

R²MoE: Representation and Expert Selection Dual-Regularized Mixture-of-Experts for Multimodal Clinical Data

Wajih Hassan Raza¹ Mya Schiess² Juan Martinez Lemus² Timothy Michael Ellmore³
Charles Green² Claudio Soto² Xin Fu¹ Renjie Hu¹

¹University of Houston

²University of Texas Health Center at Houston - McGovern Medical School

³City College of New York

Abstract

Accurate prediction of complex diseases and patient outcomes depends on effectively combining heterogeneous clinical modalities—such as neuroimaging, genomics, and free-text notes. Although multimodal fusion methods, especially Mixture-of-Experts (MoE) architectures have shown strong performance on large clinical cohorts, naively applying standard MoE models to clinical multimodal inputs exposes two fundamental limitations. First, the strong variability in different modalities introduce highly non-uniform input patterns, which make it difficult for the gating network to learn stable and reliable expert assignments. Second, the resulting latent embeddings often lack coherent semantic structure and adequate inter-class separation, hindering robust clinical prediction. We propose R²MoE, a dual-regularized MoE framework that addresses these challenges. A specialized expert regularization anchors expert behavior to modality-driven patterns, preventing routing collapse and promoting stable expert coalitions. A complementary representation regularization shapes the fused latent embedding into a semantically structured and class-discriminative space. Furthermore, we aggregate layer-wise routing weights into a gating-weight fingerprint, capturing patient-specific computational pathways that reveal clinically meaningful subtype structure and enhance downstream prediction. Across ADNI and MIMIC-IV, R²MoE outperforms the best existing multimodal models, achieving relative improvements of 6.1% and 3.3% in classification accuracy, respectively, while also yielding more interpretable structure in the learned representation space.

1. Introduction

Accurate modeling of complex diseases such as Alzheimer’s and Parkinson’s, as well as patient outcomes such as mortality prediction, requires integrating diverse sources of patient information, including neu-

roimaging, genomics, biospecimens, clinical assessments, and clinical notes [11]. Each modality captures a distinct yet complementary view of patient health: imaging reflects structural or functional changes in organs, genomics encodes molecular risk factors, and clinical notes capture unstructured physician observations [24]. When combined, these modalities provide a more comprehensive and biologically grounded understanding of disease progression [3, 17], and have been shown to improve prediction of disease onset, progression, and treatment response compared to unimodal approaches [18, 23].

Given the substantial potential of multimodal integration, extensive research has explored fusion strategies, from early feature concatenation [26] to cross-attention transformers [1] and adaptive gating for modality weighting [29]. More recently, sparsely activated Mixture-of-Experts (MoE) architectures [30, 31] have emerged as a powerful paradigm for scalable and adaptive multimodal learning. By dynamically routing inputs to specialized subnetworks, MoEs can tailor computation to data-specific characteristics and achieve strong performance across diverse prediction tasks. In particular, MoE-based models have demonstrated remarkable results on representative large multimodal clinical datasets, such as ADNI for Alzheimer’s disease prediction and MIMIC-IV for patient mortality prediction [43].

However, MoE architectures were originally developed for large-scale language models (LLMs). When applied directly to clinical multimodal data—fundamentally different from language—they face two critical challenges. First, in LLMs, the gate routes based on token-level embeddings within a continuous and homogeneous semantic space. In contrast, clinical inputs are high-dimensional composites of heterogeneous modalities with discrete and uneven distributions. This makes routing boundaries difficult to learn: clinically similar patients may be sent to different experts, weakening specialization and interpretability. [2, 17]. Second, in natural language processing, experts often self-

organize into meaningful functional domains (e.g., syntax, semantics, contextual reasoning). However, in clinical multimodal data, where no shared semantic stream exists, expert behavior is frequently driven by other dominant artifacts—such as patterns of modality missingness—rather than underlying clinical or pathological factors [7, 27]. Without explicit regularization of expert selection, a naive application of MoE to clinical multimodal learning often creates a fragmented and non-discriminative latent space, with routing decisions dominated by noisy or overrepresented modalities. This instability ultimately degrades predictive performance and limits the model’s generalizability.

Another overlooked limitation of MoEs is that they do not inherently ensure semantically structured latent representations. Even with stable routing, the expert-generated embeddings may remain heterogeneous—samples from the same disease class can scatter, while those from different classes may overlap in latent space—hindering the model’s ability to capture shared disease semantics and clinically meaningful relationships [5, 9, 41].

To overcome this issue, it is necessary to further regularize the latent representation of MoE models. By imposing discriminative constraints—such as promoting intra-class compactness and inter-class separability—the model learns to organize patient embeddings in a geometrically coherent manner. This regularization aligns multimodal representations across experts and modalities, reinforces class boundaries, and ultimately enhances both the robustness and generalizability of clinical classification. Therefore, there is a critical need to design an MoE architecture with regularized latent representation and expert selection specifically tailored for clinical data. We propose a dual regularized MoE model ($\mathbf{R}^2\text{MoE}$), with several novel contributions:

- (i) We introduce an expert selection regularization that provides a stable inductive bias for expert specialization and prevents routing collapse, thereby stabilizing expert assignment under heterogeneous and missing modalities.

- (ii) We develop a representation regularization that imposes discriminative structure on the latent space of the MoE, ensuring semantically meaningful and clinically coherent embeddings.

- (iii) We leverage the expert routing weights to construct a Routing Fingerprint descriptor that captures layer-wise expert usage patterns for each patient, revealing interpretable subtype structure and further improving classification performance.

We evaluate the performance of $\mathbf{R}^2\text{MoE}$ against state-of-the-art models on two large multimodal clinical datasets, ADNI and MIMIC-IV, demonstrating improved predictive accuracy. We further visualize the learned representation space and routing patterns, which reveal meaningful clinical structure and interpretability.

2. Related Works

2.1. Multimodal MoE Models in Healthcare

Multimodal learning has become increasingly crucial for integrating heterogeneous medical data such as neuroimaging, genomics, and EHRs. Early approaches based on feature concatenation or canonical correlation analysis [20] fail to capture nonlinear dependencies. Most recent fusion strategies: attention-based [8], hierarchical integration [21], and adaptive gating [22], still assume full modality availability—which is rare in clinical settings. Missing data is often handled via imputation [19] or distillation [13], which risk bias or oversmoothing. Meanwhile, Mixture-of-Experts (MoE) architectures offer scalable specialization via sparse routing. The sparsely-gated MoE [31] and later variants [30] have shown strong performance in large-scale vision and language tasks. Medical applications have explored MoE-based fusion for multitask and heterogeneous data [6, 39], but routing is optimized only for prediction. Few works [10, 43] address missing modalities but leaves the latent space unconstrained, limiting structure and interpretability. Importantly, existing MoE multimodal models do not jointly regularize expert selection and fused representations, leaving expert behavior weakly constrained. Prior work also treats routing purely as an internal mechanism, ignoring the representational value of expert-weight distributions. To our knowledge, no prior method leverages these routing patterns or the conditional computation pathways they encode, despite their potential to reveal clinically meaningful heterogeneity and subtype-specific structure.

