Rethinking Medical High-Modality Learning under Missingness —A Long-Tailed Distribution Perspective

Chenwei Wu*, Zitao Shuai*, Liyue Shen
University of Michigan, Ann Arbor
liyues@umich.edu

Abstract

Medical multi-modal learning is critical for integrating information from a large set of diverse modalities. However, when leveraging a high number of modalities in real clinical applications, it is often impractical to obtain full-modality observations for every patient due to data collection constraints, a problem we refer to as 'High-Modality Learning under Missingness'. In this study, we identify that such missingness inherently induces an exponential growth in possible modality combinations, followed by long-tail distributions of modality combinations due to varying modality availability. While prior work overlooked this critical phenomenon, we find this long-tailed distribution leads to significant underperformance on tail modality combination groups. Our empirical analysis attributes this problem to two fundamental issues: 1) gradient inconsistency, where tail groups' gradient updates diverge from the overall optimization direction; 2) concept shifts, where each modality combination requires distinct fusion functions. To address these challenges, we propose ***REMIND****, a unified framework that* ***RE****thinks* ***M****ult****I****modal learN****i****ng under high-mo****D****ality missingness from a long-tail perspective. Our core idea is to propose a novel group-specialized Mixture-of-Experts architecture that scalably learns group-specific multi-modal fusion functions for arbitrary modality combinations, while simultaneously leveraging a group distributionally robust optimization strategy to upweight underrepresented modality combinations. Extensive experiments on real-world medical datasets show that our framework consistently outperforms state-of-the-art methods, and robustly generalizes across various medical multi-modal learning applications under high-modality missingness.*

1. Introduction

Multi-modal learning [49] has emerged as a significant problem for numerous medical applications [1, 16], where integrating diverse data modalities is critical for comprehensive patient assessment and clinical decision-making. These multi-modal applications often operate in a *high-modality learning* regime [19], where the number and diversity of modalities per patient can be particularly high, ranging from medical imaging, clinical notes, to laboratory measurements. In real-world clinical scenarios, it is hard to guarantee fully observed high-modality data for each patient sample due to practical constraints in data collection protocol, such as high financial costs and radiation exposure (*e.g.*, PET scans), patient discomfort from invasive procedures (*e.g.*, biopsy), or technical failures [46]. Consequently, practical multi-modal medical datasets are usually incomplete, characterized by numerous random missing modalities and arbitrary modality combinations [42].

High-modality settings under such missingness often result in long-tailed distributions of modality combinations (MCs), driven by the exponential increase in possible combinations as the number of modalities grows, alongside significant variability in the availability of individual modalities. Take the Fundus Photography, Psychological Assessment, Retina Characteristics, and Multi-modal Imaging (FPRM) dataset [45] as an example, as shown in Figure 1(a), basic modality combinations such as Electronic Health Records (EHRs) and fundus imaging are widely available across most patients; while more complex combinations involving less accessible modalities (*e.g.*, EHR+3D scans+Fundus) occur infrequently due to the higher cost and specialized technical skills required to acquire high-quality data. Thus, the frequency distribution of modality combinations always exhibits a pronounced "long-tailed" pattern (Figure 1(b)). Despite recent research efforts to address the multi-modal missingness through both imputation-based approaches, such as learnable modality embedding banks [9, 42], and imputation-free techniques like cross-modal knowledge distillation [38], existing methods largely overlook this critical imbalance of modality combination groups.

In this work, we revisit multi-modal learning in high-modality settings with significant missingness, through the

*Equal contribution.

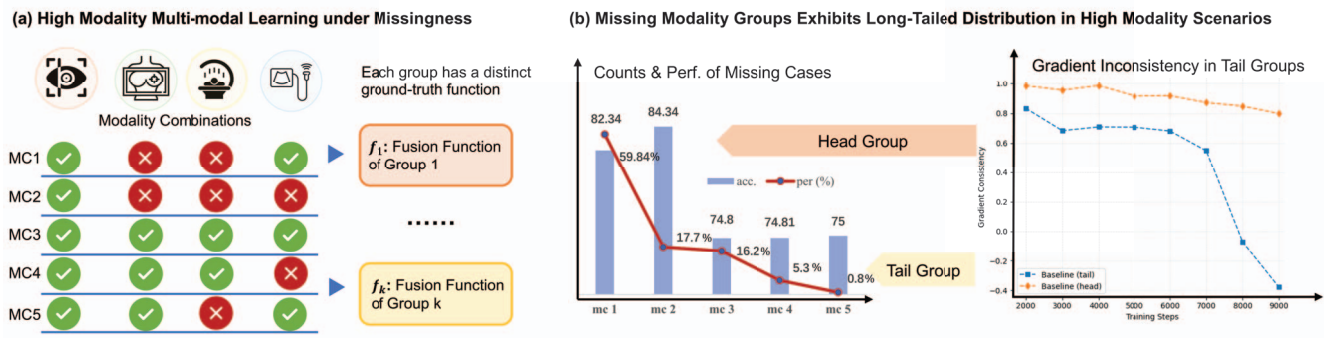

Figure 1. **Missing modalities in high-modality multi-modal learning – A long-tailed distribution view.**

lens of long-tailed distribution modeling. We conduct experiments on several datasets using the state-of-the-art methods [9, 39, 42] for high-modality learning, and observe that the model performance on long-tailed modality combinations is often inferior to that on head groups, as shown in Figure 1(b) and Sec. 5. This suggests that the model is not adequately trained to perform effective fusion involving less accessible modality combinations. To unveil the underlying mechanisms, we hypothesize that this is due to the inconsistent optimization of each group's performance, and further analyze the gradient dynamics of tail groups *v.s.* head groups. As shown in Figure 1(b), we observed a phenomenon of ***gradient inconsistency***, where long-tailed groups' gradients diverge from the overall optimization direction, which governs the model parameters' updates. This underscores the need for a robust approach to learn effective multi-modal fusion across all possible modality combinations including tail groups.

