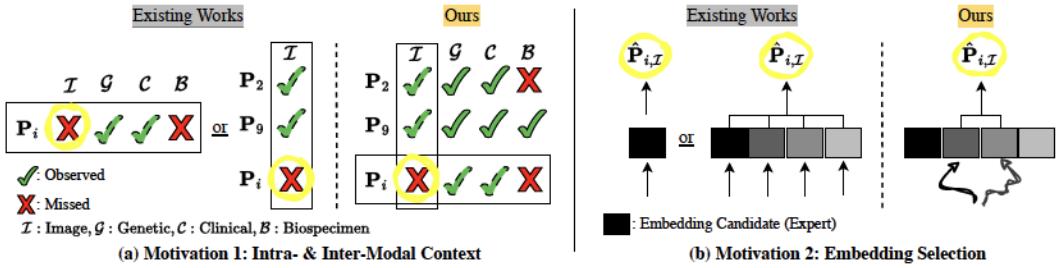
SPARSE MOE AS A NEW RETRIEVER: ADDRESSING MISSING MODALITY PROBLEM IN INCOMPLETE MULTIMODAL DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

In multimodal machine learning, effectively addressing the missing modality scenario is crucial for improving performance in downstream tasks such as in medical contexts where data may be incomplete. Although some attempts have been made to effectively retrieve embeddings for missing modalities, two main bottlenecks remain: the consideration of both intra- and inter-modal context, and the cost of embedding selection, where embeddings often lack modality-specific knowledge.. In response, we propose `MoE-Retriever`[1], a novel framework inspired by the design principles of Sparse Mixture of Experts (SMoE). First,`MoE-Retriever` samples the relevant data from modality combinations, using a so-called supporting group to construct intra-modal inputs while incorporating inter-modal inputs. These inputs are then processed by Multi-Head Attention, after which the SMoE Router automatically selects the most relevant expert, i.e., the embedding candidate to be retrieved. Comprehensive experiments on both medical and general multimodal datasets demonstrate the robustness and generalizability of `MoE-Retriever`, marking a significant step forward in embedding retrieval methods for incomplete multimodal data.

1 INTRODUCTION



(a) Motion 1: Intra- & Inter-Modal Context (b) Motion 2: Embedding Selection

Figure 1: Motivation of this work. (a) **Motivation 1: Intra- & Inter-Modal Context**: Existing works typically consider either the intra-modal context (between samples with the same missing modality, such as $\mathbf{P}_{2,\mathcal{I}}$, $\mathbf{P}_{9,\mathcal{I}}$) or the inter-modal context (within a sample's observed modalities, such as $\mathbf{P}_{i,\mathcal{G}}$, $\mathbf{P}_{i,\mathcal{C}}$). In contrast, our work considers both contexts simultaneously to effectively retrieve the most relevant embedding. (b) **Motivation 2: Embedding Selection**: When retrieving the most relevant embedding ($\hat{\mathbf{P}}_{i,\mathcal{I}}$), existing approaches either use a single static embedding or combine multiple embeddings with simple methods (e.g., sum, average, attention), which makes it difficult to obtain specialized knowledge and requires activation of each embedding candidate every time. In contrast, our work leverages the design principles of SMoE, using a router to automatically select the most relevant experts through top-k selection in a sparse and efficient manner.

In the realm of multimodal machine learning, effectively handling the *missing modality scenario* has become a pivotal challenge for enhancing downstream task performance (Baltrušaitis et al., 2018; Guo et al., 2019; Wu et al., 2024a). In practical cases such as clinical and biological settings, modalities

[1]Source code can be found in the Supplementary Material.

such as imaging, genetic, and clinical data often contain missing entries due to varying acquisition times, costs, or patient-specific factors (Ma et al., 2021; Zhang et al., 2022a;b; Wang et al., 2023). To address this, prior approaches primarily focus on two strategies: imputing missing features directly within the input feature space or employing learnable embedding to represent missing features in the latent space. The former often involves some rule-based prior, such as using the population mean to perform imputation. This method does not scale with data, as the imputation method remain fixed when the underlying distribution changes. In contrast, recent research has increasingly turned toward the latter — leveraging learnable embedding to provide more adaptive and context-aware representations for missing modalities (Zhang et al., 2022b; Wang et al., 2023; Zhang et al., 2022a; Wu et al., 2024b; Han et al., 2024). However, despite their promise, these learnable embedding-based methods still face several critical limitations.

Intra- & Inter-Modal Context. As illustrated in Figure 1 (a), current methods inadequately address both intra-modal and inter-modal contexts when supplementing missing modalities, often focusing on one or the other. In intra-modal scenarios, the goal is to retrieve embeddings for the missing (target) modality by identifying similar samples (Malitesta et al., 2024). However, existing works often choose unimodal approaches that primarily address intra-modal context, failing to personalize the sample's heterogeneous context. Conversely, in inter-modal scenarios, it is assumed that modality-invariant and modality-specific information exists across input modalities, suggesting that missing modalities can be imputed from the sample's specific observed modalities (Zhang et al., 2022b; Wang et al., 2023). However, these works do not carefully consider intra-sample information while proceeding with multi-modal fusion. As a result, focusing solely on either intra-modal or inter-modal context leads to incomplete supplementation and limits the model's ability to effectively leverage the rich multimodal information available in real-world datasets. This highlights the need for a more holistic approach that integrates both perspectives for more accurate and robust imputation of missing modalities.

Embedding Selection. Figure 1 (b) illustrates the current state of embedding retrieval. Current methods either treat the learnable or retrieved embeddings as a single embedding (Wang et al., 2023; Han et al., 2024) or use diverse embeddings but require activating all candidates every time a retrieval is performed, using operations like summation, averaging, or attention mechanisms. These methods can incur a high computational cost as the number of samples or modalities grows, and they lack the ability to adapt to diverse observed modality combinations, treating all potential scenarios equally regardless of the specific context. This uniformity in handling observed modalities limits the capacity for more nuanced and context-specific supplementation. For instance, specific knowledge may be required when certain input modality combinations are present, which is crucial for improving downstream task performance.

Our Approach. To address these challenges, we propose `MoE-Retriever`, a novel framework for embedding retrieval given a incomplete multimodal data. The main idea of `MoE-Retriever` is to borrow the desgin principle from the Sparse Mixture of Experts (SMoE), which activates most relevant experts (i.e., embedding candidates) given a specific intra- and inter-modal context within in a router in a sparse manner. To achieve this, we first begin with genertaing supporting group which is based on given modality combination and aim for target (missing) modality, responsible for sampling intra-modal samples. Next by incoporating inter-modal samples and via Multi-head attention within this incoproated inputs and router with experts which both includes the shared and modality-specific experts finally retreives the most relevant embedding for target modality. Extensive experiments on two medical datasets (ADNI, MIMIC) and two general machine learning datasets (ENRICO, CMU-MOSI) validate the efficacy and generalizability of our proposed method, consistently demonstrating its robust performance across various multimodal settings.

- We highlight that current intra- or inter- modal or single or multiple-but-lacking specialized knwoledge brings the bottleneck into incomplete multimodal embedding retrieval.
- We propose `MoE-Retriever`, borrowing the design principle of Sparse Mixture of Experts design, which inputs the both intra-modal inter-sample and inter-modal intra-sample contexts and retrieve most relevant embedding from modality-specific and shared experts.
- Our comprehensive experimental evaluations on the medicinal dataset and machine learning datasets, showcase the effectiveness and portability of `MoE-Retriever`.

2 Related Work

Multimodal Learning with Missing Modality. Multimodal learning has garnered increasing attention in the machine learning community, particularly in the medical domain, where clinical data is inherently multimodal (Khader et al., 2023; Steyaert et al., 2023). However, in real-world clinical practice, missing modalities are a common challenge (Zhou et al., 2023; Liu et al., 2023). To address this issue, one straightforward approach is to leverage generative models to impute the missing modalities (Pan et al., 2021; Zhang et al., 2024). Nonetheless, generative modeling of another distribution is a ill-posed problem (Zhang et al., 2022a). In contrast, non-generative approaches have emerged, utilizing techniques such as graph-based modeling (Wu et al., 2024b), and modality fusion (Zhang et al., 2022b; Wang et al., 2023; Yao et al., 2024). While these methods can harness both inter-patient and intra-patient information, they face challenges related to scalability and struggle to handle fleximodal scenarios (Han et al., 2024), where any combination of modalities may be present. To improve scalability, FuseMoE (Han et al., 2024) introduced a sparse Mixture-of-Experts (MoE) model aims to be robust to any combination of missing modality scenario. However, despite its scalability advantages, FuseMoE do not explicitly account both the inter-patient and intra-patient relationships simultaneously, limiting its ability to fully utilize the multimodal context of clinical data.