2.2. Latent Space Structuring

A key limitation of multimodal fusion methods is the lack of explicit structure in the latent space. Representation learning frameworks such as masked autoencoders (MAEs) [12] and contrastive self-supervision [33] align heterogeneous modalities while auxiliary reconstruction losses preserve modality-specific detail, improve robustness to missing data, and retain clinically meaningful signals [14, 32]. Clustering-based objectives additionally promote embeddings that reflect subgroups such as disease subtypes or outcome categories [36]. Deep clustering methods [16] have been applied to medical imaging and EHRs, enhancing interpretability and supporting biomarker discovery [4]. Yet these objectives are often used only as auxiliary components or post-hoc analyses rather than integrated into the core model. Critically, existing MoE-based multimodal architectures do not use clustering signals to influence routing or stabilize expert specialization. As a result, MoE routing remains unguided, allowing expert functions to drift or collapse and preventing the emergence of consistent subtype structure.

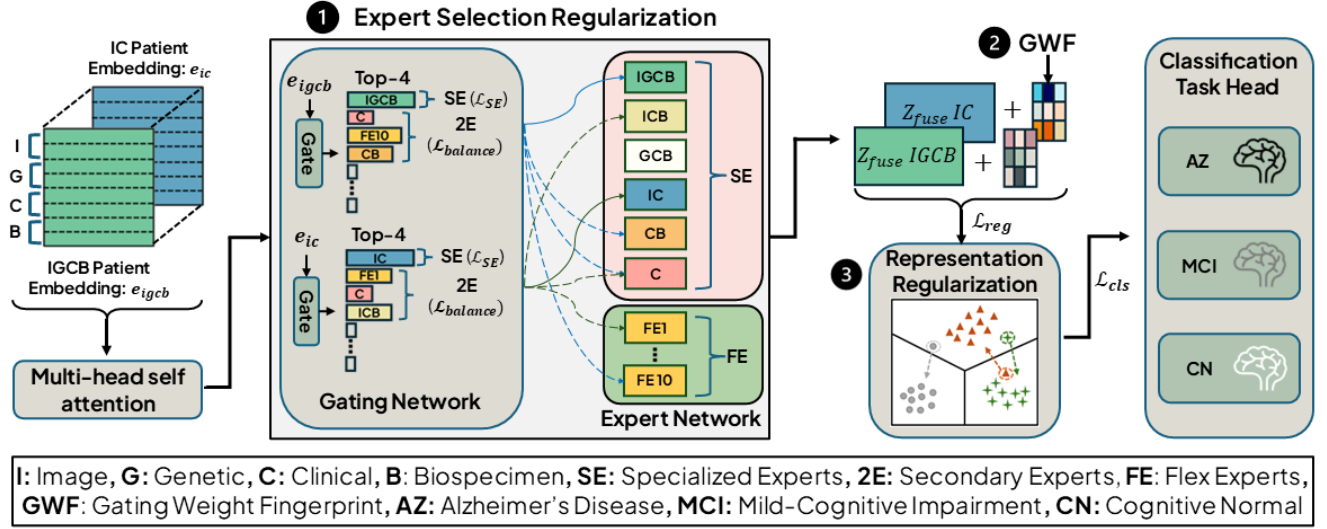


Figure 1. **Overview of the proposed R²MoE framework.** Multimodal tokens are routed through modality-combination Specialized Experts (SE) and Flex Experts (FE). The resulting top-k routing scores form a Gating Weight Fingerprint (GWF), which is fused with each patient's fused embedding (computed as a weighted combination of selected expert outputs) for final disease classification

3. Methods

3.1. Overview

We design our R²MoE framework based on the characteristics of clinical multimodal data—most patients contain only a subset of available modalities. For example, in the ADNI database, some patients have only Imaging and Clinical (IC) information, while only a small portion possess all modalities (IGCB): Imaging, Genomics, Clinical and Biospecimens. Our hypothesis is that it is essential to train a set of *Specialized Experts (SEs)* as shown in Fig. 1 ① to handle patients with specific modality combinations, as each combination carries unique statistical and biological characteristics that should be modeled distinctly. At the same time, we believe that cooperation among experts is also necessary. For instance, some experts focus on patients with only the IC combination, while others handle both IC and IGB combinations. These shared experts are referred to as *Flex Experts (FEs)* as shown in ①. FEs capture common structures shared across different modality combinations, allow gradients to be shared among combinations to prevent SEs from over-specialization or collapse, and provide a flexible fallback mechanism that stabilizes gating when rare modality combinations occur. To ensure that, we utilize Expert Selection Regularization, where SEs and FEs are formed through two regularization objectives: $\mathcal{L}_{SE}$ and $\mathcal{L}_{balance}$.

Through the MoE layer, we obtain the learned patient latent representation $\mathbf{z}_{fuse}$. At the same time, we retain each patient's activation weights across different experts, we call it Gating Weight Fingerprint (GWF), shown in ②. Our key hypothesis is that $\mathbf{z}_{fuse}$ and GWF capture complementary perspectives: the latent embedding $\mathbf{z}_{fuse}$ reflects what the

patient is—the semantic content and multimodal features of the patient; while the GWF reflect how the model processes the patient and essentially constructs a cross-layer, cross-expert meta-feature that characterizes how the model computes for a given patient at different levels of abstraction.

Finally, we use Representation Regularization to apply the triplet ($\mathcal{L}_{trip}$) and contrastive loss ($\mathcal{L}_{con}$) on the joint latent space of $\mathbf{z}_{fuse} + GWF$ to regularize a discriminative and geometrically coherent structure among patient embeddings, as shown in ③. The resulting geometric constraints backpropagate through MoE layers, progressively enhancing the model's representational and predictive capacity.

3.2. Preliminaries

We consider a multimodal medical dataset with N patients, each presenting an arbitrary subset of modalities $\mathcal{M}$. For each patient i and modality $m \in \mathcal{M}$, the raw input $\mathbf{x}_{i,m}$ is accompanied by a binary mask $o_{i,m} \in \{0, 1\}$ indicating availability, reflecting the irregular acquisition patterns common in clinical practice. Observed modalities are processed by modality-specific encoders ϕ_m , producing token sequences $\mathbf{E}_{i,m} = \phi_m(\mathbf{x}_{i,m}) \in \mathbb{R}^{T \times d}$, where T is the number of tokens, and d is the hidden dimension.

