Flex-MoE: Modeling Arbitrary Modality Combination via the Flexible Mixture-of-Experts

Sukwon Yun[1], **Inyoung Choi**[2], **Jie Peng**[3], **Yangfan Wu**[3], **Jingxuan Bao**[2], **Qiyiwen Zhang**[2], **Jiayi Xin**[2], **Qi Long**[2], **Tianlong Chen**[1]

[1]University of North Carolina at Chapel Hill
[2]University of Pennsylvania
[3]University of Science and Technology of China

{swyun, tianlong}@cs.unc.edu, {inyoungc, jiayixin}@seas.upenn.edu,
{pengjieb, ustc_wyf}@mail.ustc.edu.cn
{jingxuan.bao, qiyiwen.zhang}@pennmedicine.upenn.edu, qlong@upenn.edu

Abstract

Multimodal learning has gained increasing importance across various fields, offering the ability to integrate data from diverse sources such as images, text, and personalized records, which are frequently observed in medical domains. However, in scenarios where some modalities are missing, many existing frameworks struggle to accommodate arbitrary modality combinations, often relying heavily on a single modality or complete data. This oversight of potential modality combinations limits their applicability in real-world situations. To address this challenge, we propose Flex-MoE (Flexible Mixture-of-Experts), a new framework designed to flexibly incorporate arbitrary modality combinations while maintaining robustness to missing data. The core idea of Flex-MoE is to first address missing modalities using a new missing modality bank that integrates observed modality combinations with the corresponding missing ones. This is followed by a uniquely designed Sparse MoE framework. Specifically, Flex-MoE first trains experts using samples with all modalities to inject generalized knowledge through the generalized router ($\mathcal{G}$-Router). The $\mathcal{S}$-Router then specializes in handling fewer modality combinations by assigning the top-1 gate to the expert corresponding to the observed modality combination. We evaluate Flex-MoE on the ADNI dataset, which encompasses four modalities in the Alzheimer's Disease domain, as well as on the MIMIC-IV dataset. The results demonstrate the effectiveness of Flex-MoE, highlighting its ability to model arbitrary modality combinations in diverse missing modality scenarios. Code is available at: https://github.com/UNITES-Lab/flex-moe.

1 Introduction

In many fields, including healthcare, language, and vision, multimodal learning [6, 75, 37, 44] has emerged as a crucial approach for integrating data from multiple sources such as clinical records, imaging, and genetic data. Multimodal data enables more comprehensive analysis and decision-making, offering the potential for improved diagnosis and prediction in various applications [59, 33, 68]. However, a prominent challenge across these domains is the missing modality scenario [76, 60], where not all modalities are consistently available for every instance due to diverse reasons such as individualized data collection protocols or the variable availability of certain modalities.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-3135

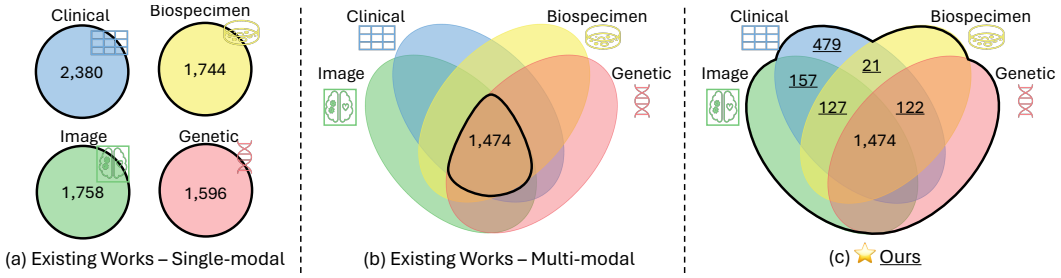

Figure 2: Data statistics from a real-world multimodal dataset (e.g., the Alzheimer's Disease Neuroimaging Initiative (ADNI)), where patients exhibit unique combinations of available modalities. Existing approaches focus on either (a) single-modality data or (b) complete multimodal data, losing the potential to leverage other combinations. Our approach incorporates all possible modality combinations, offering a more robust solution to the missing modality scenario.

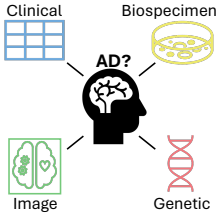

Figure 1: Multimodal AD.

As an representative example, in Alzheimer's Disease (AD) [4], one of the most prevalent neurodegenerative disorders, handling this inherently multimodal data is crucial for accurate diagnosis (Figure 1). AD datasets often include a combination of clinical symptoms, imaging data [42], and genetic profiles [48]. However, in real-world clinical settings, not all these modalities are readily available for each patient. Some data, such as clinical and imaging data, may be available from routine visits, whereas other data, such as genetic or biospecimen information, may require additional time to collect. This leads to incomplete datasets, which poses a challenge for existing models that tend to either rely heavily on single modalities or only utilize complete data, thereby missing the opportunity to leverage the full potential of multimodal learning (Figure 2).

Single Modality and Complete Data Reliance. The reliance on single-modality data or complete data across many frameworks is a significant limitation in real-world scenarios, where missing data is the norm rather than the exception. As seen in Figure 2, many current models either work with single-modality data or focus on the intersection of modalities, neglecting the potential contribution of partially available modalities. In healthcare, particularly for diseases such as AD, this often leads to missed opportunities in diagnosis and treatment due to the inability to fully exploit multimodal data when some modalities are missing.

Oversight of Modality Combinations. Beyond the challenge of missing modalities, there is also the need to model the interactions between available modalities properly. Different combinations of modalities can provide complementary information, and each combination may hold unique significance for downstream tasks. For example, in AD diagnosis, combining biospecimen data and imaging data can reveal key insights: cerebrospinal fluid biomarkers may indicate early signs of AD [21], while functional MRI can highlight cognitive impairments [42]. Hence, it is essential to develop models that not only handle missing modalities but also effectively utilize the available modality combinations.

Given the general challenge of the missing modality scenario in multimodal learning, we propose a novel framework, `Flex-MoE` (Flexible Mixture-of-Experts), to flexibly incorporate arbitrary modality combinations while maintaining robustness to missing data. `Flex-MoE` first sort samples based on the available modalities and process them through modality-specific encoders. For missing modalities, we introduce a learnable missing modality bank, which provides learnable embeddings for missing modalities based on the observed ones. This approach ensures that the model can handle incomplete datasets effectively. Our framework also builds upon Sparse Mixture-of-Experts (SMoE) design, allowing us to generalize the expert knowledge from complete data (samples with all modalities) through the $\mathcal{G}$-Router, followed by a specialized $\mathcal{S}$-Router for handling fewer modality combinations. Each expert becomes specialized in handling different modality combinations, ensuring that the model can effectively process any combination of modalities. We demonstrate the effectiveness of `Flex-MoE` through comprehensive experiments on several real-world datasets, including the Alzheimer's Disease Neuroimaging Initiative (ADNI), which involves four key modalities for AD stage prediction, and the MIMIC-IV dataset. The results confirm the robustness of `Flex-MoE` in diverse missing modality scenarios.

The contributions of this work can be summarized as follows:

- ⋆ We introduce a flexible framework that effectively incorporates arbitrary modality combinations and addresses the missing modality scenario across various domains.
- ⋆ `Flex-MoE` features a novel approach, including a missing modality bank and generalized and specialized expert training, which ensures robustness to missing modality scenario.
- ⋆ Extensive experiments on real-world datasets, including ADNI and MIMIC-IV, showcase the consistent and robust performance of `Flex-MoE` in handling diverse modality combinations.

2 Related Works

Single Modality Approach In many fields, deep learning models often rely on single modality data for tasks such as classification [15, 13, 31, 71], diagnosis [56, 72], or prediction [12, 61, 34]. While effective in certain cases, these approaches fail to capture the potential synergies between different data sources, especially in contexts where multiple modalities are available. In the Alzheimer's Disease domain, many studies focus on specific modalities. For instance, image-based approaches include a VGG19 model [43] that diagnoses early-stage AD from MRI scans and a modified ResNet18 architecture [45] that predicts AD progression using fMRI data. Other studies focus on genomics, such as DLG [36] for classifying AD patients and SWAT-CNN [26] for discovering AD-associated genetic variants. In the biospecimen modality, a deep learning-assisted spectroscopy platform [29] diagnoses AD by analyzing blood-based amyloid-beta and metabolite biomarkers. Regarding clinical data, a deep learning model [5] outperforms earlier machine learning techniques in classifying AD patients. However, since AD data is inherently multimodal, methods based on a single modality are suboptimal, missing the potential to leverage interactions between different modalities.

Multimodal Approach Across multiple fields, multimodal learning has become increasingly valuable for its ability to integrate and capture dynamics within and across different modalities, providing richer and more comprehensive representations of data. Approaches such as the Tensor Fusion Network [74], Multimodal Transformer [58], and Multimodal Adaptation Gate [53] highlight the effectiveness of combining multiple data sources. Recently, sparse mixture-of-experts-based methods, such as [44, 8, 19], have been introduced to enhance modality interactions, though these methods are still relatively unexplored in the AD domain due to the complexity of handling various modality combinations. In AD research, some works have emerged to leverage multimodal data, such as [46] and [59], which integrated a deep learning framework that combines imaging, genetic, and clinical data, achieving superior AD staging accuracy. Another multimodal model [33], incorporating longitudinal and cross-sectional data, provided more accurate AD predictions. While multimodal AD studies have shown significant progress, the challenge of missing modalities, especially in the context of how to effectively cope with modality combinations, remains largely underexplored.