Sparse Mixture-of-Experts (SMoE). SMoE (Shazeer et al., 2017) builds on the traditional Mixture-of-Experts (MoE) model (Jacobs et al., 1991; Jordan & Jacobs, 1994; Chen et al., 1999; Yuksel et al., 2012) by introducing sparsity, which enhances both computational efficiency and model performance. By selectively activating only the most relevant experts for a specific task, SMoE minimizes overhead and improves scalability, making it particularly useful for complex, high-dimensional datasets across various applications. It has been widely applied in both vision (Riquelme et al., 2021; Lou et al., 2021; Ahmed et al., 2016; Wang et al., 2020; Yang et al., 2019; Abbas & Andreopoulos, 2020) and language processing (Lepikhin et al., 2021; Kim et al., 2021; Zhou et al., 2022; Zhang et al., 2021; Zuo et al., 2022; Jiang et al., 2021). Its capacity to dynamically allocate different network parts to specific tasks (Ma et al., 2018; Aoki et al., 2021; Hazimeh et al., 2021; Chen et al., 2023) or data modalities (Kudugunta et al., 2021) has been explored for various applications (Mustafa et al., 2022). Research shows its effectiveness in areas like classification tasks for digital number recognition (Hazimeh et al., 2021) and medical signal processing (Aoki et al., 2021). However, the current use of SMoE is often biased toward its role as a backbone design, typically integrated into Transformer architectures to improve embedding representations in fusion or prediction layers. Its potential for more effective use, such as serving as a retriever or supplementing missing embeddings to bridge the feature space and encoder space, remains underexplored.

3 Method

3.1 Preliminaries and Notations

Motivation behind bringing SMoE design. In the context of incomplete multimodal data, only the observed features in the raw feature space can pass through the modality-specific encoder. This raises a critical question: *how can we effectively handle samples with missing modalities to provide robust embeddings for the missing features?* Ensuring that the embedding space, followed by the fusion and prediction layers, remains trainable through continuous gradient flow is essential. It is important to note that different samples exhibit varying combinations of observed modalities, which necessitates a personalized approach capable of handling each sample's unique environment, such as its specific modality combination.

To address this challenge, we introduce the design principles of SMoE (Shazeer et al., 2017). Given a pool of diverse experts (i.e., trainable feed-forward networks), the SMoE architecture enables the automatic and sparse activation of different experts, each specializing in certain knowledge, based on the input scenario. This dynamic routing mechanism effectively mitigates the limitations of static, one-size-fits-all designs, where learnable embeddings are constrained to a single expert or a fixed combination of embeddings without a router. In such static setups, embeddings for missing modalities are often selected at random, leading to suboptimal performance for downstream tasks.

Notation. Formally, SMoE consists of multiple experts, denoted as $\mathcal{E}_1, \ldots, \mathcal{E}_{|\mathcal{E}|}$, where $|\mathcal{E}|$ represents the total number of experts, and a router, $\mathcal{R}$, which governs the routing mechanism, sparsely selecting

the top-k experts. For a given embedding or token $\mathbf{x}$, the router $\mathcal{R}$ activates the top-k experts based on the highest scores derived from a softmax function applied to the outputs of a learnable gating function, $g(\cdot)$, typically modeled as a one or two-layer MLP. The router's output, $\mathcal{R}(\mathbf{x})_i$, indicates the selection of the i-th expert. This process is formally described as follows:

$$\begin{aligned} \mathbf{y} &= \sum_{i=1}^{|E|} \mathcal{R}(\mathbf{x})_i \cdot \mathcal{E}_i(\mathbf{x}), \\ \mathcal{R}(\mathbf{x}) &= \text{Top-K}(\text{softmax}(g(\mathbf{x})), k), \\ \text{TopK}(\mathbf{v}, k) &= \begin{cases} \mathbf{v}, & \text{if } \mathbf{v} \text{ is in the top } k, \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{1}$$

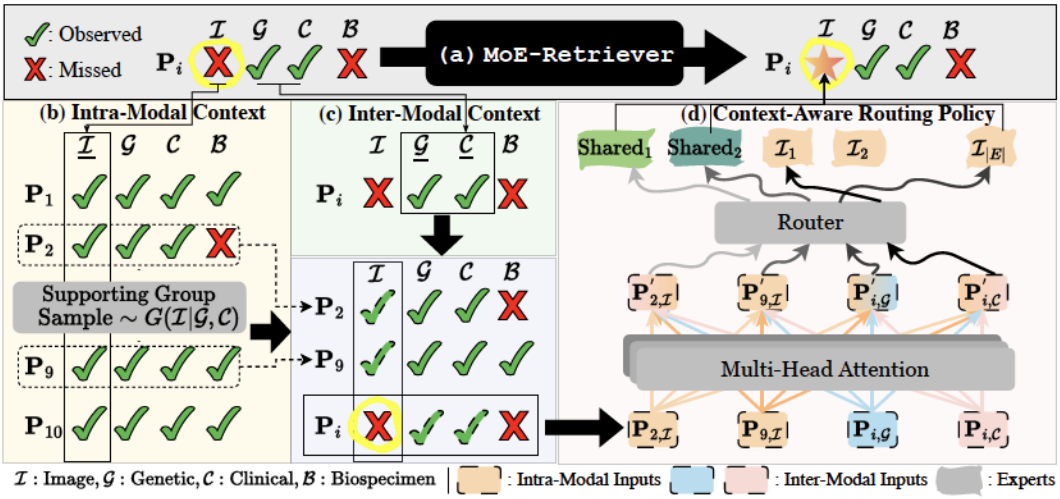


Figure 2: Overall illustration of MoE-Retriever. (a) **The role of MoE-Retriever.** Given a sample ($\mathbf{P}_i$) with a missing modality, $\mathcal{I}$ (Image), our goal is to retrieve the most relevant embedding ($\mathbf{P}_{i,\mathcal{I}}$) by considering two contextual factors. First, we focus on (b) **Intra-Modal Context**, which seeks to find embeddings within the same modality as the missing one ($\mathcal{I}$) to reflect similar contextual knowledge. To achieve this, we define a supporting group ($G(\mathcal{I}|\mathcal{G},\mathcal{C})$), where the target modality ($\mathcal{I}$) and the sample's observed modalities ($\mathcal{G},\mathcal{C}$) form a sufficient context for grouping. After sampling from this group, we incorporate the sample's specific (c) **Inter-Modal Context**, leveraging the observed modalities. We then proceed to (d) **Context-Aware Routing Policy**, which first applies multi-head attention and adopts the SMoE framework. Here, the router (top-1 selection in this example) selects the most relevant experts given two intra- and inter-modal inputs. After integrating all the embeddings, the final embedding is regarded as the retrieved embedding for the sample i's missing modality $\mathcal{I}$, denoted as $\mathbf{P}_{i,\mathcal{I}}$. For retrieving an embedding for another missing modality, $\mathcal{B}$, the supporting group would be updated to $G(\mathcal{B}|\mathcal{G},\mathcal{C})$, and the intra-modal embeddings would consist of $\mathbf{P}_{\cdot,\mathcal{B}}$, with the expert selection adapted accordingly to $\{\mathcal{B}_1,\ldots,\mathcal{B}_E\}$.

3.2 MoE-Retriever

The overall framework of MoE-Retriever, along with the detailed procedure, is illustrated in Figure 2. In essence, the key idea behind MoE-Retriever is to retrieve the most relevant embedding for the missing modality by leveraging two contexts: (1) **Intra-Modal Context**, which samples similar examples from a well-defined supporting group based on the observed modality combination (Sec 3.2.1), and (2) **Inter-Modal Context**, which considers the sample-specific heterogeneous combination of observed modalities (Sec 3.2.2). The next step is (3) **Context-Aware Routing**, where the expert pool is designed modality-specifically, using both contexts to effectively supplement the target (i.e., missing) modality. Finally, the selected experts and their linear combination with the inputs are integrated into a single embedding, which is regarded as the final retrieved embedding (Sec 3.2.3).

3.2.1 INTRA-MODAL CONTEXT

We begin with the *intra-modal context* (column-wise context in Figure 2), where intra-modal refers to the homogeneous modality that matches the target modality we aim to supplement. The rationale for incorporating this context is that, by forming a pool of similar samples, we can capture patterns directly observed across patients, without requiring any additional preprocessing. The observed pattern can be represented as a *modality combination*, which reflects similar trends or patterns, i.e., knowledge observed across the samples. Empirically, samples (e.g., patients) with similar observed modality combinations have shown exhibit analogous characteristics. For instance, patients who lack the image modality but possess both genetic and clinical modalities may be more likely to display correlations with certain domain-specific traits, such as early-stage diagnosis, mild cognitive impairment, or slower progression rates, often associated with genetic risk factors like the APOE $\epsilon 4$ allele (Dubois et al., 2023; Jack Jr et al., 2018; Lambert et al., 2013).

To effectively sample from an intra-modal sample pool, we first need to define a modality combination-specific pool, which we denote as the *supporting group*. The core idea behind the supporting group is that, given an observed modality combination and a target (missing) modality, the corresponding group must include the observed modalities as well as the target modality to support the patient's intra-modal pool. For example, if a sample contains the modalities '$\mathcal{GC}$' and we aim to impute the modality '$\mathcal{I}$' (as illustrated in Figure 2), the supporting group should include samples with 'GC' as well as the missing modality '$\mathcal{I}$'. Consequently, the supporting group would comprise samples with modality combinations such as '$\mathcal{IGC}$' or '$\mathcal{IGCB}$'.