Missing Modality Handling: The missing modalities are replaced using a learned mask-conditioned embeddings $\mathbf{E}_m^{mask}$. $\mathbf{E}_m^{mask}$ are learnable based on full-modality sample embeddings, which evolve during training. When an input modality is missing (e.g., I), the remaining modalities pass through the modality-specific encoders. We then apply scale-based matching to compare the available embeddings (G, B, C) with full-modality cases. The $\mathbf{E}_m^{mask}$ for the missing modality (I) is then computed based on the best match-

ing full-modality embeddings. This ensures that the encoders still only train on the observed samples, and are not impacted by lower-quality imputed samples, while keeping the final merged token size same. All available tokens for each patient i are merged into:

$$\mathcal{T}_i = \{\mathbf{E}_{i,m} \cdot o_{i,m} + \mathbf{E}_m^{\text{mask}} \cdot (1 - o_{i,m}) \mid m \in \mathcal{M}\},$$

which are passed through a Transformer block augmented with a sparse MoE layer. Given a specific token $\mathbf{t}_i$ from the available token set $\mathcal{T}_i$, the router g outputs logits $\mathbf{s}_i = g(\mathbf{t}_i) \in \mathbb{R}^E$ and sparse weights:

$$\pi_i = \text{TopK}(\text{softmax}(\mathbf{s}_i), k), \quad k = 4.$$

Each expert f_e receives the same token and produces an expert-specific output $\mathbf{u}_{i,e} = f_e(\mathbf{t}_i)$. The final fused representation is then a weighted sum:

$$\mathbf{z}_{\text{fuse},i} = \sum_{e=1}^E \pi_{i,e} \mathbf{u}_{i,e},$$

which is then used by a final classifier to predict $\hat{y}_i$.

3.3. R²MoE Framework

3.3.1. Expert Selection Regularization

Based on our observations that patients in multimodal clinical datasets frequently present with different subsets of modalities, we introduce *Specialized Experts (SEs)* to learn patient data with specific modality combinations. This design effectively decomposes the overall multimodal distribution into more homogeneous subspaces and is to approximate combination-specific mappings:

$$f_n^{\text{SE}}(X) \approx f^*(X, C_n) \approx \mathbb{E}[Y \mid X, C = C_n],$$

where C_n denotes the n -th modality combination (e.g., $\{I, C\}$ or $\{I, G, B\}$). $f^*(X, C_n)$ refers to the optimal prediction given input X , $\mathbb{E}[Y \mid X, C = C_n]$ refers to the conditional expectation of the clinical label Y given the input X and the event that the patient belongs to modality combination C_n .

To ensure that SEs consistently assume their intended roles, we regularize the MoE routing so that the first selected expert always corresponds to the SE matching the patient's observed modality subset. Let $ob^*(i)$ denote the index of the SE associated with the modality combination of patient i . This anchoring is imposed through a cross-entropy (CE) loss between the router logits $\mathbf{s}_i$ and the target SE index:

$$\mathcal{L}_{\text{SE}} = \text{CE}(\mathbf{s}_i, ob^*(i)).$$

The remaining experts, which are not tied to any specific modality subset, are referred to as *Flex Experts (FEs)*. Unlike SEs that specialize in a single modality combination,

FEs are trained without regularization and their usage is determined entirely by the MoE gating mechanism, allowing them to automatically capture structure that is shared across multiple modality combinations. In other words, FEs learn patterns that are invariant or smoothly varying over the space of modality subsets C . Formally, an FE f_ℓ^{FE} can be interpreted as approximating the combination-averaged behavior of the underlying target function:

$$f_\ell^{\text{FE}}(X) \approx \mathbb{E}_C[f^*(X, C)] = \sum_c p(c \mid X) f^*(X, c),$$

where $f_\ell^{\text{FE}}(X)$ denotes the mapping learned by the ℓ -th Flex Expert when processing input X . $\mathbb{E}_C[f^*(X, C)]$ represents the expected behavior of the ideal target function f^* averaged over all possible modality combinations C . $p(c \mid X)$ is the probability that a patient with input X belongs to modality combination c ; this reflects real-world acquisition patterns such as the likelihood of missing or available modalities. $\sum_c p(c \mid X) f^*(X, c)$ expresses the weighted average of the optimal functions across all modality combinations, capturing cross-combination structure.

Every gate routing selects the Top- k experts, including the Top-1 regularized Specialized Expert $f_{ob^*}^{\text{SE}}(X_i)$ and the $(k - 1)$ secondary experts, which may include either additional SEs or FEs. The final fused representation is computed as a weighted combination of the outputs from these k experts:

$$\mathbf{z}_{\text{fuse},i} = \pi_{i,ob^*} f_{ob^*}^{\text{SE}}(X_i) + \sum_{j=1}^{k-1} \pi_{i,j} f_j(X_i).$$

Balance the $(k - 1)$ secondary experts utilization: After the $f_{ob^*}^{\text{SE}}$ expert is selected, the remaining $(k - 1)$ secondary experts must have balanced utilization to ensure that all secondary experts receive adequate gradient updates, capture complementary cross-combination structure, and maintain stable routing behavior under heterogeneous and incomplete clinical modalities. For each expert f^e , we compute its *importance* and *load* when it is selected as secondary experts as:

$$\text{IMP}_{f^e} = \sum_i \pi_{i,e}. \quad \text{LOAD}_e = \sum_i 1(\pi_{i,e} > 0),$$

We then penalize imbalance:

$$\mathcal{L}_{\text{balance}} = \left(\frac{\sigma(\text{IMP})}{\mu(\text{IMP})} \right)^2 + \left(\frac{\sigma(\text{LOAD})}{\mu(\text{LOAD})} \right)^2,$$

where **IMP** and **LOAD** are the vectors of the all experts importance and load respectively, σ is the standard deviation, and μ is the mean.

3.3.2. Gating Weight Fingerprint

A central hypothesis in our design is that the $\mathbf{z}_{\text{fuse}}$ and the routing weights encode complementary aspects of patient

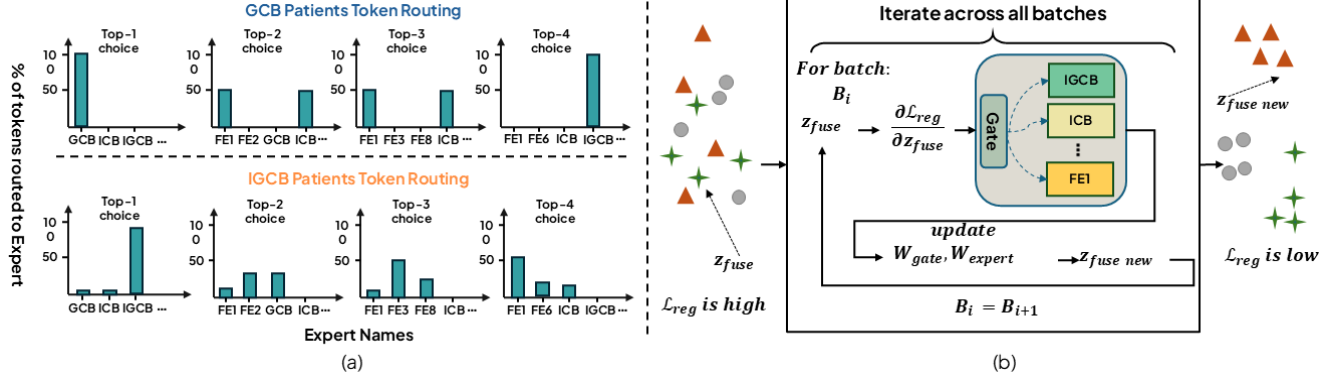


Figure 2. (a) **Top-k expert-selection for GCB and IGCB patients in the last MoE Layer:** secondary expert selection have unique patterns for each patient type. (b) **Representation Regularization:** Supervised contrastive and triplet losses encourage globally cohesive and locally separated class clusters in z_{fuse} . Gradients backpropagate through the MoE layer, stabilizing routing to align with the clustered geometry.

representation. While z_{fuse} reflects the semantic information derived from each patient’s multimodal data, the routing weights describe how the model processes the patient—i.e., the computational pathway taken through the expert network. Aggregating these routing decisions across layers produces a cross-layer, cross-expert meta-feature, a “computation fingerprint” that reveals structural patterns, such as subtype-specific routing profiles, not captured by z_{fuse} alone, making the two representations jointly more discriminative and informative. Fig. 2a illustrates this behavior for two modality combinations. As shown for a GCB patient, the top-1 route is consistently assigned to the designated SE due to $\mathcal{L}_{SE}$.