3 Methods

3.1 Preliminaries and Notations

Why Sparse Mixture-of-Experts? Given a multimodal nature, we choose to utilize Sparse Mixture-of-Experts (SMoE) [55] due to its computational efficiency and its ability to handle multimodal data by effectively alleviating the gradient conflict optimization issue between modalities [50]. To briefly introduce SMoE, Traditional Mixture-of-Experts (MoE) models [23, 28, 9, 70] evolved by incorporating sparsity into their structure, optimizing computational efficiency and model performance. SMoE selectively activates only the most relevant experts for a given task, reducing overhead and improving scalability. This innovation is particularly beneficial in handling complex, high-dimensional datasets across diverse applications. It has been widely used in vision [54, 40, 16, 2, 18, 62, 69, 1, 49] and language processing [35, 30, 78, 77, 79, 25] fields, dynamically assigning different parts of the network to specific tasks [41, 3, 20, 11] or data modalities [32, 44]. Research has shown its effectiveness in classification tasks in digital number recognition [20] and medical signal processing [3]. In this work, we further explore the use of SMoE to model arbitrary modality combinations and address the missing modality scenario.

Notation. Formally, the SMoE consists of multiple experts, denoted as $f_1, f_2, \ldots, f_{|E|}$, where $|E|$ represents the total number of experts, and a router, $\mathcal{R}$, which is responsible for the routing mechanism and sparsely selects the top-k experts. For a given embedding or token $\mathbf{x}$, the router $\mathcal{R}$

 https://doi.org/10.52202/079017-3135

engages the top-k experts based on the highest scores obtained from softmax function with learnable gating function, $g(\cdot)$ (usually one or two layer MLP), and output $\mathcal{R}(\mathbf{x})_i$, where i denotes the expert index. This process can be described as follows:

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

3.2 Our approach: `Flex-MoE`

In this section, we present our novel algorithm, `Flex-MoE`, specifically designed to flexibly address the challenge of missing modalities in the multimodal domain. We start by sorting the samples based on their number of observed modalities. Following a modality-specific encoder, we supplement the embeddings for missing parts via missing modality bank completion. This effectively manages missing modalities by learning embedding banks that capture the information specific to observed modality combinations. Next, a Transformer coupled with an SMoE layer is employed. We introduce an expert generalization and specialization step to optimize modality utilization by fully leveraging samples with complete modalities and obtaining modality combination-specific knowledge through samples with fewer modalities. A comprehensive illustration of `Flex-MoE` is provided in Figure 3. Throughout the details in the following section, while our work is exemplified through the AD domain for predicting AD stages using four representative modalities—image, clinical, biospecimen, and genetic—it is important to note that `Flex-MoE` can be generalized to any other multimodal domain.

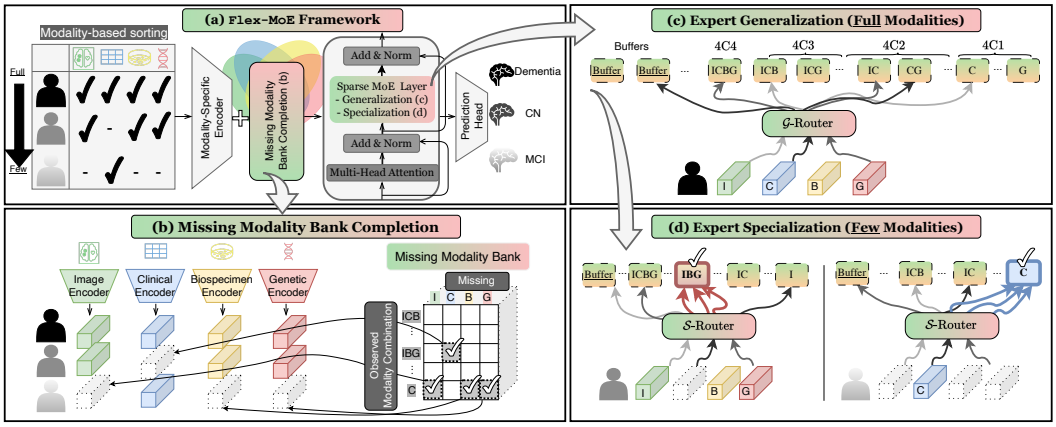

Figure 3: The comprehensive illustration of our proposed methodology, `Flex-MoE`. (a) Overall framework of `Flex-MoE`. Given samples with diverse modality combinations, we first sort the samples based on their number of available modalities in descending order, and then pass through the modality-specific encoder. (b) Each encoder is only trained with their available samples. For the missing embeddings, we introduce a missing modality bank containing learnable embeddings given the observed modality combination with their corresponding missing modality index. Equipped with this embedding, `Flex-MoE` passes through the Transformer where the FFN layer is replaced with a Sparse MoE layer. Here, (c) full modality samples take charge of training generalized experts in a balanced manner via $\mathcal{G}$-router, then (d) the remaining few modality combinations further specialize the expert knowledge with $\mathcal{S}$-Router, which fixes the top-1 gate as the corresponding observed modality combination expert. In this figure, top-2 selection of experts is illustrated as an example.

3.2.1 Missing Modality Bank Completion

Given a set of samples with their own modalities, it is straightforward to pass them through modality-specific encoders, such as a 3D-CNN for MRI images. However, we are dealing with a *missing*

modality scenario in multimodal data, where specific modalities are often missing based on their observed modality combinations. Thus, it is common to use padded or imputed inputs for the corresponding missing modalities in a multimodal setting. This approach becomes troublesome when considering interactions between modalities. For instance, given a batch with samples, some samples might have image, genetic, and clinical modalities, while others might be missing image and genetic modalities. In such cases, the image encoder and genetic encoder would take zero-padded or hypothesized imputed inputs derived from the missing samples, which are synthetic and of lower quality than the observed ones. This negatively affects the training of modality-specific encoders. Additionally, heavy imputation for each modality in a multimodal setting increases time complexity, which is not desirable in clinical settings.

Given this situation, we propose training each encoder *solely with observed samples* to fully leverage the potential of the encoder by using only observed inputs. Unlike existing approaches, our design principle considers modality combinations to ensure robust and flexible training and effective handling of missing modalities. As illustrated in Figure 1, each patient has diverse symptoms and personalized treatments, leading to variations in available modalities. For example, patients might lack image and genomic modalities (i.e., possessing only biospecimen and clinical modalities) due to various reasons such as patient conditions, resource limitations, or specific clinical settings [22, 51]. Imputing these missing modalities must be handled within this context, rather than applying a global learnable representative embedding for each modality without considering the observed environment.

Therefore, we propose a learnable missing modality bank. Given the number of modality combinations without fully observed scenarios, the total cases would be, given a modality set $\mathcal{M} = \{\mathcal{I}, \mathcal{C}, \mathcal{B}, \mathcal{G}\}$, $\sum_{m=1}^{|\mathcal{M}|-1} \binom{|\mathcal{M}|}{m} = 2^{|\mathcal{M}|} - 1$. The resulting missing modality bank can be defined as $\mathbf{B} \in \mathbb{R}^{2^{|\mathcal{M}|}-1 \times |\mathcal{M}|}$. By using this bank, the concatenated embedding of all modalities for patient i, $\mathbf{h}_i = [\mathbf{e}_i^{\mathcal{I}}, \mathbf{e}_i^{\mathcal{C}}, \mathbf{e}_i^{\mathcal{B}}, \mathbf{e}_i^{\mathcal{G}}]$ would be represented as follows:

$$\mathbf{e}_i^m = \begin{cases} \text{Encoder}^m(i) & \text{if modality } m \text{ is observed in sample } i \\ \mathbf{B}_{\mathcal{M}\setminus m, m} & \text{otherwise} \end{cases}, \quad \forall m \in \{\mathcal{I}, \mathcal{C}, \mathcal{B}, \mathcal{G}\} \tag{2}$$

where $\mathbf{e}_i^m \in \mathbb{R}^d$ denotes the embedding from the corresponding modality m for patient i, and d is the hidden dimension. Here, the main idea of the missing modality bank completion is to supplement missing modalities from a predefined bank, ensuring robust data integration with observed ones. For example, if a patient lacks clinical data but has imaging, biospecimen, and genetic data, the observed modalities pass through their specific encoders. The missing clinical embedding is supplemented from the missing modality bank, indexed by the observed modalities (e.g., $\mathbf{B}_{\mathcal{M}\setminus m, m} = \mathbf{B}_{\{\mathcal{I},\mathcal{G},\mathcal{B}\},\mathcal{C}}$). By doing so, the encoder for each modality can be trained without encountering non-observed, incomplete input features. Then, we move on to the Transformer layer, where the FFN layer is replaced by an SMoE layer, a common approach in the SMoE domain [24, 38, 10].

3.2.2 Expert Generalization & Specialization

While adopting the SMoE backbone, it is important to note that our environment differs from concurrent SMoE studies, especially in terms of multimodal learning with missing modalities. In this context, choosing the most relevant tokens is challenging for experts, since the significance, i.e., the quality of input information, varies with the missing modalities. This significantly motivates us to take a distinct approach from concurrent SMoE studies, where the input token is derived from fully observed scenarios. To address the unique challenges of the missing modality scenario, we propose a modality combination-specific MoE design. Specifically, we assign expert indices based on all possible modality combinations. For example, 'IGCB' is assigned as 0, 'IGC' as 1, ..., up to 'B' as 14. The remaining experts act as buffers, allowing the Router to select the most relevant top-k experts and activate them automatically. This approach leaves room for flexibility and maintains the initial intuition of the MoE design.