Formally, as an example from Figure 2 , let the set of modalities be $\mathcal{M} = \{\mathcal{I}, \mathcal{G}, \mathcal{C}, \mathcal{B}\}$. With a specific modality combination $mc \in \mathcal{MC} = \{\mathcal{I}, (\mathcal{I}, \mathcal{G}), (\mathcal{I}, \mathcal{G}, \mathcal{C}), \ldots, \mathcal{G}, (\mathcal{G}, \mathcal{C}), \ldots, (\mathcal{I}, \mathcal{G}, \mathcal{C}, \mathcal{B})\}$, where the total number of combinations in $\mathcal{MC}$ is $|\mathcal{MC}| = \sum_{m=1}^{|\mathcal{M}|-1} \binom{|\mathcal{M}|}{m} = 2^{|\mathcal{M}|} - 1$, the supporting group G consists of the samples that satisfy the following constraints:

$$\begin{aligned} G(j \mid \mathcal{T}, mc) &= \{j \in \{1, 2, \ldots, N\} \mid mc_j \in \mathcal{X}(\mathcal{S} \mid \mathcal{T}, mc)\} \\ \text{where } \mathcal{X}(\mathcal{S} \mid \mathcal{T}, mc) &= \{S \subseteq \mathcal{M} \mid (mc \subseteq S) \wedge (\mathcal{T} \in S)\} \ \ \forall \mathcal{T} \in \mathcal{M}, \ \ \forall mc \in \mathcal{MC} \end{aligned} \tag{2}$$

where $G(j \mid \mathcal{T}, mc)$ denotes the set of sample indices among total sample size N, derived from the set of possible modality combinations $\mathcal{X}(\mathcal{S} \mid \mathcal{T}, mc)$ for a given target modality $\mathcal{T}$ and modality combination mc. In this context, the satisfying $\mathcal{S}$ denotes any arbitrary set of modality combinations that satisfies the constraint of including both mc (i.e., $(mc \subseteq S)$) and (i.e., $\wedge$) the target modality $\mathcal{T}$ as subsets (i.e., $(\mathcal{T} \in S)$). Given the supporting group G, we sample[2] intra-modal examples that assist in the final retrieval by SMoE by referring to similar examples within the homogeneous modality.

3.2.2 INTER-MODAL CONTEXT

Beyond intra-modal context, we now consider another critical dimension: inter-modal context (illustrated row-wise in Figure 2). This approach allows us to incorporate personalized context specific to a given sample that would be missed by only considering intra-modal context. As a real-world example, this perspective is particularly meaningful in multimodal medical scenarios such as Alzheimer's diagnosis. When genetic (G) and clinical (C) data are available but imaging (I) is missing (case of Figure 2), it may suggest the patient is in the early stages of the disease, where less invasive and more accessible modalities are prioritized. Imaging, typically more expensive, may be reserved for later stages when symptoms progress (Dubois et al., 2023; Li et al., 2022). Additionally, genetic and clinical data alone can provide valuable early insights, guiding initial interventions before resorting to costly imaging techniques (Kim, 2023).

Formally, to consider inter-modal context, we directly focus on the observed modalities, i.e., mc (e.g., $(\mathcal{G}, \mathcal{C})$) for a sample index, i. By doing so, we integrate these sample-specific heterogeneous modality combinations, which will serve as input for the inter-modal examples in the final retrieval by SMoE, referring to the personalized context within the heterogeneous modalities.

[2]For the number of samples, we used a count that matches the observed modalities of the samples (i.e., $|mc|$) to ensure a balanced impact of both. They may vary and can be treated as a hyperparameter for flexibility. However, empirical observations indicate that varying the number of intra-modal samples has only a marginal effect on model performance.

3.2.3 CONTEXT-AWARE ROUTING POLICY

Now, given two contexts, i.e., intra-modal and inter-modal, we proceed with context-aware routing via the SMoE design. The goal of this routing is to retrieve the most relevant expert given an input combination that includes both homogeneous and heterogeneous modality information. For each embedding (i.e., token) input to the router, the router is trained to select the most relevant expert that can benefit the downstream task. The selected experts are expected to specialize in handling the specific input modalities.

The context-aware router design is detailed as follows:

$$\hat{\mathbf{P}}_{i,\mathcal{T}} = \sum_{e=1}^{|E|} \mathcal{R}(\mathbf{x})_e \cdot \mathcal{E}_e^{\mathcal{T}}(\mathbf{x}) \tag{3}$$
$$\text{where } \mathbf{x} \in \{\mathbf{P}'_{i_{\text{intra}},\mathcal{T}} \cup \mathbf{P}'_{i,mc}\}, \ \ \forall i_{\text{intra}} \in G(\mathcal{T} \mid mc), \forall \mathcal{T} \in \mathcal{M}, \ \ \forall mc \in \mathcal{MC}$$

where $\hat{\mathbf{P}}_{i,\mathcal{T}}$ is the predicted retrieved embedding for sample i's missing modality $\mathcal{T}$. $\mathcal{R}(\cdot)$ denotes the router responsible for top-k expert selection, as defined in Equation 1, given an input embedding or token. Here, the input of SMoE, $\mathbf{x}$ includes (i.e., $\cup$) both intra-modal examples ($\mathbf{P}'_{i_{\text{intra}},\mathcal{T}}$) and inter-modal examples ($\mathbf{P}'_{i,mc}$). $\mathbf{P}' = \text{MHA}(\mathbf{P})$, where $\mathbf{P}$ represents the embedding after passing through the modality-specific encoder from raw feature space. This denotes the embedding or token after undergoing Multi-Head Attention (MHA), i.e., Cross-Attention, enabling interaction between tokens. Thus, tokens are endowed with not only self-modality knowledge but also inter-modal harmonization before being passed to the SMoE router.

For the expert design, $\mathcal{E}_e^{\mathcal{T}}(\mathbf{x})$ represents the modality-specific expert , where each expert corresponds to a distinct FFN layer, is distinct and newly introduced in `MoE-Retriever` to enhance context-awareness, particularly in handling missing modality scenarios. Notably, the retrieval target differs for each modality combination in various samples, leading us to allocate specific expert indices for each target modality. For instance, if there are 32 experts and four modalities, each modality will have its own pool of 8 experts. Additionally, to enhance flexibility and generalizability, we include shared experts (denoted as 'Shared' in Figure 2), expecting that common knowledge can be leveraged across different modalities. The number of shared experts is controlled by the hyperparameter b, and we elaborate on this design in Section 4.4.

After retrieving the most relevant embedding for each missing modality, we proceed to the subsequent fusion layer[3], followed by the prediction head for the downstream task. Since gradients flow continuously from the input features to the output predictions, this enables end-to-end training.

3.3 OVERALL ALGORITHM

To summarize, the overall algorithm of `MoE-Retriever` is detailed in Algorithm 1.

[3]The fusion layer can be based on diverse architectures, such as Transformers or even an SMoE layer. To ensure generalizability, we choose a vanilla Transformer encoder as our fusion layer and explore alternative backbones in the Experiments section.

Algorithm 1 The overall procedure of `MoE-Retriever`.

Input: Samples, $i \leq N$, Supporting Group, $G(\mathcal{T} \mid mc)$, Modality Set, $\mathcal{M}$, Modality Combination Set, mc
Output: Retrieved Embedding for Missing Modality, $\mathcal{T}$
for $i = 1, \cdots, N$ **do**
  if $|mc_i| < |\mathcal{M}|$:
    for $t \in \mathcal{T}_i$ **do**
      $\mathbf{x} = []$
      /* `Intra-Modal Context` */
      Samples $\sim G(t \mid mc_i)$
      for $j \in$ Samples **do**
        x.append($\mathbf{P}_{j,\mathcal{T}}$)
      end for
      /* `Inter-Modal Context` */
      for $mc \in mc_i$ **do**
        x.append($\mathbf{P}_{i,mc}$)
      end for
      /* `Context-Aware Routing Policy` */
      $\mathbf{x} \leftarrow \text{MHA}(\mathbf{x})$
      $\hat{\mathbf{P}}_{i,\mathcal{T}} \leftarrow \text{SMoE}(\mathbf{x}, \mathcal{R}, \mathcal{E}^{\mathcal{T}}, \text{top-k})$
    end for
end for

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Multimodal Medical Datasets. We evaluate `MoE-Retriever` on two real-world multimodal medical datasets. *ADNI Dataset*: The Alzheimer's Disease Neuroimaging Initiative (ADNI) is pivotal for Alzheimer's Disease (AD) research, aggregating multimodal data on disease evolution and biomarkers including four modalities, image modality (MRI and PET scans), genetic profiles, clinical metrics, and biospecimen samples, with open access for research standardization (Weiner et al., 2010; 2017). After preprocessing, we extract 2380 samples and target a three-tier classification task of AD stages: Dementia, Cognitively Normal (CN), or Mild Cognitive Impairment (MCI). *MIMIC Dataset*: MIMIC-IV (Medical Information Mart for Intensive Care IV) is sourced from critical care units, offers both structured (demographics, vitals, labs, medications) and unstructured data (clinical notes). For the experiments, we extracted labs results, clinical notes, and ICD 9 codes from 9,003 patient records to predict a binary classification of one-year mortality prediction (Johnson et al., 2023). For the detailed preprocessing for ADNI and MIMIC dataset, please refer to Appendix A.1.