What we want to emphasize is that although the remaining $(k-1)$ secondary expert selections are encouraged by $\mathcal{L}_{balance}$ to use experts evenly at the global level, they never exhibit uniform expert usage for patients of the same type. Instead, for such patients, the selections consistently concentrate on a small and stable subset of experts. This behavior becomes even more pronounced in the deeper MoE layers, where routing converges to highly consistent secondary patterns and reveals fine-grained computational structure.

Patients sharing the same modality combination (e.g., GCB) have tokens drawn from similar distributions $t_i \sim \mathcal{P}_{GCB}$. Consequently, the corresponding gating logits $s_i = g(t_i)$ also follow an aligned distribution, $s_i \sim g(\mathcal{P}_{IGCB})$. As a result, the secondary routing distribution $\text{softmax}(s_i)$ consistently places higher probability on the same small subset of experts across patients, rather than spreading mass over the entire expert pool. For example in Fig. 2a, the top-2 choice for GCB patients is distributed between only the FE1 and ICB experts. This pattern reflects a stable “expert coalition”—a cooperative group of SEs and FEs that naturally forms through gradient descent. This concentrated routing pattern indicates that the MoE spontaneously discovers a preferred computational pathway for this modality-

combination, revealing an emergent and highly reproducible structure that underlies the computation fingerprint.

This fingerprint also enhances downstream classification. Routing decisions are quasi-discrete, amplifying differences between patient groups and providing a more discriminative signal than z_{fuse} alone. Routing captures information unavailable to the fused embedding, such as which experts handle imaging-dominant cases or which coalitions respond to specific multimodal configurations. Because routing integrates information across layers, it encodes hierarchical computational patterns that often correspond to clinically meaningful phenotype variations. When two subtypes consistently map to different coalitions, the classifier separates them more effectively—even when embeddings overlap or data are noisy. Thus, the computation fingerprint exposes subtype-specific processing pathways that significantly improve robustness and discriminative power.

Based on the above idea, we extract all secondary routing weights from the final two MoE layers—i.e., the $(k-1)$ non-SE expert activations—and concatenate them to form a compact descriptor of the patient’s layer-wise expert engagement, we define this as the Gating Weight Fingerprint (GWF). As shown on the right side of Fig. 1, this GWF is appended to the fused embedding z_{fuse} to construct the final latent representation used for downstream classification.

3.3.3. Representation Regularization

Although the SE-FE expert structure helps capture modality-specific and shared patterns, these components alone do not guarantee that the fused embedding z_{fuse} is semantically well-organized. Without explicit representation regularization, MoE tends to produce geometrically unstable latent spaces, where samples of the same disease class may drift apart and samples of different classes may overlap. When patient embeddings are arranged with clear intra-class compactness and inter-class separation, the experts re-

ceive cleaner and coherent training signals, which promotes meaningful specialization rather than redundant behavior.

Therefore we introduces representation regularization using supervised contrastive and triplet objectives, which impose class-consistent structure on $\mathbf{z}_{\text{fuse}}$ and yield clearer supervision signals for the MoE. Fig. 2b illustrates how these objectives reshape the latent space and produce gradients that reinforce stable routing behavior.

Contrastive supervision: We use a prototype-based contrastive objective that pulls each embedding toward its class centroid:

$$\mathcal{L}_{\text{con}} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \boldsymbol{\mu}_{y_i})/\tau)}{\sum_{c=1}^C \exp(\text{sim}(\mathbf{z}_i, \boldsymbol{\mu}_c)/\tau)},$$

where $\mathbf{z}_i$ is the fused embedding, $\boldsymbol{\mu}_c$ is the centroid of the negative class c , sim is the cosine similarity, and $\tau > 0$ is a temperature coefficient controlling distribution sharpness. This loss enforces global semantic structure, pulling same class embeddings together and pushing different classes apart—ensuring that class identity remains consistent.

Triplet-margin refinement: To sharpen local decision boundaries, we incorporate a semi-hard triplet loss:

$$\mathcal{L}_{\text{trip}} = \frac{1}{|\mathcal{T}|} \sum_{(i,p,n) \in \mathcal{T}} [\|\mathbf{z}_i - \mathbf{z}_p\|_2 - \|\mathbf{z}_i - \mathbf{z}_n\|_2 + \alpha]_+,$$

where (i, p, n) are *semi-hard* triplets—i.e., negatives ($\mathbf{z}_n$) that are farther than positives ($\mathbf{z}_p$) but lie within a margin window—and $\alpha > 0$ is the enforced margin. Triplets refine the fine-scale geometry within and across classes, promoting clear separation even when modality sets differ. Semi-hard sampling avoids degenerate triplets while providing informative gradients that guide boundary refinement in latent space. This approach yields smoother boundaries and more stable clusters, preserving discriminative representations.

Effect on MoE specialization: The contrastive loss defines the *global cluster structure*, while the triplet loss sculpts its *local geometry*, making R²MoE’s fused representation simultaneously cohesive and separable across clinical classes. As each expert receives supervision from samples routed to it, the representation-regularized structure forces experts to specialize on distinct regions of the latent space, reducing redundancy and sharpening their functional roles. A more structured and stable latent space ultimately strengthens the downstream classification signal and improves overall predictive performance. We combine the two objectives into a unified regularizer:

$$\mathcal{L}_{\text{reg}} = \mathcal{L}_{\text{con}} + \lambda_{\text{trip}} \mathcal{L}_{\text{trip}}.$$

Unlike prior work where clustering is used only as a post-hoc diagnostic tool, R²MoE integrates representation-level regularization directly into MoE training, ensuring clinically meaningful and interpretable latent geometry.