Generalization It now becomes clear why the samples used for training `Flex-MoE` are sorted in descending order. Inspired by curriculum learning [7, 63], where easy samples are presented first and more challenging samples appear later, we regard the level of difficulty as the number of missing modalities. We first train our SMoE layer with easy samples, where all modalities are fully observed. Using this intersection as a gold standard, we initially train all the experts in the MoE model. The procedure essentially follows the vanilla SMoE design as described in Equation 1, but with one key

 https://doi.org/10.52202/079017-3135

difference: the input tokens consist only of inputs where all modalities are fully observed. Hence, we refer to this router as the Generalized Router, $\mathcal{G}$-Router. This approach leverages the completeness of information in these samples, which should be fully utilized before specializing the experts in their respective areas. To ensure balanced activation of the experts initially, which will later specialize, we incorporate the load and importance balancing loss [55], which will later be exemplified in Equation 4.

Specialization Once the experts are initially trained using fully observed samples, we aim to specialize each expert, which is the key advantage of the MoE design. We leverage the remaining samples, which encompass diverse modality combination configurations. Each modality combination requires its own specialized expertise. For instance, samples with Image, Biospecimen, and Genetic data will have a corresponding expert index activated through the top-1 gating mechanism to fully utilize the specialized knowledge of that expert (i.e., expert 'IBG' in Figure 3). To effectively specialize the modality combination-specific experts, we propose a Specialized Router design, $\mathcal{S}$-Router, which encompasses following technical novelties. First, to facilitate targeted expert selection when an input token is provided, we avoid manually replacing the selected routing policy with our preferred choice in a post-hoc manner, which would stop the continuous gradient flow. Instead, we innovatively introduce a cross-entropy loss between the top-1 expert selection and the targeted expert indices for each token by the S-Router. Formally, this can be described as follows:

$$\mathcal{L}_{ce} = -\sum_{j=1}^{n} \mathcal{MC}(\mathbf{x}_j) \log(\max(\mathcal{S}\text{-Router}(\mathbf{x}_j))) \tag{3}$$

where $\mathcal{MC}(\mathbf{x}_j)$ denotes the modality combination index of a given token $\mathbf{x}_j$ in a total of n tokens. $\max(\mathcal{S}\text{-Router}(\mathbf{x}_j))$ denotes the maximum prediction probability of the corresponding activated expert index, which corresponds to the probability of the top-1 expert index. By comparing these two, the router is trained to activate the corresponding expert index for a given input token with a certain modality combination. Thus, the specialized knowledge inherent in specific modality combinations is naturally contained within the target expert.

Moreover, when calculating load and importance balancing loss [55], we specifically compute the loss for the *remaining* top-k-1 experts, as the top-1 selection is manually handled and thus considered biased rather than balanced. We aim for the selection of the remaining k-1 experts to occur in a balanced manner, allowing interaction with other related modality combinations. Formarlly, it can be expressed as follows:

$$\begin{aligned} \mathcal{L}_{\text{balance}} &= \text{CV}^2\left(\sum_{j}^{N} \text{importance}_j\right) + \text{CV}^2\left(\sum_{j}^{N} \text{load}_j\right) \\ \text{where}\ \ \text{importance}_e &= \sum_{i}^{N} g_{ie},\ \ \text{load}_e = \sum_{i}^{N} \delta(g_{ie} > 0), \quad \forall e \in E \setminus e_{\text{top-1}} \end{aligned} \tag{4}$$

where $\text{CV}^2(x) = \left(\frac{\sigma(x)}{\mu(x)}\right)^2$, $\sigma(x)$ is the standard deviation of x, $\mu(x)$ is the mean of x, g_{ie} is the gate value for sample i with expert index e as discussed in Equation 1, and $\delta(\cdot > 0)$ is an indicator function that is 1 when the inner value is greater than 0. $E \setminus e_{\text{top-1}}$ denotes the set of expert indices excluding the top-1 selected expert index. This ensures that the resulting MoE model not only retains global knowledge but also incorporates specialized expert knowledge tailored to specific modality combinations. During the inference phase, the specified expert index for a particular modality combination can be activated alongside other experts, enabling predictions to consider both the specific modality combination and intersections with other modalities.

Finally, the output embeddings of the Sparse MoE layer pass through a 1-layer MLP prediction head to predict one of the three stages of AD, i.e., Dementia, CN, or MCI. To further facilitate a curriculum-learning approach, we first use warm-up epochs with sorted samples, followed by shuffled samples for the remaining epochs. This strategy enhances generalizability across diverse input samples, enabling better handling of variability during the inference phase.

4 Experiments

4.1 Experimental Setting

ADNI Dataset Alzheimer's Disease Neuroimaging Initiative (ADNI) is a landmark multimodal AD dataset that tracks disease progression and pathological changes, comprising of comprehensive imaging, genetic, clinical, and biospecimen data ([64], [67]). The imaging data in ADNI includes magnetic resonance imaging (MRI) and positron emission tomogrpahy (PET). The genetic data includes a variety of genetic information, including genotyping data such as APOE genotyping and single nucleotide polymorphisms. The clinical data includes demographics, physical examinations, and cognitive assessments. Biospecimens such as blood, urine, and cerebrospinal fluid are also collected. ADNI has established standardized multi-center protocols and provides open access to qualified researchers, making it a gold-standard resource in the field ([65], [66]). Before integrating all modalities, to address the initial missing data within each modality, we applied simple mean imputation [39] for each column. For more detailed data table with preprocessing steps for each modality, please refer to Appendix A.1.

MIMIC-IV Dataset We use the Medical Information Mart for Intensive Care IV (MIMIC-IV) database [27], 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. In order to align the experimental setup with the ADNI data, which does not contain temporal data, we take the last visit for each patient.

Baselines We compare the performance of `Flex-MoE` with various state-of-the-art baselines across modality-specific, e.g., image or genetic, and multimodal approaches. For the image-only modality, we first experimented with 3D MRI scans by utilizing a 3D CNN [17] and an architecture that combines 3D CNN and 3D CLSTM [68]. To decrease computational complexity, we also extracted 2D slices from the 3D volumes. For 2D MRI scans, we implemented the VGG architecture with pre-trained weights and applied layer-wise transfer learning [43], as well as a modified ResNet-18 network [45]. For the genetic-only approach, we employed a ResNet-34 based architecture to handle the high-dimensional genetic data [36]. In ADNI dataset, we further implemented domain speicfic baselines, such as auto-encoder and 3D CNN-based architecture that incorporates imaging, genetic, and clinical data [59], and a GRU-based architecture that considers imaging, genetic, clinical, and biospecimen data [33]. Moreover, we include ShaeSpec [60], which utilizes a spectral attention mechanism to emphasize important features across modalities, and mmFormer [76], which is based on transformer-based multimodal fusion with an attention mechanism. For multimodal approaches in both ADNI and MIMIC-IV, we incorporate the recent FuseMOE [19] model, which directly integrates multimodal data through a mixture of experts strategy, as the most straightforward baseline. Additionally, we compare the following methods: MulT [57], which captures cross-modal interactions through cross-attention mechanisms; MAG [52], which fuses multimodal features by mapping them to an adaptation vector; TF [73], which combines multimodal embedding sub-networks and a tensor fusion layer; and LIMoE [44], which addresses training stability in multimodal learning using entropy regularization based on contrastive learning.

Experimental Settings. To ensure a fair comparison with other baselines, we used the best hyper-parameter settings provided in the original papers. If not available, we tuned the learning rate in 1e-3, 1e-4, 1e-5, the hidden dimension in 64, 128, 256, and the batch size in 8, 16. For our proposed method, we searched the number of experts in 16, 32, and Top-k in 2, 3, 4. We set the coefficient of the sum of additional losses (importance and load balancing) combined with our cross-entropy loss to 0.01, scaling it within the task classification loss. For the dataset split, we chose 70% for training, with the remaining 30% split evenly between validation and test sets (15% each). It is important to note that, to share the same inference space, where single and multimodal baselines should both be able to predict, we opted to choose the intersection as the test and validation sets. This means that during the training phase, the dataset can be incomplete. For the multi-modal baselines, if they had the ability to impute or interact with other modalities, we leveraged their methods. Otherwise, we used zero-padding to facilitate batch-wise training. For single-modal and multi-modal baselines

 https://doi.org/10.52202/079017-3135

that solely work on the intersection region, we filtered that data and used it during training. All experiments were conducted using NVIDIA A100 GPUs. Each experiment was run three times with different seeds to ensure reproducibility, and the results were averaged. The optimal hyperparameter settings for `Flex-MoE` can be found in Appendix A.2.

Table 1: Performance on ADNI dataset with ACC metric across different models and modality combinations, given the Image ($\mathcal{I}$,), Genetic ($\mathcal{G}$,), Clinical ($\mathcal{C}$,), and Biospecimen ($\mathcal{B}$,) modalities. $\mathcal{MC}$ denotes observed modality combination.