Additional Multimodal Datasets. To demonstrate the generalizbility of `MoE-Retriever` toward other real-world multimodal domain, we use two general multimodal datasets. *CMU-MOSI Dataset*: The Multimodal Corpus of Sentiment Intensity (CMU-MOSI) dataset comprising 2,199 annotated video clips, advances affect recognition through detailed sentiment analysis on a scale from -3 to +3, utilizing YouTube vlogs for real-world sentiment expression research (Zadeh et al., 2016). *ENRICO Dataset*: The Enhanced Rico (ENRICO) (Leiva et al., 2020) dataset is a collection of 1,460 Android app screens, each comprising an image along with the set of apps and their respective locations. This dataset is organized into 20 distinct design categories, which focuses on a classification tasks to identify different design motifs.

Baselines. We compare `MoE-Retriever` against various state-of-the-art baselines from three categories. (1) **feature modeling methods**: mmFormer (Zhang et al., 2022b) and ShaSpec (Wang et al., 2023)). (2) **graph-based approaches**: MUSE (Wu et al., 2024b) and M3Care (Zhang et al., 2022a). (3) **MoE-based method**: FuseMoE (Han et al., 2024). For details regarding modality-specific encoders setting, please refer to Appendix A.2.

Implementations. To ensure a fair comparison with other baselines, we utilized the optimal hyper-parameter settings provided in the original papers. For dataset split, we choose 70% for training, 15% as validation set, and the remaining 15% for testing. Both the ADNI and MIMIC datasets contain missing data. For the CMU-MOSI and ENRICO datasets, we applied random dropping with probability of 0.3 for each modality independently to simulate missing modality scenarios. Given the

incomplete nature of the datasets, if a baseline implementation could impute or interact with other modalities, we leveraged those methods. Otherwise, we used zero-padding to support batch-wise training. All experiments were conducted on NVIDIA A100 GPUs. Each experiment was run three times with different seeds to ensure reproducibility, and the results were averaged.

4.2 PRIMARY RESULTS

Dataset	Modality	Metric	mmFormer	ShaSpec	M3Care	MUSE	FuseMoE	MoE-Retriever
ADNI	$\mathcal{I}$+$\mathcal{G}$	Acc.	50.42±4.98	54.81±4.47	48.69±4.03	43.90±2.59	60.41±0.87	**61.09**±2.12
		F1	46.66±2.40	54.43±4.11	40.29±6.49	26.83±2.68	61.04±0.95	**62.10**±1.12
	$\mathcal{I}$+$\mathcal{G}$+$\mathcal{C}$	Acc.	51.73±1.40	58.36±1.65	48.97±2.45	45.04±2.65	60.97±1.32	**63.12**±1.19
		F1	49.97±1.89	52.69±4.99	43.55±6.24	37.21±2.61	61.30±1.07	**62.17**±2.90
	$\mathcal{I}$+$\mathcal{G}$+$\mathcal{C}$+$\mathcal{B}$	Acc.	55.46±1.05	59.94±2.25	54.68±0.70	52.24±2.61	59.52±1.00	**64.52**±2.55
		F1	46.94±0.31	59.94±1.88	46.09±2.29	43.07±2.01	59.55±1.60	**63.80**±2.96
MIMIC	$\mathcal{L}$+$\mathcal{N}$	Acc.	77.37±0.00	77.37±0.15	76.14±0.46	**77.40**±1.12	60.50±3.82	76.82±3.02
		F1	43.62±0.00	55.19±1.52	45.26±0.44	51.53±1.90	52.79±1.32	**58.06**±2.19
	$\mathcal{L}$+$\mathcal{C}$	Acc.	77.37±0.00	77.37±0.13	76.76±0.59	**77.40**±1.12	63.31±3.21	77.20±0.47
		F1	43.62±0.00	57.32±0.52	43.92±0.52	51.53±1.90	54.78±0.91	**57.73**±0.64
	$\mathcal{N}$+$\mathcal{C}$	Acc.	77.37±0.00	77.40±0.03	77.26±0.35	77.32±1.13	64.77±0.36	**77.45**±0.14
		F1	43.62±0.00	54.59±0.65	45.31±1.22	51.53±1.90	55.54±0.60	**56.65**±1.23
	$\mathcal{L}$+$\mathcal{N}$+$\mathcal{C}$	Acc.	77.37±0.00	**77.40**±0.09	76.04±0.70	77.40±1.12	63.90±1.72	76.59±0.07
		F1	43.62±0.00	55.79±0.94	45.43±1.17	51.25±1.87	55.38±0.16	**59.74**±0.81

Table 1: Performance comparison in ADNI and MIMIC Datasets. Image ($\mathcal{I}$), Genetic ($\mathcal{G}$), Clinical ($\mathcal{C}$), and Biospecimen ($\mathcal{B}$) modalities are used for ADNI dataset. For ADNI dataset, we use the image modality as a central reference, and sequentially added genetic, clinical, and finally all four modalities. Lab ($\mathcal{L}$), Notes ($\mathcal{N}$), and Code ($\mathcal{C}$) modalities are used in MIMIC dataset. We report Accuracy (Acc.) and F1-Macro (F1) scores.

Dataset	Modality	mmFormer	ShaSpec	M3Care	MUSE	FuseMoE	MoE-Retriever
ENRICO	$\mathcal{S}$+$\mathcal{W}$	36.19±0.98	21.03±0.32	19.06±5.17	36.01±2.81	36.99±6.83	**38.24**±1.16
CMU-MOSI	$\mathcal{V}$+$\mathcal{A}$	42.23±0.00	50.91±1.63	42.23±0.00	44.64±1.94	47.46±2.36	**53.12**±2.26
	$\mathcal{V}$+$\mathcal{T}$	62.20±0.90	60.01±1.44	42.12±0.14	52.54±1.92	63.77±1.62	**65.74**±0.55
	$\mathcal{A}$+$\mathcal{T}$	65.65±0.63	65.09±1.02	47.05±6.83	50.82±1.91	61.33±0.93	**66.13**±0.69
	$\mathcal{V}$+$\mathcal{A}$+$\mathcal{T}$	62.75±1.12	64.02±0.65	42.23±0.00	50.66±1.93	60.67±0.22	**65.21**±2.72

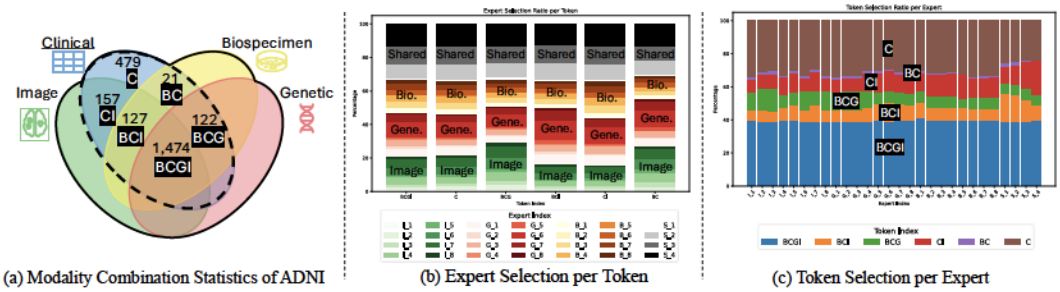
Table 2: Performance comparison in ENRICO and CMU-MOSI Datasets. Screenshot ($\mathcal{S}$), and Wireframe ($\mathcal{W}$) modalities are used for ENRICO dataset. Vision ($\mathcal{V}$), Audio ($\mathcal{A}$), and Text ($\mathcal{T}$) modalities are used in CMU-MOSI dataset. We report Accuracy (Acc.) for both datasets.

Results on ADNI and MIMIC Datasets. Table 1 presents several insights: **1)** On the ADNI dataset, among all modality combinations, `MoE-Retriever` outperforms all baselines by a notable margin. **2)** Notably, as the number of available modalities increases (e.g., $\mathcal{I}+\mathcal{G}+\mathcal{C}+\mathcal{B}$), the potential of `MoE-Retriever` grows, providing a large margin of improvement (7.64% gain compared to the best-performing model, ShaSpec, and 8.40% gain compared to the state-of-the-art model, FuseMoE). This shows that with more modalities, there is greater room for improvement, which can be attributed to the fact that a larger number of intra- and inter-modal samples facilitate the retrieval process. **3)** The two graph-based methods, M3Care (Zhang et al., 2022a) and MUSE (Wu et al., 2024b), perform the worst on the ADNI dataset. This suggests that while graph-based approaches capture intra-modal relationships between samples, they struggle due to the lack of handling inter-modal interactions, highlighting the importance of these interactions. **4)** FuseMoE (Han et al., 2024), a mixture-of-experts (MoE)-based method, achieves the best performance on the ADNI dataset but significantly underperforms on the MIMIC dataset [4]. This can be attributed to FuseMoE's reliance on a single random embedding to impute missing modalities. **5)** On the MIMIC dataset,

[4]We attempted to use the authors' code but observed unstable performance. Thus, we borrowed FuseMoE's performance on these datasets from the recent Flex-MoE paper (Yun et al., 2024).

all baseline models suffer from the label imbalance problem, resulting in either Acc or F1 scores being biased. However, `MoE-Retriever` appears to be a well-balanced model, where the F1 score, being more significant than Acc in imbalanced cases, consistently outperforms all baselines. All in all, `MoE-Retriever` achieves notable performance gains on both datasets, thanks to its ability to model intra- and inter-modal contexts and its context-aware routing policy via the SMoE design, showcasing that better-retrieved embeddings for missing modalities lead to downstream performance improvements.