3.4. Training Objective and Schedule

The full training objective is a weighted sum:

$$\mathcal{L} = \lambda_{\text{cls}} \mathcal{L}_{\text{cls}} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}} + \lambda_{\text{balance}} \mathcal{L}_{\text{balance}} + \lambda_{\text{SE}} \mathcal{L}_{\text{SE}},$$

where $\mathcal{L}_{\text{cls}}$ is the overall cross-entropy prediction loss for the classification task head, and λ is the weighted combination of each loss. To stabilize training, we adopt a two-phase schedule. In the warm-up phase, the model is optimized with only $\mathcal{L}_{\text{cls}}$, with only full-modality samples being used, and their tokens sent to all experts without routing. This initialization prevents early collapse and gives each expert a shared baseline before the router becomes active. Once stable after warm-up, the remaining loss functions are also activated for correct routing decisions. This staged optimization prevents instability caused by premature enforcement of clustering or expert selection.

4. Experiments

4.1. Experimental Settings

Datasets: We evaluate R²MoE on two large-scale multimodal medical datasets: the Alzheimer’s Disease Neuroimaging Initiative (ADNI) [37, 38] and the MIMIC-IV critical care database [15]. **ADNI** is a longitudinal Alzheimer’s disease cohort, comprising functional MRI, genetic data, clinical assessments, and biospecimen biomarkers—with each modality capturing distinct biological and cognitive axes. We use the standard preprocessed version adopted in prior multimodal studies and focus on stage prediction across cognitively normal, mild cognitive impairment, and Alzheimer’s disease subjects. ADNI highlights the difficulty of fusing high-dimensional imaging with lower-dimensional but clinically informative modalities such as genetics and cognitive scores. **MIMIC-IV** contains over 40,000 ICU patients with heterogenous medical records. We consider three modalities: structured laboratory/vitals (L), ICD-9 diagnosis codes (C), and free-text clinical notes (N). For each datapoint, we extract these modalities, and perform one-year mortality prediction. To match the non-temporal ADNI setup, we only use the last visit for each patient.

Comparison with State-of-the-art methods: We compare R²MoE to a wide range of multimodal fusion methods. (1) *Early fusion models* include ShaSpec [35], mmFormer [45], Tensor Fusion (TF) [44], and MulT [34], which concatenate or attend across features without explicit handling of missingness. (2) *Adaptive fusion methods* include MAG [28] and LIMoE [25], which assign modality weights or rely on modular expert routing. (3) *MoE-based methods* include FuseMoE [10] and FlexMoE [43], which are designed to accommodate arbitrary modality subsets via sparse expert routing. This diverse method set allows us to evaluate not

Table 1. ACC Metric performance on ADNI Dataset, across multiple models and different modality combinations, given the Image (G), Biospecimen (B), Clinical (C), and Genetic (G) modalities. $\mathcal{MC}$ represents the modality combination.

$\mathcal{MC}$	$\mathcal{I}$	$\mathcal{G}$	$\mathcal{C}$	$\mathcal{B}$	ShaSpec	mmFormer	TF	MuT	MAG	LIMoE	[42]	[40]	FuseMoE	Flex-MoE	R ² MoE
<i>Two modalities</i>															
$\mathcal{I}, \mathcal{G}$	•	•			48.09 ± 0.66	49.85 ± 4.92	59.94 ± 0.40	60.32 ± 0.95	59.94 ± 1.00	59.29 ± 0.95	53.74 ± 0.42	53.56 ± 0.83	60.41 ± 0.87	61.08 ± 0.78	63.14 ± 0.49
$\mathcal{I}, \mathcal{C}$	•		•		47.62 ± 1.81	51.96 ± 4.23	54.53 ± 0.66	50.14 ± 1.05	52.19 ± 2.90	52.38 ± 3.46	57.23 ± 1.12	59.61 ± 0.94	53.13 ± 1.97	56.49 ± 2.55	61.18 ± 0.31
$\mathcal{I}, \mathcal{B}$	•			•	50.98 ± 2.09	51.45 ± 3.53	52.57 ± 2.06	51.17 ± 2.88	52.47 ± 1.11	53.87 ± 2.75	50.82 ± 1.94	51.29 ± 2.11	49.67 ± 1.97	60.41 ± 0.26	63.08 ± 1.02
$\mathcal{G}, \mathcal{C}$		•	•		52.85 ± 2.65	49.58 ± 4.45	38.38 ± 3.03	46.03 ± 5.42	40.34 ± 6.11	35.76 ± 6.24	48.99 ± 1.24	47.56 ± 1.16	38.84 ± 2.42	60.60 ± 0.26	63.23 ± 0.89
$\mathcal{G}, \mathcal{B}$		•		•	58.54 ± 2.97	48.45 ± 4.56	42.20 ± 1.78	39.40 ± 2.91	40.52 ± 2.52	36.88 ± 5.04	45.56 ± 0.79	43.50 ± 1.02	37.91 ± 0.80	63.59 ± 1.04	65.02 ± 0.50
$\mathcal{C}, \mathcal{B}$			•	•	59.10 ± 2.69	47.71 ± 4.49	39.68 ± 2.38	44.54 ± 0.82	40.15 ± 2.58	43.98 ± 0.00	48.92 ± 1.34	46.51 ± 0.52	37.91 ± 0.80	60.50 ± 0.82	63.13 ± 0.31
<i>Three modalities</i>															
$\mathcal{I}, \mathcal{G}, \mathcal{C}$	•	•	•		49.30 ± 3.17	46.49 ± 3.57	54.06 ± 1.98	60.97 ± 0.95	61.34 ± 0.61	51.50 ± 2.25	53.76 ± 0.90	58.29 ± 2.11	60.97 ± 1.32	63.21 ± 1.73	66.28 ± 1.21
$\mathcal{I}, \mathcal{G}, \mathcal{B}$	•	•		•	52.85 ± 0.53	47.15 ± 6.43	54.44 ± 2.26	53.03 ± 1.95	54.15 ± 1.06	53.97 ± 1.08	54.91 ± 1.22	55.77 ± 1.63	52.85 ± 1.00	62.28 ± 2.75	65.97 ± 0.72
$\mathcal{I}, \mathcal{C}, \mathcal{B}$	•		•	•	52.85 ± 3.30	47.18 ± 4.68	52.29 ± 1.47	49.86 ± 1.50	53.24 ± 0.50	54.97 ± 0.90	55.11 ± 2.02	57.21 ± 1.62	49.67 ± 1.00	64.05 ± 1.78	67.10 ± 1.12
$\mathcal{G}, \mathcal{C}, \mathcal{B}$		•	•	•	62.09 ± 3.27	46.38 ± 4.24	43.33 ± 4.43	43.32 ± 6.74	37.25 ± 1.99	48.99 ± 2.62	53.74 ± 1.98	54.05 ± 1.00	34.64 ± 1.95	65.36 ± 1.38	68.03 ± 1.67
<i>Four modalities</i>															
$\mathcal{I}, \mathcal{G}, \mathcal{C}, \mathcal{B}$	•	•	•	•	52.84 ± 0.53	58.92 ± 6.58	57.24 ± 3.05	58.82 ± 0.82	61.44 ± 1.61	55.18 ± 4.22	58.52 ± 1.30	57.09 ± 1.77	59.52 ± 1.00	66.11 ± 1.14	70.14 ± 0.38

Table 2. Performance on MIMIC-IV dataset with ACC metric across different models and modality combinations, given the Lab and Vital values ($\mathcal{L}$), Clinical Notes ($\mathcal{N}$), and ICD-9 Codes ($\mathcal{C}$) modalities. $\mathcal{MC}$ denotes observed modality combination.