	Modalities				Dataset: ADNI / Metric: ACC									
$\mathcal{MC}$					[59]	[33]	ShaSpec	mmFormer	TF	MulT	MAG	LIMoE	FuseMoE	Flex-MoE
$\mathcal{I},\mathcal{G}$	•	•			54.81 ±1.45	53.59 ±2.98	48.09 ±0.66	49.85 ±4.92	59.94 ±0.40	60.32 ±0.95	59.94 ±1.00	59.29 ±0.95	60.41 ±0.87	**61.08** ±0.78
$\mathcal{I},\mathcal{C}$	•		•		44.35 ±1.99	**57.15** ±1.58	47.62 ±1.81	51.96 ±4.23	54.53 ±0.66	50.14 ±1.05	52.19 ±2.90	52.38 ±3.46	53.13 ±1.97	56.49 ±2.55
$\mathcal{I},\mathcal{B}$	•			•	40.80 ±2.94	57.61 ±1.86	50.98 ±2.09	51.45 ±3.53	52.57 ±2.06	51.17 ±2.88	52.47 ±4.11	53.87 ±2.75	49.67 ±1.97	**60.41** ±0.26
$\mathcal{G},\mathcal{C}$		•	•		51.91 ±1.39	52.85 ±2.47	52.85 ±2.65	49.58 ±4.45	38.38 ±3.03	46.03 ±5.42	40.34 ±6.11	35.76 ±6.24	38.84 ±2.42	**60.60** ±0.26
$\mathcal{G},\mathcal{B}$		•		•	45.01 ±1.30	52.66 ±3.63	58.54 ±2.97	48.45 ±4.56	42.20 ±1.78	39.40 ±2.91	40.52 ±2.52	36.88 ±5.04	37.91 ±0.80	**63.59** ±1.04
$\mathcal{C},\mathcal{B}$			•	•	44.63 ±0.92	**63.68** ±0.48	59.10 ±2.69	47.71 ±4.49	39.68 ±2.38	44.54 ±0.82	40.15 ±2.58	43.98 ±0.00	37.91 ±0.80	60.50 ±0.82
$\mathcal{I},\mathcal{G},\mathcal{C}$	•	•	•		55.12 ±2.38	54.72 ±0.28	49.30 ±3.17	46.49 ±3.57	54.06 ±1.98	60.97 ±0.95	61.34 ±0.61	53.50 ±2.25	60.97 ±1.32	**63.21** ±1.73
$\mathcal{I},\mathcal{G},\mathcal{B}$	•	•		•	56.12 ±3.44	55.28 ±3.44	52.85 ±0.53	47.15 ±6.43	54.44 ±2.26	53.03 ±1.95	54.15 ±1.06	53.97 ±1.08	52.85 ±1.00	**62.28** ±2.75
$\mathcal{I},\mathcal{C},\mathcal{B}$	•		•	•	43.79 ±0.69	60.97 ±2.60	52.85 ±3.30	47.18 ±4.68	52.29 ±1.47	49.86 ±1.50	53.24 ±0.50	54.97 ±0.00	49.67 ±1.00	**64.05** ±1.78
$\mathcal{G},\mathcal{C},\mathcal{B}$		•	•	•	45.28 ±1.85	53.87 ±3.35	62.09 ±3.27	46.38 ±4.24	43.33 ±4.43	43.32 ±6.74	37.25 ±1.99	40.99 ±2.62	34.64 ±1.95	**65.36** ±1.38
$\mathcal{I},\mathcal{G},\mathcal{C},\mathcal{B}$	•	•	•	•	52.10 ±0.99	55.64 ±1.86	52.84 ±0.53	58.92 ±6.58	57.24 ±3.05	58.82 ±0.82	61.44 ±1.61	55.18 ±4.22	59.52 ±1.00	**66.11** ±1.14

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.

	Modalities			Dataset: MIMIC-IV / Metric: ACC					
$\mathcal{MC}$				TF	MulT	MAG	LIMoE	FuseMoE	Flex-MoE
$\mathcal{L},\mathcal{N}$	•	•		60.05 ±1.96	57.96 ±7.25	62.72 ±2.36	63.80 ±1.99	60.50 ±3.82	**76.14** ±0.73
$\mathcal{L},\mathcal{C}$	•		•	64.13 ±3.39	62.47 ±2.01	60.13 ±1.97	64.89 ±1.46	63.31 ±3.21	**75.15** ±0.55
$\mathcal{N},\mathcal{C}$		•	•	60.97 ±2.36	62.23 ±2.81	59.41 ±4.15	64.27 ±4.05	64.77 ±3.05	**74.96** ±1.59
$\mathcal{L},\mathcal{N},\mathcal{C}$	•	•	•	63.11 ±2.17	64.62 ±0.44	62.87 ±2.50	61.61 ±2.37	63.90 ±1.72	**76.81** ±0.90

4.2 Primary Results

In Table 1 and Table 2, we provide a comprehensive comparison of `Flex-MoE` with various multimodal baselines. We have the following observations: **(1)** Overall, `Flex-MoE` performs effectively in diverse multimodal settings, fully harnessing its potential as more modalities become available. This is supported by the large margin of improvement (7.6% and 11.07% over the best performing baselines, MAG and the most recent work FuseMoE, respectively, in full modality settings in Table 1). **(2)** Although the recently proposed FuseMoE [19] suggested its ability to handle missing scenarios, the lack of effective modality combination creates a bottleneck in such AD domain, even performing worse when a smaller number of modalities is used (FuseMoE performs better with three modalities than with full modalities), which is not optimal given the diverse missing modality scenarios. **(3)** Despite its specific characteristics in the AD domain [33, 59], the reliance on intersection data and the lack of consideration for how missing modalities relate to observed modality combinations have been overlooked. **(4)** Overall, the performance gain derived from `Flex-MoE` can be attributed to its unique ability to cope with diverse modality combinations through a missing modality bank, and its capability to fully harness the knowledge of samples via a generalization followed by a specialization step for experts. For additional results on different metrics, such as Macro-F1 and AUC, please refer to Appendix A.3.

4.3 Effectiveness of Modality Combination Consideration

To validate the effectiveness of the two essential modules of `Flex-MoE`—the *missing modality bank* and the *unique SMoE design*—under a missing modality scenario, we evaluate them followingly.

First, to evaluate the effectiveness of the missing modality bank introduced in Figure 3 (b), we assess whether it captures relevant embedding information given an observed modality combination. Specifically, we validate this by examining the inter-relationship between modalities, focusing on

the consistency of how the missing modality bank handles missing information. In Figure 4, we measure the cosine similarity between observed modality combinations. The key observation is that **(1)** with more overlapping modality combination, it tends to share more similar embedding information. This is evident in the left side of Figure 4, where the full modality scenario (ICBG) shows higher similarity with ICB and CBG (0.56 and 0.58, respectively) compared to IC and CB (0.46 and 0.45). The clinical modality (C) is most similar across combinations, which aligns with the dataset characteristic that clinical data is present in all input combinations, as shown in Figure 2. On the right side of Figure 4, the similarity between missing modalities is shown. When **(2)** modality G is missing, it is more similar to the cases where C and B are missing compared to I, suggesting that certain missing modalities share more commonality in how they are handled by the model. This insight underscores the importance of careful consideration when modeling missing modalities and demonstrates how the missing modality bank effectively captures necessary embedding information in terms of modality combination.

Furthermore, in Figure 5, we show the activation ratio of input modality combinations across each expert index, i.e., possible modality combinations. We observe two key findings: **(1)** Thanks to expert generalization using full modality samples (BCGI), generalized knowledge is distributed across all experts. This allows each expert to leverage commonly shared knowledge while activating the most relevant inputs for the downstream task. **(2)** Expert specialization enables each expert to acquire specialized knowledge. For instance, in the case of the BCG expert, the two most activated input tokens were BCGI and its corresponding token, BCG. Similarly, for the BCI and CI experts, they not only possess general knowledge from BCGI but also hold their own specialized knowledge from BCI and CI, respectively, to effectively handle these inputs. This demonstrates that when a certain modality combination is provided, the top-1 expert is successfully selected, allowing it to supplement its specialized and necessary information. Overall, these routing experiments demonstrate that `Flex-MoE` contains both globally generalized and locally expert-specific knowledge, achieved by leveraging samples with both full and fewer modalities.

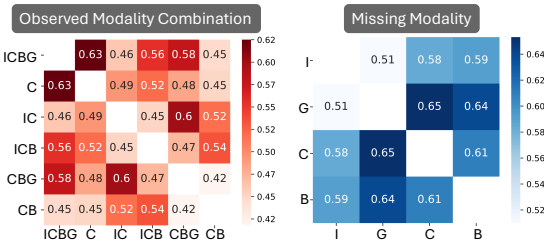

Figure 4: Cosine similarity between observed modality combination and missing modality, corresponding to row and column in missing modality bank.

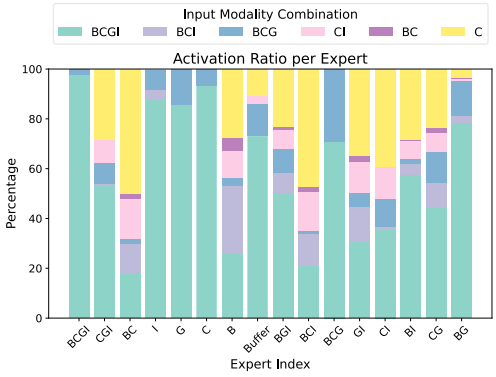

Figure 5: Modality combination activation ratio.

4.4 Comprehensive Evaluation

Ablation Study. In this section, we investigate the crucial components that contribute most positively to the performance gain of `Flex-MoE`. From Table 3, we observe that **(1)** when both expert specialization and generalization are absent, the performance drop is most severe. Additionally, **(2)** the performance decline in the embedding bank negatively affects overall performance, indicating that the missing modality bank combined with expert generalization and specialization is crucial for handling missing modality scenarios. Furthermore, (3) the sorting based on descending order appear most effective as the expert generalization occurs first withi the full modality samples.