Results on ENRICO and CMU-MOSI Datasets. Table 2 shows the performance across generalized domains: design motifs for the ENRICO dataset and sentiment analysis for the CMU-MOSI dataset. We observe that **1)** `MoE-Retriever` outperforms current multimodal baselines, demonstrating its generalizability across diverse multimodal domains. Specifically, in the CMU-MOSI dataset, we observe **2)** that as the number of modalities increases, the performance of existing baselines improves, but the increase does not surpass that of `MoE-Retriever`, highlighting its effectiveness as a strong benchmark model for various domains and modality combinations.



(a) Modality Combination Statistics of ADNI (b) Expert Selection per Token (c) Token Selection per Expert

Figure 3: (a) Statistics of modality combinations observed in the ADNI dataset. We observe that although the ADNI dataset comprises four modalities, the modality combinations are not as diverse, showing only six unique regions. Notably, all modality combinations include the clinical modality. (b) Given an input token (i.e., modality combination), we track the expert selection ratio based on the modality combination. Alternatively, (c) from the expert's perspective, we provide how each expert selects the input token and their relative ratio. The backbone illustration of (a) is adapted from (Yun et al., 2024).

In-depth Analysis. To gain a deeper understanding of how `MoE-Retriever` functions and contributes to embedding retrieval, we provide an in-depth analysis using the ADNI dataset in Figure 3. First, as shown in Figure 3 (a), we observe six unique modality combination regions. Interestingly, the clinical modality is present in all combinations, indicating that the input token will always include the clinical ($\mathcal{C}$) modality. This also suggests that the missing modality, i.e., the target modality, will often include $\mathcal{I}$, $\mathcal{G}$, or $\mathcal{B}$, depending on its interaction with other modalities.

Next, after training `MoE-Retriever` , we track the activation ratio from both token and expert perspectives. In Figure 3 (b), we observe: **1)** `MoE-Retriever` successfully learns which modality should be selected and imputed. For example, when the token index is given as BCG, which lacks the $\mathcal{I}$ modality, the majority of tokens select image-specific experts, ranging from I_1 to I_8. **2)** This imputation tendency is also observed when the input token is BCI or BC, naturally incorporating the missing modality. This indicates that both the router and the experts are equipped with the knowledge of how to handle different input modality combinations. **3)** It is also notable that shared experts are frequently selected among activated experts, suggesting that these shared experts have learned and contain common knowledge that can interact with various modalities, aligning with the motivation behind designing shared experts as a buffer.

4) In Figure 3 (c), which shows the token selection ratio from the expert's perspective, it is expected that $BCGI$ is widely chosen by the experts, as this full modality combination is the majority in the ADNI dataset. This combination is frequently sampled through the supporting group, serving as a reference for missing cases. **5)** We also observe that experts select the necessary inputs, such as B_3, B_4, B_5, which most often select tokens like CI. **6)** In summary, by equipping the router and experts with the knowledge to select the most relevant embedding candidates, missing embeddings are

effectively retrieved to interact with other modalities. This, in turn, boosts performance in downstream tasks by leveraging intra- and inter-modal context and a context-aware routing policy.

4.4 Ablation Study

Table 3: Ablation Study.

Model Variants	Acc.	F1
MoE-Retriever	**64.52**$\pm$2.55	**63.80**$\pm$2.96
	($\lvert E_{\mathcal{T}}\rvert$=8, $\lvert E_{Sh.}\rvert$=4, $\lvert\mathcal{R}\rvert$=1)	
w/o Intra-Modal Context	61.26$\pm$2.33	61.80$\pm$1.67
w/o Inter-Modal Context	60.97$\pm$1.50	61.60$\pm$0.78
w/o Context-Aware Routing	62.34$\pm$1.25	63.11$\pm$2.11
$\lvert E_{\mathcal{T}}\rvert$=8, $\lvert E_{Sh.}\rvert$=1, $\lvert\mathcal{R}\rvert$=1	60.60$\pm$1.32	59.70$\pm$1.26
$\lvert E_{\mathcal{T}}\rvert$=8, $\lvert E_{Sh.}\rvert$=2, $\lvert\mathcal{R}\rvert$=1	63.77$\pm$1.35	62.92$\pm$0.28
$\lvert E_{\mathcal{T}}\rvert$=8, $\lvert E_{Sh.}\rvert$=8, $\lvert\mathcal{R}\rvert$=1	62.98$\pm$0.79	62.75$\pm$1.41
$\lvert E_{\mathcal{T}}\rvert$=4, $\lvert E_{Sh.}\rvert$=4, $\lvert\mathcal{R}\rvert$=1	63.14$\pm$2.47	60.88$\pm$2.21
$\lvert E_{\mathcal{T}}\rvert$=16, $\lvert E_{Sh.}\rvert$=4, $\lvert\mathcal{R}\rvert$=1	60.14$\pm$2.97	59.91$\pm$1.22
$\lvert E_{\mathcal{T}}\rvert$=8, $\lvert E_{Sh.}\rvert$=4, $\lvert\mathcal{R}\rvert$=2	61.14$\pm$1.85	61.04$\pm$1.12
$\lvert E_{\mathcal{T}}\rvert$=8, $\lvert E_{Sh.}\rvert$=4, $\lvert\mathcal{R}\rvert$=4	60.54$\pm$2.52	60.23$\pm$2.71
2 x Transformer Layer	63.34$\pm$0.97	62.79$\pm$1.31
Sparse MoE Fusion Layer	62.84$\pm$2.85	63.11$\pm$2.25

To verify the effectiveness of MoE-Retriever, we conducted an extensive ablation study using the ADNI dataset in the $\mathcal{I}+\mathcal{G}+\mathcal{C}+\mathcal{B}$ scenario. Key observations include: **1)** Regarding the core module design in MoE-Retriever, involving inter-modal context is crucial as it personalizes the specific observed modality context of each sample. **2)** When designing shared experts (E_{sh}), it is important to strike a balance in the number of shared experts. Having too many can deteriorate the acquisition of specialized knowledge required by modality-specific experts. **3)** For modality-specific experts, selecting too few or too many experts can lead to suboptimal results, emphasizing the need for a balanced number, such as eight. **4)** For the router design, utilizing a single router to handle both intra- and inter-modal contexts proved to be sufficient. The more examples it encounters during training, the more knowledge it is able to accumulate. **5)** In the subsequent fusion layer, we experimented with both a vanilla transformer design and a version with the SMoE layer attached. However, no significant performance gain was observed, suggesting that the utilization of SMoE in embedding retrieval was sufficient.

4.5 Computational Efficiency

In Figure 4, we compare the inference time for a single epoch, computational cost, and the number of parameters for each model across different modality configurations in the ADNI dataset. The results show that MoE-Retriever outperforms in all three computational dimensions: **1)** Mean Time, **2)** GFLOPs, and **3)** Number of Parameters, thanks to the adoption of the SMoE design. Notably, as the modality combinations increase, the efficiency is maintained, highlighting the advantage of SMoE, which sparsely activates the relevant parameters. This represents a significant step forward in embedding retrieval design.

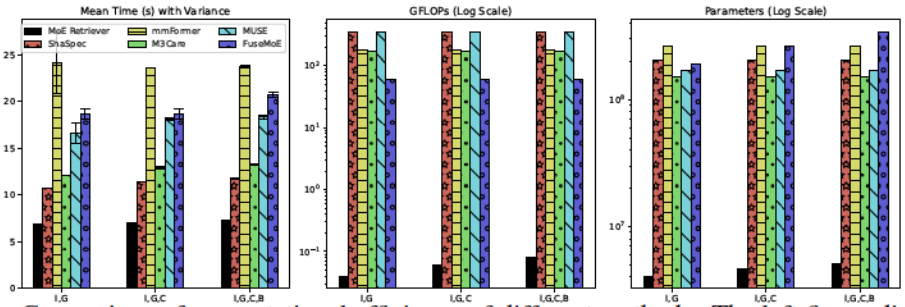


Figure 4: Comparison of computational efficiency of different methods. The left figure displays the averaged inference time for a single epoch of testing data, with error bar showing the variance. The middle plot illustrates the computational cost in GFLOPs (floating-point operations per second divided by 10^9), while the right figure shows the number of parameters on a logarithmic scale. The FLOPs and GFLOPs are computed using the fvcore package.