$\mathcal{MC}$	$\mathcal{L}$	$\mathcal{N}$	$\mathcal{C}$	TF	MuT	MAG	LIMoE	[42]	[40]	FuseMoE	Flex-MoE	R ² MoE
$\mathcal{L}, \mathcal{N}$	•	•		60.05 ± 1.96	57.96 ± 7.25	62.72 ± 2.36	63.80 ± 1.99	60.22 ± 1.07	61.14 ± 0.54	60.50 ± 3.82	76.14 ± 0.73	78.84 ± 1.06
$\mathcal{L}, \mathcal{C}$	•		•	64.13 ± 3.39	62.47 ± 2.01	60.13 ± 1.97	64.89 ± 1.46	61.29 ± 1.21	61.45 ± 0.67	63.31 ± 3.21	75.15 ± 0.55	77.90 ± 0.71
$\mathcal{N}, \mathcal{C}$		•	•	60.97 ± 2.36	62.23 ± 2.81	59.41 ± 4.15	64.27 ± 4.05	58.10 ± 0.97	59.89 ± 0.31	64.77 ± 3.05	74.96 ± 1.59	77.06 ± 0.42
$\mathcal{L}, \mathcal{N}, \mathcal{C}$	•	•	•	63.11 ± 2.17	64.62 ± 0.44	62.87 ± 2.50	61.61 ± 2.37	61.52 ± 0.90	60.01 ± 0.76	63.90 ± 1.72	76.81 ± 0.90	79.40 ± 0.86

only R²MoE’s classification performance but also the added value of dual-regularization, which no prior method integrates end-to-end.

Training setup: For all considered methods, we either adopted the best hyperparameters reported in their original papers or performed grid searches when such details were unavailable. Learning rates were tuned over $\{10^{-3}, 10^{-4}, 10^{-5}\}$, hidden dimensions over $\{64, 128, 256\}$, and batch sizes over $\{8, 16, 32\}$. For R²MoE, we additionally searched over the number of experts $\{8, 16, 32\}$ and Top- k routing choices $\{1, 2, 3, 4\}$. Loss coefficients were tuned on the validation set: the representation losses λ_{reg} were varied in between $\{0.05 - 0.5\}$. All models were trained using the AdamW optimizer with cosine learning rate decay and weight decay of 10^{-3} . We set the embedding dimension to 1024 for R²MoE, with $E = 16$ experts (6 specialized and 10 flex experts) unless otherwise specified. Top- k routing was set to 4 in most experiments. Each model was trained for a maximum of 50 epochs with early stopping based on validation accuracy. Each experiment was repeated three times with different random seeds to account for variability, and we report the mean and standard deviation across runs.

4.2. Primary Results

Classification: Tables I and II show our accuracy results for ADNI and MIMIC-IV datasets to compare R²MoE with other models. As shown, R²MoE consistently outperforms all other methods across both datasets. On ADNI, it achieves higher accuracy for every modality configura-

tion—for example, 63.08% vs 60.41% on (I,B)—with the margin increasing as more modalities are included, reaching 66.28% on (I,G,C) and 70.14% with all four modalities. These improvements confirm that dual-regularization to MoE structure enables R²MoE to exploit complementary signals rather than simply concatenating or routing modalities. On MIMIC-IV, R²MoE also shows strong robustness to noisy, irregular data: it reaches 78.84% on (L,C), a 3.42% improvement over FlexMoE, and maintains this advantage with all three modalities (L,N,C). This indicates that R²MoE prevents high-dimensional clinical notes from dominating the fused representation and preserves structured signals such as labs and codes, making it well-suited for heterogeneous ICU populations where missing and noisy data are common.

Latent space structure: Beyond accuracy, we evaluate latent structure using variance ratio and silhouette score for both ADNI and MIMIC-IV datasets, shown in Table III. R²MoE achieves 1.464 variance ratio and 0.357 silhouette score for ADNI dataset, which shows improved performance over all other schemes. These metrics confirm that R²MoE produces tighter, better-separated clusters aligned with disease classes. Visualizations in Fig. 3 corroborate this: embeddings from R²MoE form well-defined, class-specific manifolds, whereas embeddings produced by other methods comparatively remain diffuse with overlapping clusters. This validates our central claim: representation regularization does not merely improve prediction but also yield clinically interpretable latent spaces. For medical applications, where models must support both diagnosis

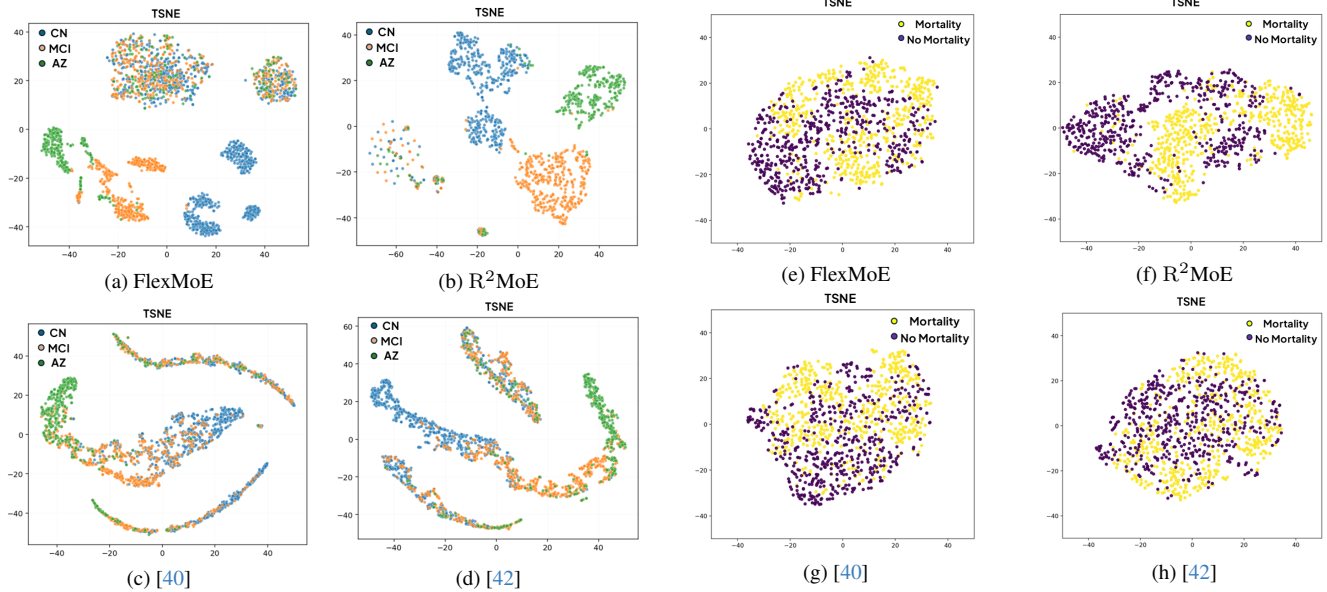


Figure 3. **ADNI and MIMIC latent-space comparisons using t-SNE projections.** Experiments across both ADNI (left) and MIMIC-IV (right) shows that R²MoE consistently achieves the most coherent and well-structured latent space.