Sensitivity Study. In Figure 6, we also varied the hyperparameters used in this study. We examined the number of experts, number of SMoE layers and top-k selection. We found that **(1)** employing many experts does not always guraantee a higher performance compared to its increase in complexity, showing using 16 experts apeear to be a suitable choice to equip fine-grained specialized knowledge. **(2)** Using sing a single layer of the SMoE was most effective, as stacking more layers or adding a Transformer block caused an overload to parameter learning. Additionally, **(3)** compared to the commonly used top-2 gating network in concurrent SMoE studies, we found that top-4 selection was

https://doi.org/10.52202/079017-3135

Table 3: Ablation study of `Flex-MoE`.

	ACC	F1
`Flex-MoE`	**66.11**	**64.73**
w/o ES	62.75	60.79
w/o {ES + EG}	62.49	60.07
w/o embedding bank	63.87	62.48
w/o sorting - random	62.65	60.70
w/o sorting - ascending	63.87	62.22

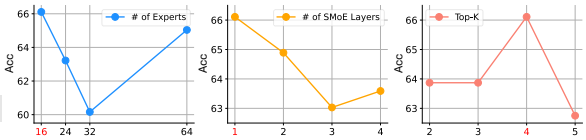

Figure 6: Sensitivity analysis of `Flex-MoE`. The hyper-parameters include the number of experts, the number of SMoE layers and Top-k expert selection. For the experiment, ADNI dataset with full modalities is used.

the most effective. This is because manually assigning the top-1 expert index to the target modality combination leaves more room for better harmonization with the SMoE design.

Complexity Study. In Table 4, we further verify the benefits of utilizing the SMoE design in terms of mean time per iteration, GFLOPs, and the number of parameters compared to the baselines. We observed the following: **(1)** Compared to recent baseline FuseMoE, `Flex-MoE` achieves notable efficiency gains (e.g., 22.74%, 1.15%, and 89.17% gain in mean time, GFLOPs, and # of parameters, respectively.) while delivering higher performance. **(2)** Although TF appears to be a lightweight design in the $\mathcal{I}, \mathcal{G}$ and $\mathcal{I}, \mathcal{G}, \mathcal{C}$ settings, it trades off computational efficiency with significantly lower performance compared to `Flex-MoE`. **(3)** Notably, as the number of modalities increases, existing models tend to become more complex in terms of GFLOPs and the number of parameters to manage the additional complexity. However, `Flex-MoE` remains robust and efficient, maintaining higher performance due to its effective use of sparsely activated experts, brought by the SMoE framework.

Modality	Metric	TF	MulT	MAG	LIMOE	FuseMoE	`Flex-MoE`
$\mathcal{I}, \mathcal{G}$	Mean Time (s) (↓)	12.40	12.85	11.64	12.65	18.68	12.73
	GFLOPs (↓)	59.05	59.24	59.06	59.24	59.74	59.06
	# of Parameters (↓)	33,370,898	37,343,683	36,454,595	37,344,707	264,680,387	36,516,807
	Accuracy (↑)	59.94 ±0.40	60.32 ±0.95	59.94 ±1.01	59.29 ±0.95	60.41 ±0.87	**61.08** ±0.78
$\mathcal{I}, \mathcal{G}, \mathcal{C}$	Mean Time (s) (↓)	13.80	23.28	14.55	14.64	18.68	14.53
	GFLOPs (↓)	59.05	59.59	59.06	59.32	59.74	59.06
	# of Parameters (↓)	34,424,162	40,185,923	36,504,643	37,960,643	340,929,475	36,685,511
	Accuracy (↑)	54.06 ±1.98	60.97 ±0.95	61.34 ±0.61	53.50 ±2.25	60.97 ±1.32	**63.21** ±1.73
$\mathcal{I}, \mathcal{G}, \mathcal{C}, \mathcal{B}$	Mean Time (s) (↓)	15.83	38.70	16.04	17.96	20.71	16.00
	GFLOPs (↓)	59.39	60.12	59.06	59.41	59.76	59.07
	# of Parameters (↓)	119,483,922	46,409,667	36,504,643	38,638,531	340,929,475	36,916,167
	Accuracy (↑)	57.24 ±0.35	58.82 ±0.82	61.44 ±1.16	55.18 ±2.42	59.52 ±1.00	**66.11** ±1.14

Table 4: Complexity comparsion of mean time, GFLOPs, and # of parameters in ADNI dataset.

5 Conclusion

While multimodal learning brings new opportunities and challenges across various domains, including medical fields, existing approaches struggle to handle arbitrary modality combinations, especially in missing modality scenarios, often relying on single modalities or complete datasets. In this work, we propose a flexible multimodal learning framework, `Flex-MoE`, capable of managing arbitrary subsets of available modalities. By carefully considering modality combination, it leverages a learnable embedding bank to capture missing modality information and utilizes a unique SMoE design to enhance expert generalization and specialization. Extensive experiments on the representative ADNI and MIMIC-IV datasets validate its effectiveness in handling diverse modality combinations. Future work includes extending the framework to explore the scaling laws of available modalities, which in turn presents numerous modality combinations, offering significant room for further improvement.

Societal Impact and Limitation: The proposed algorithm has the potential to significantly improve early diagnosis and treatment outcomes for patients, reducing the burden on healthcare systems. However, its effectiveness can be limited by the availability of comprehensive and high-quality patient data, and there may be challenges in integrating this tool into existing clinical workflows.

Acknowledgement This work is supported by RF1-AG063481, R01-AG071174, Gemma Academic Program, and OpenAI Researcher Access Program.

References

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

[2] K. Ahmed, M. H. Baig, and L. Torresani. Network of experts for large-scale image categorization. In *European Conference on Computer Vision*, pages 516–532. Springer, 2016.

[3] R. Aoki, F. Tung, and G. L. Oliveira. Heterogeneous multi-task learning with expert diversity. *CoRR*, abs/2106.10595, 2021. URL `https://arxiv.org/abs/2106.10595`.

[4] A. Association et al. 2016 alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 12 (4):459–509, 2016.

[5] J. Ávila-Jiménez, V. Cantón-Habas, M. del Pilar Carrera-González, M. Rich-Ruiz, and S. Ventura. A deep learning model for alzheimer's disease diagnosis based on patient clinical records. *Computers in Biology and Medicine*, 169:107814, 2024.

[6] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[7] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

[8] B. Cao, Y. Sun, P. Zhu, and Q. Hu. Multi-modal gated mixture of local-to-global experts for dynamic image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23555–23564, 2023.

[9] K. Chen, L. Xu, and H. Chi. Improved learning algorithms for mixture of experts in multiclass classification. *Neural networks*, 12(9):1229–1252, 1999.

[10] T. Chen, Z. Zhang, A. Jaiswal, S. Liu, and Z. Wang. Sparse moe as the new dropout: Scaling dense and self-slimmable transformers, 2023.

[11] Z. Chen, Y. Shen, M. Ding, Z. Chen, H. Zhao, E. G. Learned-Miller, and C. 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)*, pages 11828–11837, June 2023.

[12] J. M. Choi, M. Christman, and R. Sengupta. Personalized video relighting with an at-home light stage, 2024. URL `https://arxiv.org/abs/2311.08843`.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[14] J. Doshi, G. Erus, Y. Ou, S. M. Resnick, R. C. Gur, R. E. Gur, T. D. Satterthwaite, S. Furth, C. Davatzikos, A. N. 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.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. 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=YicbFdNTTy`.

[16] D. Eigen, M. Ranzato, and I. Sutskever. Learning factored representations in a deep mixture of experts. *arXiv preprint arXiv:1312.4314*, 2013.

[17] S. Esmaeilzadeh, D. I. Belivanis, K. M. Pohl, and E. 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*, pages 337–345. Springer, 2018.

 https://doi.org/10.52202/079017-3135

[18] S. Gross, M. Ranzato, and A. Szlam. Hard mixtures of experts for large scale weakly supervised vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6865–6873, 2017.

[19] X. Han, H. Nguyen, C. Harris, N. Ho, and S. Saria. Fusemoe: Mixture-of-experts transformers for fleximodal fusion. *arXiv preprint arXiv:2402.03226*, 2024.

[20] H. Hazimeh, Z. Zhao, A. Chowdhery, M. Sathiamoorthy, Y. Chen, R. Mazumder, L. Hong, and E. H. Chi. Dselect-k: Differentiable selection in the mixture of experts with applications to multi-task learning. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 29335–29347, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f5ac21cd0ef1b88e9848571aeb53551a-Abstract.html.

[21] C. R. Jack and D. M. Holtzman. Biomarker modeling of alzheimer's disease. *Neuron*, 80(6): 1347–1358, 2013.

[22] C. R. Jack Jr, D. A. Bennett, K. Blennow, M. C. Carrillo, B. Dunn, S. B. Haeberlein, D. M. Holtzman, W. Jagust, F. Jessen, J. Karlawish, et al. Nia-aa research framework: toward a biological definition of alzheimer's disease. *Alzheimer's & Dementia*, 14(4):535–562, 2018.

[23] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

[24] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mixtral of experts, 2024.

[25] H. Jiang, K. Zhan, J. Qu, Y. Wu, Z. Fei, X. Zhang, L. Chen, Z. Dou, X. Qiu, Z. Guo, et al. Towards more effective and economic sparsely-activated model. *arXiv preprint arXiv:2110.07431*, 2021.

[26] T. Jo, K. Nho, P. Bice, A. J. Saykin, and A. D. N. Initiative. Deep learning-based identification of genetic variants: application to alzheimer's disease classification. *Briefings in Bioinformatics*, 23(2):bbac022, 2022.