5 Conclusion

In this work, we propose MoE-Retriever, a novel framework inspired by the SMoE design that uniquely integrates both intra-modal and inter-modal contexts. By utilizing a modality combination based supporting group for intra-modal context and modality-specific expert which also include shared experts, MoE-Retriever effectively selects the most relevant expertes, i.e., embeddings tailored to specific missing modality scenarios. Our extensive experiments on both medical and general machine learning datasets demonstrate that MoE-Retriever not only enhances accuracy and robustness in missing modality scenarios but also exhibits scalability and computational efficiency.

REFERENCES

Alhabib Abbas and Yiannis Andreopoulos. Biased mixtures of experts: Enabling computer vision inference under data transfer limitations. *IEEE Transactions on Image Processing*, 29:7656–7667, 2020.

Karim Ahmed, Mohammad Haris Baig, and Lorenzo Torresani. Network of experts for large-scale image categorization. In *European Conference on Computer Vision*, pp. 516–532. Springer, 2016.

Raquel Aoki, Frederick Tung, and Gabriel L. Oliveira. Heterogeneous multi-task learning with expert diversity. *CoRR*, abs/2106.10595, 2021. URL https://arxiv.org/abs/2106.10595.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443, 2018.

Ke Chen, Lei Xu, and Huisheng Chi. Improved learning algorithms for mixture of experts in multiclass classification. *Neural networks*, 12(9):1229–1252, 1999.

Zitian Chen, Yikang Shen, Mingyu Ding, Zhenfang Chen, Hengshuang Zhao, Erik G. Learned-Miller, and Chuang Gan. Mod-squad: Designing mixtures of experts as modular multi-task learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11828–11837, June 2023.

Jimit Doshi, Guray Erus, Yangming Ou, Susan M Resnick, Ruben C Gur, Raquel E Gur, Theodore D Satterthwaite, Susan Furth, Christos Davatzikos, Alzheimer's Neuroimaging Initiative, et al. Muse: Multi-atlas region segmentation utilizing ensembles of registration algorithms and parameters, and locally optimal atlas selection. *Neuroimage*, 127:186–195, 2016.

Bruno Dubois, Christine AF von Arnim, Nerida Burnie, Sasha Bozeat, and Jeffrey Cummings. Biomarkers in alzheimer's disease: role in early and differential diagnosis and recognition of atypical variants. *Alzheimer's Research & Therapy*, 15(1):175, 2023.

Soheil Esmaeilzadeh, Dimitrios Ioannis Belivanis, Kilian M Pohl, and Ehsan Adeli. End-to-end alzheimer's disease diagnosis and biomarker identification. In *Machine Learning in Medical Imaging: 9th International Workshop, MLMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 9*, pp. 337–345. Springer, 2018.

Wenzhong Guo, Jianwen Wang, and Shiping Wang. Deep multimodal representation learning: A survey. *Ieee Access*, 7:63373–63394, 2019.

Xing Han, Huy Nguyen, Carl Harris, Nhat Ho, and Suchi Saria. Fusemoe: Mixture-of-experts transformers for fleximodal fusion. *arXiv preprint arXiv:2402.03226*, 2024.

Hussein Hazimeh, Zhe Zhao, Aakanksha Chowdhery, Maheswaran Sathiamoorthy, Yihua Chen, Rahul Mazumder, Lichan Hong, and Ed H. Chi. Dselect-k: Differentiable selection in the mixture of experts with applications to multi-task learning. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 29335–29347, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f5ac21cd0ef1b88e9848571aeb53551a-Abstract.html.

Clifford R Jack Jr, David A Bennett, Kaj Blennow, Maria C Carrillo, Billy Dunn, Samantha Budd Haeberlein, David M Holtzman, William Jagust, Frank Jessen, Jason Karlawish, et al. Nia-aa research framework: toward a biological definition of alzheimer's disease. *Alzheimer's & dementia*, 14(4):535–562, 2018.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Hao Jiang, Ke Zhan, Jianwei Qu, Yongkang Wu, Zhaoye Fei, Xinyu Zhang, Lei Chen, Zhicheng Dou, Xipeng Qiu, Zikai Guo, et al. Towards more effective and economic sparsely-activated model. *arXiv preprint arXiv:2110.07431*, 2021.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994.

Firas Khader, Jakob Nikolas Kather, Gustav Müller-Franzes, Tianci Wang, Tianyu Han, Soroosh Tayebi Arasteh, Karim Hamesch, Keno Bressem, Christoph Haarburger, Johannes Stegmaier, et al. Medical transformer for multimodal survival prediction in intensive care: integration of imaging and non-imaging data. *Scientific Reports*, 13(1):10666, 2023.

So Yeon Kim. Personalized explanations for early diagnosis of alzheimer's disease using explainable graph neural networks with population graphs. *Bioengineering*, 10(6):701, 2023.

Young Jin Kim, Ammar Ahmad Awan, Alexandre Muzio, Andrés Felipe Cruz-Salinas, Liyang Lu, Amr Hendy, Samyam Rajbhandari, Yuxiong He, and Hany Hassan Awadalla. Scalable and efficient moe training for multitask multilingual models. *CoRR*, abs/2109.10465, 2021. URL https://arxiv.org/abs/2109.10465.

Sneha Kudugunta, Yanping Huang, Ankur Bapna, Maxim Krikun, Dmitry Lepikhin, Minh-Thang Luong, and Orhan Firat. Beyond distillation: Task-level mixture-of-experts for efficient inference. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pp. 3577–3599. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-emnlp.304. URL https://doi.org/10.18653/v1/2021.findings-emnlp.304.

Jean-Charles Lambert, Carla A Ibrahim-Verbaas, Denise Harold, Adam C Naj, Rebecca Sims, Céline Bellenguez, Gyungah Jun, Anita L DeStefano, Joshua C Bis, Gary W Beecham, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. *Nature genetics*, 45(12):1452–1458, 2013.

Luis A Leiva, Asutosh Hota, and Antti Oulasvirta. Enrico: A dataset for topic modeling of mobile ui designs. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, pp. 1–4, 2020.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=qrwe7XHTmYb.

Lanlan Li, Xianfeng Yu, Can Sheng, Xueyan Jiang, Qi Zhang, Ying Han, and Jiehui Jiang. A review of brain imaging biomarker genomics in alzheimer's disease: implementation and perspectives. *Translational Neurodegeneration*, 11(1):42, 2022.

Mingxuan Liu, Siqi Li, Han Yuan, Marcus Eng Hock Ong, Yilin Ning, Feng Xie, Seyed Ehsan Saffari, Yuqing Shang, Victor Volovici, Bibhas Chakraborty, et al. Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques. *Artificial intelligence in medicine*, 142:102587, 2023.

Yuxuan Lou, Fuzhao Xue, Zangwei Zheng, and Yang You. Cross-token modeling with conditional computation. *arXiv preprint arXiv:2109.02008*, 2021.

Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H. Chi. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In Yike Guo and Faisal Farooq (eds.), *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pp. 1930–1939. ACM, 2018. doi: 10.1145/3219819.3220007. URL https://doi.org/10.1145/3219819.3220007.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2302–2310, 2021.

Daniele Malitesta, Emanuele Rossi, Claudio Pomo, Tommaso Di Noia, and Fragkiskos D Malliaros. Do we really need to drop items with missing modalities in multimodal recommendation? *arXiv preprint arXiv:2408.11767*, 2024.

Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts. *CoRR*, abs/2206.02770, 2022. doi: 10.48550/arXiv.2206.02770. URL https://doi.org/10.48550/arXiv.2206.02770.

Yangming Ou, Aristeidis Sotiras, Nikos Paragios, and Christos Davatzikos. Dramms: Deformable registration via attribute matching and mutual-saliency weighting. *Medical image analysis*, 15(4): 622–639, 2011.

Yongsheng Pan, Mingxia Liu, Yong Xia, and Dinggang Shen. Disease-image-specific learning for diagnosis-oriented neuroimage synthesis with incomplete multi-modality data. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):6839–6853, 2021.

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. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 8583–8595, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/48237d9f2dea8c74c2a72126cf63d933-Abstract.html.

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. *arXiv preprint arXiv:1701.06538*, 2017.

Sandra Steyaert, Marija Pizurica, Divya Nagaraj, Priya Khandelwal, Tina Hernandez-Boussard, Andrew J Gentles, and Olivier Gevaert. Multimodal data fusion for cancer biomarker discovery with deep learning. *Nature machine intelligence*, 5(4):351–362, 2023.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15878–15887, 2023.

Xin Wang, Fisher Yu, Lisa Dunlap, Yi-An Ma, Ruth Wang, Azalia Mirhoseini, Trevor Darrell, and Joseph E Gonzalez. Deep mixture of experts via shallow embedding. In *Uncertainty in artificial intelligence*, pp. 552–562. PMLR, 2020.

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, et al. The alzheimer's disease neuroimaging initiative: progress report and future plans. *Alzheimer's & Dementia*, 6(3): 202–211, 2010.

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, et al. The alzheimer's disease neuroimaging initiative 3: Continued innovation for clinical trial improvement. *Alzheimer's & Dementia*, 13(5):561–571, 2017.