Table 3. Comparison of R²MoE with other baselines across Variance Ratio and Silhouette Score for ADNI and MIMIC-IV.

Score	ADNI				MIMIC-IV			
	[42]	[40]	Flex-MoE	R ² MoE	[42]	[40]	Flex-MoE	R ² MoE
Variance Ratio	0.654	0.781	0.736	1.464	0.794	0.886	0.912	1.152
Silhouette	0.087	0.136	0.112	0.357	0.092	0.148	0.152	0.346

Table 4. Ablation study of R²MoE on ADNI (four-modality setting). Mean of accuracy and F1-score over three runs.

Method	Acc (%)	F1 (%)
R ² MoE (full)	70.14	69.31
w/o $\mathcal{L}_{\text{reg}}$	63.23	61.60
w/o $\mathcal{L}_{\text{SE}}$	63.87	62.09
w/o $\mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{se}}$	62.87	61.44
w/o $\mathcal{L}_{\text{balance}}$	65.94	63.56
w/o Gating Weight Fingerprint	69.26	68.52

and discovery, this structural interpretability is as important as raw accuracy.

4.3. Ablation Study

Table 4 summarizes the contribution of each component of R²MoE on ADNI (four-modality setting). The full model achieves the highest accuracy and F1 score. Removing $\mathcal{L}_{\text{reg}}$ yields the largest drop, while eliminating $\mathcal{L}_{\text{SE}}$ also reduces performance (Acc: 63.87%), indicating that supervising expert outputs sharpens specialization. Removing both further degrades performance, confirming that expert and representation regularization act in a complementary manner.

Balanced routing matters as well: removing $\mathcal{L}_{\text{balance}}$ lowers accuracy to 65.94%, showing that preventing secondary-route collapse is essential. Finally, removing the GWF decreases performance (Acc: 69.26%), demonstrating that the routing-level structural signal contributes meaningfully beyond the fused representation. Overall, these ablations validate that R²MoE’s gains arise from the coordination of dual expert and representation regularization, alongside routing-level fingerprints.

5. Conclusion

We introduced R²MoE, a dual-regularized Mixture-of-Experts framework for multimodal clinical prediction under heterogeneous and incomplete modality conditions. R²MoE combines two complementary forms of regularization: (i) *expert-selection regularization*, which anchors each modality-combination to its designated SE and enforces balanced secondary routing, and (ii) *representation regularization*, where contrastive and triplet objectives shape the fused embedding and reinforce consistent expert specialization. In addition, the GWF captures stable, subtype-revealing routing patterns that emerge in deeper MoE layers. Together, these components allow R²MoE to outperform the best existing multimodal models, across ADNI and MIMIC-IV datasets, while producing robust and discriminative fused representations.

Acknowledgment

This work is partially supported by NSF grants CCF-2504152 and CNS-2107057.

References

- [1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text, 2021. [1](#)
- [2] Rawan AlSaad, Alaa Abd-alrazaq, Sabri Boughorbel, Arfan Ahmed, Max-Antoine Renault, Rafat Damseh, and Javaid Sheikh. Multimodal large language models in health care: Applications, challenges, and future outlook. *J Med Internet Res*, 26:e59505, 2024. [1](#)
- [3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy, 2017. [1](#)
- [4] Adrià Casamitjana, Paula Petrone, José Luis Molinuevo, Juan Domingo Gispert, and Verónica Vilaplana. Shared latent structures between imaging features and biomarkers in early stages of alzheimer’s disease: A predictive study. *IEEE Journal of Biomedical and Health Informatics*, 24(2):365–376, 2020. [2](#)
- [5] Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Deep adaptive image clustering. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5880–5888, 2017. [2](#)
- [6] Zitian Chen, Yikang Shen, Mingyu Ding, Zhenfang Chen, Hengshuang Zhao, Erik Learned-Miller, and Chuang Gan. Mod-squad: Designing mixture of experts as modular multi-task learners, 2022. [2](#)
- [7] Can Cui, Zuhayr Asad, William F. Dean, Isabelle T. Smith, Christopher Madden, Shunxing Bao, Bennett A. Landman, Joseph T. Roland, Lori A. Coburn, Keith T. Wilson, Jeffrey P. Zwerner, Shilin Zhao, Lee E. Wheless, and Yuankai Huo. Multi-modal learning with missing data for cancer diagnosis using histopathological and genomic data. In *Medical Imaging 2022: Computer-Aided Diagnosis*, page 120331D. International Society for Optics and Photonics, SPIE, 2022. [2](#)
- [8] Joy Dhar, Nayyar Zaidi, Maryam Haghighat, Puneet Goyal, Sudipta Roy, Azadeh Alavi, and Vikas Kumar. Multimodal fusion learning with dual attention for medical imaging, 2024. [2](#)
- [9] Ching Fang, Christopher Sandino, Behrooz Mahasseni, Juri Minxha, Hadi Pouransari, Erdrin Azemi, Ali Moin, and Ellen Zippi. Promoting cross-modal representations to improve multimodal foundation models for physiological signals, 2024. [2](#)
- [10] Xing Han, Huy Nguyen, Carl Harris, Nhat Ho, and Suchi Saria. Fusemoe: Mixture-of-experts transformers for flexi-modal fusion, 2025. [2](#), [6](#)
- [11] Yan Hao, Chao Cheng, Juanjuan Li, Hongwen Li, Xingsi Di, Xiaoxia Zeng, Shoumei Jin, Xiaodong Han, Chongsong Liu, Qianqian Wang, Bingying Luo, Xianhai Zeng, and Ke Li. Multimodal integration in health care: Development with applications in disease management. *J Med Internet Res*, 27:e76557, 2025. [1](#)
- [12] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. [2](#)
- [13] Jue Jiang, Andreas Rimner, Joseph O. Deasy, and Harini Veeraraghavan. Unpaired cross-modality educed distillation (cmcdl) for medical image segmentation. *IEEE Transactions on Medical Imaging*, 41(5):1057–1068, 2022. [2](#)
- [14] Yang Jiang, Shuang Zhang, and Jianning Chi. Multi-modal brain tumor data completion based on reconstruction consistency loss. *Journal of Digital Imaging*, 36(4):1794–1807, 2023. [2](#)
- [15] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019. [6](#)
- [16] Ibna Kowsar, Shourav B. Rabbani, Kazi Fuad B. Akhter, and Manar D. Samad. Deep clustering of electronic health records tabular data for clinical interpretation. In *2023 IEEE International Conference on Telecommunications and Photonics (ICTP)*, pages 01–05, 2023. [2](#)
- [17] Felix Krones, Umar Marikkar, Guy Parsons, Adam Szmul, and Adam Mahdi. Review of multimodal machine learning approaches in healthcare. *Information Fusion*, 114:102690, 2025. [1](#)
- [18] Garam Lee, Kwangsik Nho, Byungkon Kang, Kyung-Ah Sohn, and Dokyoon Kim. Predicting alzheimer’s disease progression using multi-modal deep learning approach. *Scientific Reports*, 9(1):1952, 2019. [1](#)
- [19] Kwanhyung Lee, Soojeong Lee, Sangchul Hahn, Heejung Hyun, Edward Choi, Byungeun Ahn, and Joohyung Lee. Learning missing modal electronic health records with unified multi-modal data embedding and modality-aware attention, 2023. [2](#)
- [20] Baiying Lei, Siping Chen, Dong Ni, and Tianfu Wang. Discriminative learning for alzheimer’s disease diagnosis via canonical correlation analysis and multimodal fusion. *Frontiers in Aging Neuroscience*, Volume 8 - 2016, 2016. [2](#)
- [21] Yihao Li, Mostafa El Habib Daho, Pierre-Henri Conze, Rachid Zeglache, 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. [2](#)
- [22] Congcong Liu and Fangfang Ye. A review of multimodal medical data fusion techniques for personalized medicine. In *Proceedings of the 4th International Conference on Biomedical and Intelligent Systems*, page 338–347, New York, NY, USA, 2025. Association for Computing Machinery. [2](#)
- [23] Siqi Liu, Sidong Liu, Weidong Cai, Hangyu Che, Sonia Pujol, Ron Kikinis, Dagan Feng, Michael J. Fulham, and ADNI. Multimodal neuroimaging feature learning for multi-class diagnosis of alzheimer’s disease. *IEEE Transactions on Biomedical Engineering*, 62(4):1132–1140, 2015. [1](#)
- [24] Farida Mohsen, Hazrat Ali, Nady El Hajj, and Zubair Shah. Artificial intelligence-based methods for fusion of electronic health records and imaging data. *Scientific Reports*, 12(1):17981, 2022. [1](#)
- [25] Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts, 2022. [6](#)