[27] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, and et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6 (317):1–8, 2019. doi: 10.1038/s41597-019-0322-0. URL https://doi.org/10.1038/s41597-019-0322-0.

[28] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994.

[29] M. Kim, S. Huh, H. J. Park, S. H. Cho, M.-Y. Lee, S. Jo, and Y. S. Jung. Surface-functionalized sers platform for deep learning-assisted diagnosis of alzheimer's disease. *Biosensors and Bioelectronics*, page 116128, 2024.

[30] Y. J. Kim, A. A. Awan, A. Muzio, A. F. Cruz-Salinas, L. Lu, A. Hendy, S. Rajbhandari, Y. He, and H. H. Awadalla. Scalable and efficient moe training for multitask multilingual models. *CoRR*, abs/2109.10465, 2021. URL https://arxiv.org/abs/2109.10465.

[31] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks, 2017. URL https://arxiv.org/abs/1609.02907.

[32] S. Kudugunta, Y. Huang, A. Bapna, M. Krikun, D. Lepikhin, M. Luong, and O. Firat. Beyond distillation: Task-level mixture-of-experts for efficient inference. In M. Moens, X. Huang, L. Specia, and S. W. Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 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.

[33] G. Lee, K. Nho, B. Kang, K.-A. Sohn, and D. Kim. Predicting alzheimer's disease progression using multi-modal deep learning approach. *Scientific reports*, 9(1):1952, 2019.

[34] J. Lee, S. Yun, Y. Kim, T. Chen, M. Kellis, and C. Park. Single-cell rna sequencing data imputation using bi-level feature propagation. *Briefings in Bioinformatics*, 25(3):bbae209, 2024.

[35] D. Lepikhin, H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, and Z. 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.

[36] L. Li, Y. Huang, Y. Han, and J. Jiang. Use of deep learning genomics to discriminate alzheimer's disease and healthy controls. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 5788–5791. IEEE, 2021.

[37] P. P. Liang, Y. Lyu, X. Fan, Z. Wu, Y. Cheng, J. Wu, L. Chen, P. Wu, M. A. Lee, Y. Zhu, R. Salakhutdinov, and L. Morency. Multibench: Multiscale benchmarks for multimodal representation learning. In *NeurIPS Datasets and Benchmarks*, 2021.

[38] B. Lin, Z. Tang, Y. Ye, J. Cui, B. Zhu, P. Jin, J. Huang, J. Zhang, M. Ning, and L. Yuan. Moe-llava: Mixture of experts for large vision-language models, 2024.

[39] R. J. Little and D. B. Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

[40] Y. Lou, F. Xue, Z. Zheng, and Y. You. Cross-token modeling with conditional computation. *arXiv preprint arXiv:2109.02008*, 2021.

[41] J. Ma, Z. Zhao, X. Yi, J. Chen, L. Hong, and E. H. Chi. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In Y. Guo and F. Farooq, editors, *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pages 1930–1939. ACM, 2018. doi: 10.1145/3219819.3220007. URL https://doi.org/10.1145/3219819.3220007.

[42] F. Márquez and M. A. Yassa. Neuroimaging biomarkers for alzheimer's disease. *Molecular neurodegeneration*, 14(1):21, 2019.

[43] A. Mehmood, S. Yang, Z. Feng, M. Wang, A. S. Ahmad, R. Khan, M. Maqsood, and M. Yaqub. A transfer learning approach for early diagnosis of alzheimer's disease on mri images. *Neuroscience*, 460:43–52, 2021.

[44] B. Mustafa, C. Riquelme, J. Puigcerver, R. Jenatton, and N. 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.

[45] M. Odusami, R. Maskeliūnas, R. Damaševičius, and T. Krilavičius. Analysis of features of alzheimer's disease: Detection of early stage from functional brain changes in magnetic resonance images using a finetuned resnet18 network. *Diagnostics*, 11(6):1071, 2021.

[46] M. Odusami, R. Maskeliūnas, R. Damaševičius, and S. Misra. Machine learning with multimodal neuroimaging data to classify stages of alzheimer's disease: a systematic review and meta-analysis. *Cognitive Neurodynamics*, pages 1–20, 2023.

[47] Y. Ou, A. Sotiras, N. Paragios, and C. Davatzikos. Dramms: Deformable registration via attribute matching and mutual-saliency weighting. *Medical image analysis*, 15(4):622–639, 2011.

[48] A. Papassotiropoulos, M. Fountoulakis, T. Dunckley, D. A. Stephan, and E. M. Reiman. Genetics, transcriptomics, and proteomics of alzheimer's disease. *Journal of Clinical Psychiatry*, 67 (4):652, 2006.

[49] S. Pavlitskaya, C. Hubschneider, M. Weber, R. Moritz, F. Huger, P. Schlicht, and M. Zollner. Using mixture of expert models to gain insights into semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 342–343, 2020.

 https://doi.org/10.52202/079017-3135

[50] J. Peng, K. Zhou, R. Zhou, T. Hartvigsen, Y. Zhang, Z. Wang, and T. Chen. Sparse moe as a new treatment: Addressing forgetting, fitting, learning issues in multi-modal multi-task learning, 2024. URL https://openreview.net/forum?id=bIHyMpzeuI.

[51] R. C. Petersen, P. S. Aisen, L. A. Beckett, M. C. Donohue, A. C. Gamst, D. J. Harvey, C. Jack Jr, W. J. Jagust, L. M. Shaw, A. W. Toga, et al. Alzheimer's disease neuroimaging initiative (adni) clinical characterization. *Neurology*, 74(3):201–209, 2010.

[52] W. Rahman, M. K. Hasan, S. Lee, A. Bagher Zadeh, C. Mao, L.-P. Morency, and E. Hoque. Integrating multimodal information in large pretrained transformers. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2359–2369, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.214. URL https://aclanthology.org/2020.acl-main.214.

[53] W. Rahman, M. K. Hasan, S. Lee, A. Zadeh, C. Mao, L.-P. Morency, and E. Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 2359. NIH Public Access, 2020.

[54] C. Riquelme, J. Puigcerver, B. Mustafa, M. Neumann, R. Jenatton, A. S. Pinto, D. Keysers, and N. Houlsby. Scaling vision with sparse mixture of experts. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 8583–8595, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/48237d9f2dea8c74c2a72126cf63d933-Abstract.html.

[55] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[56] X. Tang, J. Zhang, Y. He, X. Zhang, Z. Lin, S. Partarrieu, E. B. Hanna, Z. Ren, H. Shen, Y. Yang, et al. Explainable multi-task learning for multi-modality biological data analysis. *Nature Communications*, 14(1):2546, 2023.

[57] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In A. Korhonen, D. Traum, and L. Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1656. URL https://aclanthology.org/P19-1656.

[58] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019.

[59] J. Venugopalan, L. Tong, H. R. Hassanzadeh, and M. D. Wang. Multimodal deep learning models for early detection of alzheimer's disease stage. *Scientific reports*, 11(1):3254, 2021.

[60] H. Wang, Y. Chen, C. Ma, J. Avery, L. Hull, and G. Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling, 2024. URL https://arxiv.org/abs/2307.14126.

[61] J. Wang, A. Ma, Y. Chang, J. Gong, Y. Jiang, R. Qi, C. Wang, H. Fu, Q. Ma, and D. Xu. scgnn is a novel graph neural network framework for single-cell rna-seq analyses. *Nature communications*, 12(1):1882, 2021.

[62] X. Wang, F. Yu, L. Dunlap, Y.-A. Ma, R. Wang, A. Mirhoseini, T. Darrell, and J. E. Gonzalez. Deep mixture of experts via shallow embedding. In *Uncertainty in artificial intelligence*, pages 552–562. PMLR, 2020.

[63] X. Wang, Y. Chen, and W. Zhu. A survey on curriculum learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):4555–4576, 2021.

[64] M. W. Weiner, P. S. Aisen, C. R. Jack Jr, W. J. Jagust, J. Q. Trojanowski, L. Shaw, A. J. Saykin, J. C. Morris, N. Cairns, L. A. Beckett, et al. The alzheimer's disease neuroimaging initiative: progress report and future plans. *Alzheimer's & Dementia*, 6(3):202–211, 2010.

[65] M. W. Weiner, D. P. Veitch, P. S. Aisen, L. A. Beckett, N. J. Cairns, R. C. Green, D. Harvey, C. R. Jack, W. Jagust, E. Liu, et al. The alzheimer's disease neuroimaging initiative: a review of papers published since its inception. *Alzheimer's & Dementia*, 9(5):e111–e194, 2013.

[66] M. W. Weiner, D. P. Veitch, P. S. Aisen, L. A. Beckett, N. J. Cairns, J. Cedarbaum, M. C. Donohue, R. C. Green, D. Harvey, C. R. Jack Jr, et al. Impact of the alzheimer's disease neuroimaging initiative, 2004 to 2014. *Alzheimer's & Dementia*, 11(7):865–884, 2015.

[67] M. W. Weiner, D. P. Veitch, P. S. Aisen, L. A. Beckett, N. J. Cairns, R. C. Green, D. Harvey, C. R. Jack Jr, W. Jagust, J. 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.

[68] Z. Xia, G. Yue, Y. Xu, C. Feng, M. Yang, T. Wang, and B. Lei. A novel end-to-end hybrid network for alzheimer's disease detection using 3d cnn and 3d clstm. In *2020 IEEE 17th international symposium on biomedical imaging (ISBI)*, pages 1–4. IEEE, 2020.