Renjie Wu, Hu Wang, and Hsiang-Ting Chen. A comprehensive survey on deep multimodal learning with missing modality. *arXiv preprint arXiv:2409.07825*, 2024a.

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

Brandon Yang, Gabriel Bender, Quoc V Le, and Jiquan Ngiam. Condconv: Conditionally parameterized convolutions for efficient inference. *Advances in Neural Information Processing Systems*, 32, 2019.

Wenfang Yao, Kejing Yin, William K Cheung, Jia Liu, and Jing Qin. Drfuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16416–16424, 2024.

Seniha Esen Yuksel, Joseph N. Wilson, and Paul D. Gader. Twenty years of mixture of experts. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8):1177–1193, 2012. doi: 10.1109/TNNLS.2012.2200299.

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. *arXiv preprint arXiv:2410.08245*, 2024.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016.

Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao. M3care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2418–2428, 2022a.

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. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 107–117. Springer, 2022b.

Yue Zhang, Chengtao Peng, Qiuli Wang, Dan Song, Kaiyan Li, and S Kevin Zhou. Unified multimodal image synthesis for missing modality imputation. *IEEE Transactions on Medical Imaging*, 2024.

Zhengyan Zhang, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. Moefication: Conditional computation of transformer models for efficient inference. *arXiv preprint arXiv:2110.01786*, 2021.

Tongxue Zhou, Su Ruan, and Haigen Hu. A literature survey of mr-based brain tumor segmentation with missing modalities. *Computerized Medical Imaging and Graphics*, 104:102167, 2023.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew Dai, Zhifeng Chen, Quoc Le, and James Laudon. Mixture-of-experts with expert choice routing. *arXiv preprint arXiv:2202.09368*, 2022.

Simiao Zuo, Xiaodong Liu, Jian Jiao, Young Jin Kim, Hany Hassan, Ruofei Zhang, Jianfeng Gao, and Tuo Zhao. Taming sparsely activated transformer with stochastic experts. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=B72HXs80q4.

A Appendix

A.1 Detailed Preprocessing

We followed the same preprocessing procedure of the ADNI dataset and MIMIC dataset, as described in Flex-MoE (Yun et al., 2024).

A.1.1 Detailed Data Preprocessing in ADNI

Image Modality To preprocess the image data, we first applied a correction for magnetic field intensity inhomogeneity to ensure consistency and reliability across MRI images. Next, we used the MUSE (Multiatlas Region Segmentation Utilizing Ensembles of Registration Algorithms and Parameters) method to segment gray matter tissue, the primary focus of this study (Doshi et al., 2016). This technique involves utilizing multiple atlases to extract the most accurate region-of-interest values from the segmented gray matter. Afterward, voxel-wise volumetric maps of tissue regions

were created by spatially aligning skull-stripped images to a template in the Montreal Neurological Institute (MNI) space, using a registration method (Ou et al., 2011).

Genetic Modality We obtained SNP (single nucleotide polymorphisms) data from the ADNI 1, GO/2, and 3 studies, and pre-processed it as follows. First, SNP data from these studies were aligned to a unified reference build using Liftover `https://liftover.broadinstitute.org/`, converting all data to NCBI build 37 (UCSC hg19). Next, we aligned strands based on the 1000 Genome Project phase 3, using McCarthy Group Tools `https://www.well.ox.ac.uk/~wrayner/tools/`. Linkage disequilibrium (LD) pruning was then applied with parameters (50, 5, 0.1) to remove highly correlated SNPs, reducing the total SNPs from 565, 989 to 144, 746. Imputation was performed on this pruned set using the Michigan Imputation Server `https://imputationserver.sph.umich.edu/index.html#!`, and the resulting SNP data was recoded as $\{0, 1, 2\}$.

Biospecimen Modality Biospecimen data was extracted from several ADNI-provided csv files. CSF Aβ1-42 and Aβ1-40 data were taken from ISOPROSTANE_09May2024.csv, Total Tau and Phosphorylated Tau from UPENNBIOMK_ROCHE_ELECSYS_09May2024.csv, Plasma Neurofilament Light Chain data from batemanlab_20221118_09May2024.csv, and ApoE genotype data from APOERES_09May2024.csv. Numerical data was scaled using a MinMax scaler to a range of -1 to 1, while categorical data was one-hot encoded. For missing values, we imputed the mean for numerical fields and the mode for categorical fields.

Clinical Modality Clinical data was extracted from ADNI csv files, including MEDHIST_09May2024.csv, NEUROEXM_09May2024.csv, PTDEMOG_09May2024.csv, RECCMEDS_09May2024.csv, and VITALS_09May2024.csv. During preprocessing, we excluded the columns 'PTCOGBEG,' 'PTADDX,' and 'PTADBEG,' which contain direct Alzheimer's Disease diagnosis information. Numerical data was scaled using a MinMax scaler (-1 to 1), while categorical data was one-hot encoded. Missing values were imputed by using the mean for numerical columns and the mode for categorical columns.

A.1.2 Detailed Data Preprocessing in MIMIC

Lab, Notes, Codes Modalities. For the MIMIC dataset, we use the Medical Information Mart for Intensive Care IV (MIMIC-IV) database, which contains de-identified health data for patients who were admitted to either the emergency department or stayed in critical care units of the Beth Israel Deaconess Medical Center in Boston, Massachusetts24. MIMIC-IV excludes patients under 18 years of age. We take a subset of the MIMIC-IV data, where each patient has at least more than 1 visit in the dataset as this subset corresponds to patients who likely have more serious health conditions. For each datapoint, we extract ICD-9 codes, clinical text, and labs and vital values. Using this data, we perform binary classification on one-year mortality, which foresees whether or not this patient will pass away in a year. We drop visits that occur at the same time as the patient's death.

Missingness in MIMIC dataset. Code Modality: This combines diagnosis and procedure data. There are 4 records with missing diagnoses and 1777 with missing procedures. **Note Modality**: Derived from the "text" column of the original CSV file, there are 108 records with missing notes. **Lab Modality**: This presents a more complex scenario, as it includes 2172 different measurements. If we consider all 2172 measurements as potentially missing, then technically, there is no missing data since essential measurements, like heart rate, are consistently collected for each patient. However, if we evaluate the proportion of missing values in the (9003, 2172) matrix, we find that 94.216% of the entries are NaN.

A.2 Modality-specific Encoder Settings

ADNI Dataset. For image modality, we used a customized 3D-CNN (Esmaeilzadeh et al., 2018) with hidden dimension 256 as encoder . For genomics, clinical, and biospecimen modalities, we used MLP with hidden dimension 256 as encoder. **MIMIC Dataset.** For all lab, note, and code modalities, we used LSTM with hidden dimension 256 as encoder. **ENRICO Dataset.** For both screenshot image and wireframe image modality, we used VGG11 from torchvision library with hidden dimension size 16 as encoder. **CMU-MOSI Dataset.** For both vision, audio, and text modality, we used Gated Recurrent Unit with hidden dimension 256 as encoder.

A.3 DIFFERENT NOISE LEVEL OF CMU-MOSI DATASET

Dataset	Modality	Noise Level	mmFormer	ShaSpec	M3Care	MUSE	FuseMoE	MoE-Retriever
CMU-MOSI	$\mathcal{V}$+$\mathcal{A}$	0.1	51.82±1.36	52.90±3.28	52.62±1.59	51.49±1.95	50.25±2.97	**53.19**±1.22
		0.3	50.60±2.26	46.61±2.36	47.26±2.66	50.70±2.01	47.46±1.51	**53.12**±2.34
		0.5	**49.95**±1.58	47.39±5.97	42.45±3.68	49.42±1.93	46.91±3.44	49.67±3.80
	$\mathcal{V}$+$\mathcal{T}$	0.1	62.51±1.43	65.75±1.39	69.20±0.08	49.02±1.95	63.49±0.98	**66.29**±1.99
		0.3	59.78±1.10	62.04±1.12	62.02±0.26	49.16±1.91	55.41±3.12	**63.29**±2.54
		0.5	57.85±0.65	59.37±1.36	54.63±0.76	48.71±1.92	52.39±1.73	**62.49**±1.36
	$\mathcal{A}$+$\mathcal{V}$	0.1	63.54±0.73	66.98±0.48	67.79±4.42	53.00±1.94	69.58±0.51	**69.78**±2.21
		0.3	61.40±1.17	63.07±4.19	65.77±2.69	46.93±1.93	58.11±2.24	**65.27**±0.38
		0.5	55.61±3.14	58.06±3.65	46.83±5.06	44.73±1.91	50.86±4.65	**62.14**±3.17
	$\mathcal{V}$+$\mathcal{A}$+$\mathcal{T}$	0.1	64.38±1.37	67.78±1.56	68.29±1.24	49.42±1.94	69.87±1.86	**70.36**±1.42
		0.3	59.37±0.96	63.76±2.56	63.11±5.80	49.39±1.89	63.97±0.88	**65.25**±4.06
		0.5	54.41±3.78	55.71±2.37	48.65±3.62	48.86±1.91	47.67±3.48	**58.48**±1.06

