformer for Unaligned Multimodal Language Sequences. In: Proc. Annual Meeting of the Association for Computational Linguistics (ACL), pp. 6558–6569 (2019)
27. Arevalo, J., Solorio, T., Montes-y-Gómez, M., González, F.A.: Gated multimodal units for information fusion. arXiv preprint arXiv:1702.01992 (2017)
28. Grossmann, P., Stringfield, O., El-Hachem, N., et al.: Defining the biological basis of radiomic phenotypes in lung cancer. eLife 6, e23421 (2017)
29. Xu, M., Wang, J., Tao, M., Bao, B.K., Xu, C.: CookGALIP: Recipe Controllable Generative Adversarial CLIPs With Sequential Ingredient Prompts for Food Image Generation. IEEE Transactions on Multimedia **27**, 2772–2782 (2025)
30. Li, S., Cai, Q., Zhang, F., Shu, Y., Li, G., Dong, J., Liu, L., Zhang, D.: LearnMat: Semantic-Aware Self-Supervision Fine-Grained Visual Recognition. IEEE Transactions on Image Processing **35**, 2555–2569 (2026)
31. Cai, Q., Wang, S., Zhang, F., Zhang, C., Liu, Y., Li, H., Dong, J., Zhang, D.: Multimodal artificial intelligence for disease diagnosis: Advances, applications, and challenges. Pattern Recognition **178**, 113456 (2026)