[69] B. Yang, G. Bender, Q. V. Le, and J. Ngiam. Condconv: Conditionally parameterized convolutions for efficient inference. *Advances in Neural Information Processing Systems*, 32, 2019.

[70] S. E. Yuksel, J. N. Wilson, and P. 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.

[71] S. Yun, K. Kim, K. Yoon, and C. Park. Lte4g: Long-tail experts for graph neural networks. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, CIKM '22. ACM, Oct. 2022. doi: 10.1145/3511808.3557381. URL `http://dx.doi.org/10.1145/3511808.3557381`.

[72] S. Yun, J. Peng, A. E. Trevino, C. Park, and T. Chen. Mew: Multiplexed immunofluorescence image analysis through an efficient multiplex network, 2024. URL `https://arxiv.org/abs/2407.17857`.

[73] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency. Tensor fusion network for multimodal sentiment analysis. In M. Palmer, R. Hwa, and S. Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1115. URL `https://aclanthology.org/D17-1115`.

[74] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017.

[75] Y. Zhang and D.-Y. Yeung. Multi-task learning in heterogeneous feature spaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pages 574–579, 2011.

[76] Y. Zhang, N. He, J. Yang, Y. Li, D. Wei, Y. Huang, Y. Zhang, Z. He, and Y. Zheng. mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation, 2022. URL `https://arxiv.org/abs/2206.02425`.

[77] Z. Zhang, Y. Lin, Z. Liu, P. Li, M. Sun, and J. Zhou. Moefication: Conditional computation of transformer models for efficient inference. *arXiv preprint arXiv:2110.01786*, 2021.

[78] Y. Zhou, T. Lei, H. Liu, N. Du, Y. Huang, V. Zhao, A. Dai, Z. Chen, Q. Le, and J. Laudon. Mixture-of-experts with expert choice routing. *arXiv preprint arXiv:2202.09368*, 2022.

[79] S. Zuo, X. Liu, J. Jiao, Y. J. Kim, H. Hassan, R. Zhang, J. Gao, and T. Zhao. Taming sparsely activated transformer with stochastic experts. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=B72HXs80q4`.

 https://doi.org/10.52202/079017-3135

A Appendix

A.1 Detailed Data Preprocessing in ADNI

	File Name	Data Shape	Column Examples	Missing Rate (%)
Image	Processed T1-weighted fMRI	(10853, 91, 109, 91)	N/A	N/A
Genomics	ADNI_cluster_01_forward_757LONI	(757, 567386)	rs121434621', 'GSA-rs116587930'	4.45
	ADNI_GO_2_Forward_Bin	(432, 567386)	rs3131972', 'rs386134241'	0.17
	ADNI_GO2_GWAS_2nd_orig_BIN	(361, 567386)	rs182004761', 'rs386134241'	0.29
	ADNI3_PLINK_Final	(327, 567386)	rs3131972', 'rs3937033'	0.62
	ADNI3_PLINK_FINAL_2nd	(328, 567386)	200610-37', 'rs386134241'	0.46
Clinical	MEDHIST_09May2024.csv	(3083,40)	MHSOURCE', 'MHPSYCH', 'MH2NEURL', 'MH3HEAD'	13.28
	NEUROEXM_09May2024.csv	(3873,28)	NXTREMOR', 'NXCONSCI','NXNERVE', 'NXMOTOR'	13.71
	PTDEMOG_09May2024.csv	(5073,77)	PTWORK', 'PTNOTRT','PTRTYR', 'PTHOME'	66.58
	RECCMEDS_09May2024.csv	(66622,29)	'CMROUTE', 'CMREASON', 'CMEVNUM', 'CMBGN'	35.46
	VITALS_09May2024.csv	(15381,26)	VSHEIGHT', 'VSHTUNIT', 'VSBPSYS', 'VSBPDIA'	20.98
Biospecimen	APOERES_09May2024.csv	(2737, 17)	APTESTDT', 'APGEN1','APGEN2', 'APVOLUME'	39.8
	UPENNBIOMK_ROCHE_ELECSYS_09May2024.csv	(3174, 12)	ABETA40', 'ABETA42', 'TAU', 'PTAU'	13.22

Table 5: Summary of data files with their respective shapes, column examples, and missing rates.

Image Modality We first performed magnetic field intensity inhomogeneity correction to ensure that the MRI images are uniform and reliable. Then, we used a method called MUSE (Multiatlas Region Segmentation Utilizing Ensembles of Registration Algorithms and Parameters) to segment the gray matter tissue, which was the focus of our study [14]. This process involves using multiple atlases and selecting the most optimal ones to accurately extract region-of-interest values from the segmented gray matter tissue maps. Next, voxelwise regional volumetric maps for each tissue volume were generated by spatially aligning skull-stripped images to a template residing in the Montreal Neurological Institute (MNI) space using a registration method [47].

Genetic Modality We collected SNP (single nucleotide polymorphisms) data from the ADNI 1, GO/2, and 3 studies. First, SNP data from the different phases of these studies were aligned to the same reference build using Liftover `https://liftover.broadinstitute.org/`. Specifically, all SNP data were converted to NCBI build 38 (UCSC hg38). After liftover, we merged the studies into a unified dataset. Next, linkage disequilibrium (LD) pruning, with parameters (50, 5, 0.1), was applied to filter out SNPs that were highly correlated with others. Here, the parameters represent the window size (50), the step size (5), and the r-squared threshold (0.1). The SNP data contains values of $0, 1, 2$.

Biospecimen Modality We extract biospecimen data from the following csv files provided by ADNI. The file UPENNBIOMK_ROCHE_ELECSYS_09May2024.csv was used for Total Tau and Phosphorylated Tau data. The file APOERES_09May2024.csv was used for ApoE genotype data. For numerical data, we applied a MinMax scaler to scale the values to a range of -1 to 1. For categorical data, we used one-hot encoding. For the missing values, we imputed the mean value for numerical columns and the mode for categorical columns.

Clinical Modality We extracted clinical data from the following csv files provided by ADNI: MEDHIST_09May2024.csv, NEUROEXM_09May2024.csv, PTDEMOG_09May2024.csv, RECCMEDS_09May2024.csv, VITALS_09May2024.csv. During the preprocessing of this clinical data, we excluded the columns 'PTCOGBEG,' 'PTADDX,' and 'PTADBEG' as they contain information directly related to Alzheimer's Disease diagnosis. For numerical data, we applied a MinMax scaler to scale the values to a range of -1 to 1. For categorical data, we used one-hot encoding. For the missing values, we imputed the mean value for numerical columns and the mode for categorical columns.

A.2 Hyperparameter Setting

Table 6: The hyperparameter setup for `Flex-MoE`.

	ADNI	MIMIC-IV
	$\mathcal{I},\mathcal{G},\mathcal{C},\mathcal{B}$	$\mathcal{L},\mathcal{N},\mathcal{C}$
Learning rate	0.0001	0.0001
# of Experts	16	32
# of SMoE layers	1	1
Top-K	4	3
Training Epochs	50	50
Warm-up Epochs	5	5
Hidden dimension	128	128
Batch Size	8	8
# of Attention Heads	4	4

A.3 More Primary Results

Table 7: Performance on ADNI dataset with Macro F1 metric across different models and modality combinations, given the Image ($\mathcal{I}$,), Genetic ($\mathcal{G}$,), Clinical ($\mathcal{C}$,), and Biospecimen ($\mathcal{B}$,) modalities. $\mathcal{MC}$ denotes observed modality combination.

	Modalities				Dataset: ADNI / Metric: Macro F1									
$\mathcal{MC}$					[59]	[33]	ShaSpec	mmFormer	TF	MulT	MAG	LIMoE	FuseMoE	`Flex-MoE`
$\mathcal{I},\mathcal{G}$	•	•			52.92 ±0.17	52.75 ±3.91	45.86 ±1.55	40.25 ±7.03	60.03 ±0.88	60.59 ±1.05	**61.06** ±0.79	58.75 ±0.83	61.04 ±0.95	61.05 ±1.03
$\mathcal{I},\mathcal{C}$	•		•		25.65 ±6.58	**57.36** ±1.45	47.68 ±0.80	44.31 ±8.97	54.45 ±0.90	51.88 ±1.30	52.72 ±3.39	51.84 ±0.96	53.32 ±1.47	54.16 ±2.01
$\mathcal{I},\mathcal{B}$	•			•	28.78 ±1.32	57.93 ±2.04	49.82 ±2.00	45.82 ±9.33	51.20 ±2.64	52.64 ±2.57	53.36 ±2.96	52.70 ±3.47	50.38 ±1.31	**58.50** ±0.94
$\mathcal{G},\mathcal{C}$		•	•		50.54 ±1.38	51.62 ±3.34	50.29 ±0.36	39.45 ±5.47	27.85 ±1.89	36.77 ±6.42	29.33 ±0.46	29.57 ±3.65	27.49 ±2.87	**59.44** ±0.49
$\mathcal{G},\mathcal{B}$		•		•	31.21 ±1.71	51.41 ±4.30	56.32 ±4.83	35.63 ±1.44	29.82 ±1.48	32.41 ±1.55	28.29 ±1.05	20.91 ±0.60	20.91 ±0.60	**61.65** ±1.71
$\mathcal{C},\mathcal{B}$			•	•	24.78 ±6.24	**63.46** ±0.35	55.46 ±4.06	33.09 ±5.43	29.57 ±1.99	33.22 ±0.72	27.20 ±3.53	20.36 ±0.00	29.11 ±3.83	59.13 ±1.75
$\mathcal{I},\mathcal{G},\mathcal{C}$	•	•	•		48.59 ±5.44	53.86 ±3.84	48.99 ±2.59	26.88 ±9.21	54.31 ±0.88	61.82 ±0.21	61.07 ±1.04	51.33 ±1.38	61.30 ±1.07	**61.98** ±1.04
$\mathcal{I},\mathcal{G},\mathcal{B}$	•	•		•	55.46 ±2.05	54.82 ±4.18	51.80 ±0.99	28.19 ±8.29	53.70 ±2.43	52.46 ±1.64	52.54 ±1.64	52.80 ±1.92	52.75 ±0.91	**59.45** ±3.14
$\mathcal{I},\mathcal{C},\mathcal{B}$	•		•	•	25.53 ±4.83	61.34 ±2.48	51.80 ±0.99	27.87 ±9.08	52.64 ±1.49	50.14 ±0.84	52.02 ±2.22	52.41 ±1.31	50.61 ±0.47	**61.60** ±1.46
$\mathcal{G},\mathcal{C},\mathcal{B}$		•	•	•	24.95 ±6.49	60.57 ±2.64	60.57 ±2.64	25.99 ±9.98	40.40 ±6.88	29.38 ±0.79	31.59 ±1.93	27.99 ±2.13	27.49 ±0.93	**64.15** ±1.69
$\mathcal{I},\mathcal{G},\mathcal{C},\mathcal{B}$	•	•	•	•	49.76 ±1.95	57.93 ±2.04	51.80 ±0.99	53.64 ±9.09	57.27 ±0.44	59.58 ±0.77	61.38 ±1.32	53.63 ±0.30	59.55 ±1.60	**64.73** ±2.01