- [26] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, page 689–696, Madison, WI, USA, 2011. Omnipress. 1
- [27] Jie Peng, Jenna L. Ballard, Mohan Zhang, Sukwon Yun, Jiayi Xin, Qi Long, Yanyong Zhang, and Tianlong Chen. Modalities contribute unequally: Enhancing medical multi-modal learning through adaptive modality token re-balancing. In *Forty-second International Conference on Machine Learning*, 2025. 2
- [28] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2359–2369, Online, 2020. Association for Computational Linguistics. 6
- [29] Wajih Hassan Raza, Aamir Bader Shah, Yu Wen, Yidan Shen, Juan Diego Martinez Lemus, Mya Caryn Schiess, Timothy Michael Ellmore, Renjie Hu, and Xin Fu. Neuro-moe: A transformer-based mixture-of-experts framework for multi-modal neurological disorder classification, 2025. 1
- [30] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts, 2021. 1, 2
- [31] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, 2017. 1, 2
- [32] Kai Sun, Siyan Xue, Fuchun Sun, Haoran Sun, Yu Luo, Ling Wang, Siyuan Wang, Na Guo, Lei Liu, Tian Zhao, Xinzhou Wang, Lei Yang, Shuo Jin, Jun Yan, and Jiahong Dong. Medical multimodal foundation models in clinical diagnosis and treatment: Applications, challenges, and future directions. *Artificial Intelligence in Medicine*, page 103265, 2025. 2
- [33] Aiham Taleb, Matthias Kirchler, Remo Monti, and Christoph Lippert. Contig: Self-supervised multimodal contrastive learning for medical imaging with genetics. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20876–20889, 2022. 2
- [34] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy, 2019. Association for Computational Linguistics. 6
- [35] Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling, 2024. 6
- [36] Jin Wang, Bocheng Guo, Yijie Li, Junyi Wang, Yuqian Chen, Jarrett Rushmore, Nikos Makris, Yogesh Rath, Lauren J O'Donnell, and Fan Zhang. A novel deep learning tractography fiber clustering framework for functionally consistent white matter parcellation using multimodal diffusion mri and functional mri. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2025. 2
- [37] Michael W. Weiner, Paul S. Aisen, Clifford R. Jack Jr., William J. Jagust, John Q. Trojanowski, Leslie Shaw, Andrew J. Saykin, John C. Morris, Nigel Cairns, Laurel A. Beckett, Arthur Toga, Robert Green, Sarah Walter, Holly Soares, Peter Snyder, Eric Siemers, William Potter, Patricia E. Cole, Mark Schmidt, and Alzheimer's Disease Neuroimaging Initiative. The alzheimer's disease neuroimaging initiative: Progress report and future plans. *Alzheimer's & Dementia*, 6(3):202–211.e7, 2010. 6
- [38] Michael W. Weiner, Dallas P. Veitch, Paul S. Aisen, Laurel A. Beckett, Nigel J. Cairns, Robert C. Green, Danielle Harvey, Clifford R. Jack Jr., William Jagust, John C. Morris, Ronald C. Petersen, Jennifer Salazar, Andrew J. Saykin, Leslie M. Shaw, Arthur W. Toga, John Q. Trojanowski, and Alzheimer's Disease Neuroimaging Initiative. The alzheimer's disease neuroimaging initiative 3: Continued innovation for clinical trial improvement. *Alzheimer's & Dementia*, 13(5):561–571, 2017. 6
- [39] Chenwei Wu, Zitao Shuai, Zhengxu Tang, Luning Wang, and Liye Shen. Dynamic modeling of patients, modalities and tasks via multi-modal multi-task mixture of experts. In *The Thirteenth International Conference on Learning Representations*, 2025. 2
- [40] Zhenbang Wu, Anant Dadu, Nicholas Tustison, Brian Avants, Mike Nalls, Jimeng Sun, and Faraz Faghri. Multi-modal patient representation learning with missing modalities and labels. In *The Twelfth International Conference on Learning Representations*, 2024. 7, 8
- [41] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis, 2016. 2
- [42] Yiwen Ye, Yutong Xie, Jianpeng Zhang, Ziyang Chen, Qi Wu, and Yong Xia. Continual self-supervised learning: Towards universal multi-modal medical data representation learning. pages 11114–11124, 2024. 7, 8
- [43] Sukwon Yun, Inyoung Choi, Jie Peng, Yangfan Wu, Jingxuan Bao, Qiyiwen Zhang, Jiayi Xin, Qi Long, and Tianlong Chen. Flex-moe: Modeling arbitrary modality combination via the flexible mixture-of-experts, 2024. 1, 2, 6
- [44] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark, 2017. Association for Computational Linguistics. 6
- [45] Yao Zhang, Nanjun He, Jiawei Yang, Yuexiang Li, Dong Wei, Yawen Huang, Yang Zhang, Zhiqiang He, and Yefeng Zheng. mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation, 2022. 6