Table 4: Accuracy in CMU-MOSI dataset across different modalities and noise levels

A.4 AUROC AND PRAUC RESULTS IN MIMIC DATASET

Modality	mmFormer	ShaSpec	M3Care	MUSE	FuseMoE	MoE-Retriever
$\mathcal{L}$+$\mathcal{N}$	67.28±2.49	66.08±2.87	50.63±1.44	55.22±1.85	66.05±0.98	**68.33**±0.50
$\mathcal{L}$+$\mathcal{C}$	65.26±2.17	64.61±1.18	50.09±0.93	59.15±1.81	62.53±2.44	**67.41**±0.57
$\mathcal{N}$+$\mathcal{C}$	62.71±2.31	65.61±1.43	51.06±2.64	49.64±1.86	66.61±0.65	**65.01**±1.01
$\mathcal{L}$+$\mathcal{N}$+$\mathcal{C}$	68.42±1.65	69.27±0.14	50.08±0.21	67.4±1.67	66.65±0.78	**69.39**±1.08

Table 5: AUROC Results of MIMIC with different modality combinations

Modality	mmFormer	ShaSpec	M3Care	MUSE	FuseMoE	MoE-Retriever
$\mathcal{L}$+$\mathcal{N}$	35.20±2.94	34.07±2.26	23.17±0.85	27.03±2.12	33.50±1.01	**36.46**±0.66
$\mathcal{L}$+$\mathcal{C}$	34.07±1.46	33.76±0.67	23.15±0.70	29.8±2.27	32.19±0.68	**34.50**±1.41
$\mathcal{N}$+$\mathcal{C}$	30.97±2.89	34.36±1.55	23.29±1.23	21.12±1.34	**35.24**±0.34	33.29±0.85
$\mathcal{L}$+$\mathcal{N}$+$\mathcal{C}$	36.54±1.24	36.62±1.17	22.66±0.25	35.23±2.54	34.59±1.40	**36.83**±0.10

Table 6: PRAUC Results of MIMIC with different modality combinations

A.5 HYPERPARAMETER TUNING OF BASELINE MODEL

Learning Rate	Hidden Dimension	Acc.	F1	AUROC	PRAUC
1e-4	64	64.58 ± 1.88	64.25 ± 1.55	70.51 ± 0.43	63.86 ± 1.18
	128	65.17 ± 0.80	64.87 ± 0.78	72.00 ± 2.48	64.72 ± 4.57
	256	63.67 ± 0.76	63.61 ± 0.73	72.77 ± 0.96	66.17 ± 2.24
1e-3	64	63.92 ± 1.37	63.66 ± 1.45	71.98 ± 2.33	65.93 ± 3.33
	128	64.09 ± 0.07	63.59 ± 0.14	72.46 ± 1.21	67.61 ± 2.29
	256	64.32 ± 0.61	64.18 ± 0.54	73.33 ± 1.24	66.59 ± 3.51
1e-2	64	64.51 ± 2.61	64.33 ± 2.49	72.67 ± 2.10	65.29 ± 1.75
	128	65.06 ± 3.25	64.63 ± 2.99	73.18 ± 2.71	66.62 ± 2.82
	256	63.62 ± 0.50	63.29 ± 0.54	71.81 ± 0.73	63.85 ± 1.50

Table 7: Results of different hyperparameters of CMU-MOSI Modality $\mathcal{V}$+$\mathcal{A}$+$\mathcal{T}$. Learning rate $1e-3$ and hidden dimension 128 are the optimal hyperparameter provided in the ShaSpec paper.

A.6 GRADIENT CONFLICT IN CMU-MOSI DATASET

In Table 4 of the main paper, we observe that when three modalities ($\mathcal{V}$+$\mathcal{A}$+$\mathcal{T}$) are used, the performance for all models does not reach its peak, even compared to using two modalities ($\mathcal{V}$+$\mathcal{T}$ or $\mathcal{A}$+$\mathcal{T}$). This highlights an intriguing phenomenon: *adding more modalities does not always guarantee improved performance*. To investigate this from an optimization perspective, we analyze the gradients when all modalities are provided. Specifically, we compute the derivative of the loss with respect to each modality and measure the cosine similarities between modality pairs to detect potential gradient conflicts. Higher cosine similarity indicates positive correlation between gradients, while lower values suggest conflicts.

In Figure 5(a), where dense models such as ShaSpec (Wang et al., 2023) or mmFormer (Zhang et al., 2022b) (leveraging fully connected layers or Transformers) are used in the fusion layer, $\mathcal{V}$+$\mathcal{A}$ shows positive synergy. However, other pairs, such as $\mathcal{V}$+$\mathcal{T}$ and $\mathcal{A}$+$\mathcal{T}$, exhibit less positive interactions, leading to challenges in simultaneously optimizing $\mathcal{V}$+$\mathcal{A}$+$\mathcal{T}$ and negatively impacting overall performance. In contrast, as shown in Figure 5(b), sparse models like FuseMoE (Han et al., 2024), which use SMoE in the fusion layer, demonstrate improved synergy due to SMoE's ability to selectively activate the most relevant experts, thereby reducing interference between modalities.

Finally, in our approach (`MoE-Retriever`), as shown in Figure 5(c), SMoE is applied prior to the fusion layer to retrieve missing modalities before the fusion step. This design further enhances the synergy between modalities, resulting in better optimization compared to other baselines, achieving

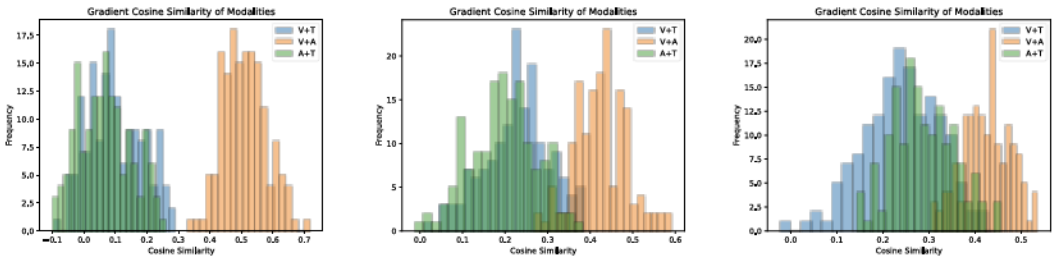
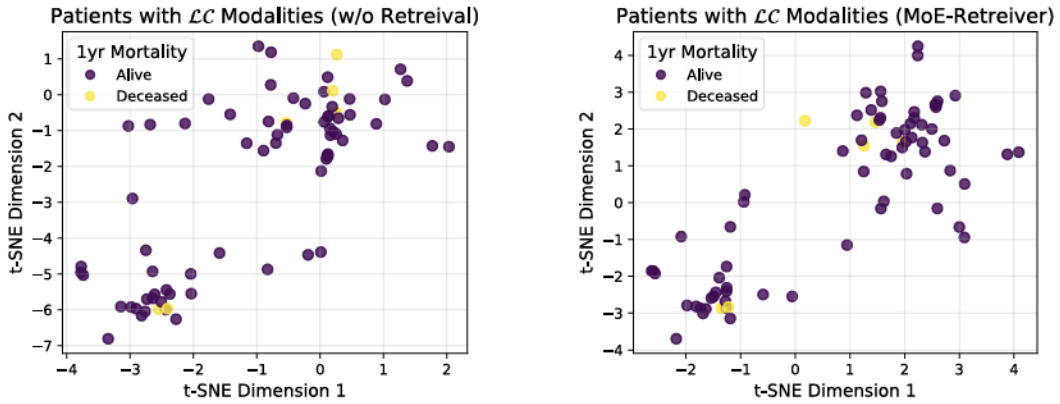


Figure 5: Gradient similarities of two paired modalities in the CMU-MOSI dataset. (a) Dense model, where the fusion layer is based on a transformer model. (b) Sparse model, where the fusion layer adopts a Sparse Mixture-of-Experts backbone. (c) Ours (`MoE-Retriever`), where SMoE is utilized prior to the fusion layer to retrieve the missing modality. Higher cosine similarity indicates that the gradient operates in a more positive (i.e., same) direction during optimization.

A.7 EFFECTIVENESS OF RETRIEVAL FROM THE MIMIC DATASET

To further verify the benefits of the feature retrieval process at the sample (i.e., patient) level, we present a t-SNE plot in Figure 6 using the MIMIC dataset for the one-year mortality prediction task. Specifically, we demonstrate patient-level embeddings, focusing on patients with $\mathcal{LC}$ modalities but missing the $\mathcal{N}$ modality. In Figure 6(a), we observe that before retrieving the $\mathcal{N}$ modality, the



(a) Before Retrieving $\mathcal{N}$ (Note) Modality (b) After Retrieving $\mathcal{N}$ (Note) Modality

Figure 6: t-SNE plot comparison on the MIMIC dataset with patients having $\mathcal{LC}$ modalities. (a) Before retrieving the $\mathcal{N}$ modality, the embeddings of alive patients are not condensed and appear more scattered. (b) After retrieving the $\mathcal{N}$ modality via `MoE-Retriever`, the embeddings of patients with the alive label become more similar to each other.