Table 8: Performance on ADNI dataset with AUC metric across different models and modality combinations, given the Image ($\mathcal{I}$,), Genetic ($\mathcal{G}$,), Clinical ($\mathcal{C}$,), and Biospecimen ($\mathcal{B}$,) modalities. $\mathcal{MC}$ denotes observed modality combination.

	Modalities				Dataset: ADNI / Metric: AUC									
$\mathcal{MC}$					[59]	[33]	ShaSpec	mmFormer	TF	MulT	MAG	LIMoE	FuseMoE	`Flex-MoE`
$\mathcal{I},\mathcal{G}$	•	•			70.04 ±0.72	70.25 ±3.26	66.07 ±1.11	68.05 ±2.04	73.45 ±1.06	70.95 ±1.76	73.14 ±0.71	71.88 ±1.14	72.37 ±1.08	**74.52** ±1.81
$\mathcal{I},\mathcal{C}$	•		•		54.52 ±2.93	**73.99** ±1.02	65.39 ±0.82	68.15 ±2.09	72.88 ±0.81	71.37 ±1.65	71.68 ±1.90	71.86 ±1.27	70.98 ±0.24	73.03 ±0.14
$\mathcal{I},\mathcal{B}$	•			•	57.02 ±8.74	76.06 ±0.85	68.86 ±0.69	68.44 ±1.98	71.70 ±0.42	72.43 ±1.84	72.82 ±1.84	71.82 ±1.65	70.59 ±1.09	**77.68** ±0.33
$\mathcal{G},\mathcal{C}$		•	•		69.24 ±0.45	66.46 ±3.29	71.38 ±0.82	65.50 ±5.45	47.57 ±1.96	61.17 ±5.61	52.00 ±1.08	51.14 ±0.60	49.23 ±1.54	**78.34** ±0.47
$\mathcal{G},\mathcal{B}$		•		•	48.91 ±5.97	69.68 ±3.58	75.29 ±2.91	64.69 ±5.26	51.32 ±2.33	53.53 ±0.68	51.36 ±1.07	51.82 ±0.30	51.82 ±0.30	**79.24** ±0.79
$\mathcal{C},\mathcal{B}$			•	•	58.41 ±5.16	79.53 ±0.34	78.73 ±0.22	63.25 ±5.89	49.82 ±2.03	64.36 ±2.80	50.20 ±1.63	48.29 ±3.21	48.82 ±0.58	**79.65** ±0.81
$\mathcal{I},\mathcal{G},\mathcal{C}$	•	•	•		69.07 ±3.39	76.05 ±0.86	76.05 ±0.86	66.70 ±4.15	66.35 ±0.86	74.24 ±0.62	71.87 ±0.84	72.77 ±1.21	70.98 ±1.06	71.14 ±0.83	**79.55** ±1.69
$\mathcal{I},\mathcal{G},\mathcal{B}$	•	•		•	70.75 ±2.30	76.02 ±0.86	69.79 ±1.39	65.91 ±2.01	72.11 ±2.08	71.88 ±0.34	72.35 ±0.32	71.70 ±0.81	72.16 ±0.57	**79.27** ±0.65
$\mathcal{I},\mathcal{C},\mathcal{B}$	•		•	•	52.98 ±5.16	76.06 ±0.85	69.79 ±1.39	66.09 ±1.67	72.04 ±0.92	71.32 ±0.16	71.97 ±1.76	72.25 ±0.65	71.20 ±1.33	**80.55** ±1.26
$\mathcal{G},\mathcal{C},\mathcal{B}$		•	•	•	56.71 ±4.72	70.06 ±0.85	79.18 ±0.63	63.49 ±4.78	69.00 ±0.66	60.85 ±5.25	51.73 ±0.41	49.36 ±0.52	48.82 ±1.02	**81.67** ±0.59
$\mathcal{I},\mathcal{G},\mathcal{C},\mathcal{B}$	•	•	•	•	69.44 ±0.84	76.06 ±0.85	69.79 ±1.39	73.93 ±5.97	69.42 ±3.20	71.09 ±0.66	71.99 ±0.54	72.05 ±0.27	71.16 ±1.01	**81.67** ±0.54

Table 9: Performance on MIMIC-IV dataset with Macro F1 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.

	Modalities			Dataset: MIMIC-IV / Metric: Macro F1					
$\mathcal{MC}$				TF	MulT	MAG	LIMoE	FuseMoE	`Flex-MoE`
$\mathcal{L},\mathcal{N}$	•	•		50.81 ±0.47	50.61 ±2.77	**53.46** ±0.26	55.19 ±1.52	52.79 ±1.32	51.29 ±1.83
$\mathcal{L},\mathcal{C}$	•		•	55.09 ±1.29	**56.33** ±1.00	54.07 ±0.98	57.32 ±0.52	54.78 ±0.91	53.85 ±1.43
$\mathcal{N},\mathcal{C}$		•	•	54.37 ±0.41	55.33 ±1.04	54.15 ±1.79	54.59 ±0.65	**55.54** ±0.60	53.02 ±3.99
$\mathcal{L},\mathcal{N},\mathcal{C}$	•	•	•	54.19 ±0.38	**58.43** ±0.22	55.04 ±1.41	55.79 ±0.94	55.38 ±1.06	53.19 ±1.28

Table 10: Performance on MIMIC-IV dataset with AUC 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.

	Modalities			Dataset: MIMIC-IV / Metric: AUC					
$\mathcal{MC}$				TF	MulT	MAG	LIMoE	FuseMoE	`Flex-MoE`
$\mathcal{L},\mathcal{N}$	•	•		56.31 ±1.00	57.10 ±0.78	58.11 ±0.83	62.64 ±1.81	58.33 ±0.36	**64.39** ±0.28
$\mathcal{L},\mathcal{C}$	•		•	60.43 ±0.76	65.12 ±2.19	60.75 ±0.20	65.14 ±0.34	62.52 ±0.39	**66.27** ±0.17
$\mathcal{N},\mathcal{C}$		•	•	61.37 ±1.33	62.18 ±0.77	61.99 ±0.25	61.34 ±0.41	61.74 ±0.31	**64.27** ±0.87
$\mathcal{L},\mathcal{N},\mathcal{C}$	•	•	•	60.66 ±0.65	67.35 ±0.18	61.29 ±0.32	65.18 ±0.60	61.67 ±0.15	**69.87** ±0.81

https://doi.org/10.52202/079017-3135

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clearly state the challenges of the problem at hand and the contributions that we make in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations and future directions of the work in the Conclusion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: Our work does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide details about the experiment implementation in Section 4.1 Experimental Setting. We provide further details about the data processing for the experiments in the Appendix. Thus, it is possible to replicate the main experimental results of the paper. We also provide the anonymized code.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

https://doi.org/10.52202/079017-3135

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The data is available online `https://adni.loni.usc.edu/about/`. The paper also provides access to the code, which is reproducible.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All details about the experimental setting, including data splits, hyperparameters etc., are included in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: For each experiment, we report the mean accuracy and F1 score, as well as the corresponding standard deviation values in Section 4. Thus, we provide appropriate information about the statistical significance.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide sufficient information on the compute resources necessary to reproduce the experiments in Section 4.1 Experimental Settings.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: All content in this paper abide by the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss Societal Impact and Limitation as a subsection under Section 5 Conclusion.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

https://doi.org/10.52202/079017-3135

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work does not use any data or models that have a high risk of misuse such as pretrained language models, image generators, or scraped datasets.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use the ADNI dataset and provide the proper citation. All materials related to the paper will adhere to the copyright policies of NeurIPS.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The details of our proposed method are outlined in Sections 3 and 4. Additionally, we release the source code of our work.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We do not include crowdsourcing experiments and research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: IRB approval was not necessary for this project.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

https://doi.org/10.52202/079017-3135

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.