Addressing the long-tailed distribution of MCs introduces unique challenges beyond classical long-tailed classification problems extensively studied in previous research [30, 35, 43]. In traditional long-tailed problems, the data label distributions differ across classes, yet the learned mapping function from inputs to output labels remains consistent and applies to all classes. In contrast, multi-modal missingness introduces a ***concept shift*** scenario, where each MC induces a fundamentally different fusion function due to the varying availability across input modalities. Different combinations produce unique cross-modal interactions and hold distinct significance for downstream tasks [20]. For example, in ICU mortality prediction, lab tests and vital signs help capture physiological state information, but adding clinical notes may create new synergistic information by contextualizing numeric abnormalities with prior patient history, requiring a fundamentally different multi-modal fusion strategy for this new modality combination. Prior works have also shown that when certain modalities are unavailable, the models cannot easily recover the distinct modality-specific information via alignment strategies [38], and thus need to rely exclusively on signals from available sources for predictive tasks [2, 12]. Additionally, given the exponentially growing number of possible MCs in high-modality settings and severely limited samples in tail groups, training separate models independently for each combination is impractical. This challenge requires a unified framework capable of dynamically learning adaptive multi-modal fusion functions across diverse modality combinations.

To tackle the above challenges, we propose a unified multi-modal learning framework, named ***REMIND***, by ***RE***thinking ***M***ult***I***modal lear***N***ing under high-mo***D***ality missingness from a long-tailed distribution perspective. Our approach consists of two primary components. First, we employ a group Distributionally Robust Optimization (DRO) strategy to up-weight modality combinations that are typically under-optimized due to limited data availability, and thus ensure robust performance even for rare modality combinations. Second, to accommodate the inherent concept shifts caused by modality missingness, we introduce a scalable architecture building on soft Mixture-of-Experts (MoE). It operates on a shared set of expert modules across all MC groups, but learns group-specific routing matrices that dynamically determine how to combine these experts based on the available modality combination. Our contribution lies three-fold:

- We are the first to formulate multi-modal learning with high-modality missingness through the lens of long-tailed distribution modeling, revealing that existing approaches fail on tailed modality combinations due to gradient inconsistency and concept shift challenges.
- We propose a novel approach combining group distributionally robust optimization with adaptive multi-modal fusion mechanisms based on soft MoE structure, enabling learning group-specific fusion functions to effectively handle imbalanced distributions of modality combinations.
- Extensive experiments on various real-world medical multi-modal datasets with significant missing data consis-

tently demonstrate substantial improvements over state-of-the-art methods, particularly on challenging tailed modality combinations, validating the effectiveness of our approach.

2. Related Work

2.1. High-Modality Learning with Missing Data

One branch of multi-modal learning methods leverages imputation [46] to handle the missingness. For example, Chen et al. [3] trained a generative model to synthesize missing medical images to create full-modality records; Hamghalam et al. [8], Liu et al. [21], Yun et al. [42] rely on imputing the latent representations of the missing modality, with the supervision signals from prediction tasks. However, due to the complexity of reconstruction learning objectives, these methods either introduce instability and noise that adversely affect the primary tasks, or suffer from limited flexibility by requiring separate networks to be trained for each missing modality case [2]. These disadvantages become increasingly severe in the high-modality scenarios. Other imputation-free [40] approaches aim to learn unified models that robustly extract task-relevant information from a large range of modality combinations. Liang et al. [18], Liu et al. [22] apply statistical constraints to learn representation spaces that are robust to possible missing modalities. However, these methods still struggle to generalize to scenarios involving more than two modalities [48]. Recent works [39, 42] focus on learning dynamical multi-modal fusion (*e.g.* via MoE) functions or modeling the dynamic relationship across modalities and tasks [10]. However, these methods overlook the under-optimization of samples from rare modality combinations and struggle with high-modality learning under missing data, where many modality combinations have only a few samples.

2.2. Long-Tailed Modeling

Long-tail learning aims to provide high precision for tail classes that have rare occurrences [24, 35], or ensure tail groups have comparable performance with the majority groups of the dataset [23, 43]. Classical long-tail learning methods focus on reweighting-based methods like resampling [30], logits-adjustment [24], distributionally robust optimization [33], or mixing up tail groups and head groups to ensure fair optimization [4]. However, these methods share an assumption that the conditional probability $P(Y|X)$ is consistent across groups [47], which does not hold in the missing modality cases. Thus, directly adapting these methods would fail to solve the high-modality learning under missingness problem.

3. Method

3.1. Problem Formulation and Modeling

Problem Formulation. Suppose we are given a multi-modal dataset with m modalities $M_1, M_2, ..., M_m$ and a prediction task Y. In practical applications, these modalities are not always completely available across different individual samples. This potentially yields at most $2^m - 1$ missing modality cases, forming a **modality combination** family consisting of different groups $\mathcal{G} = \{\{M_1,\}, \{M_1, M_2\}, ..., \{M_1, M_2, ..., M_m\}\}$. Given such a high-modality dataset with random missingness, our goal is to develop a unified multi-modal learning framework that can accurately predict Y for any arbitrary modality combination within $\mathcal{G}$ [19].

Long-Tailed Modality Combination Groups. First of all, the targeted high-modality setting involves a significantly larger number of modalities compared to classical two-modal learning scenarios (e.g., medical image and text), inherently leading to an exponential increase in possible modality patterns. For simplicity and the purpose of illustration, given a multi-modal dataset with m modalities $M_1, M_2, \ldots, M_m$, we assume each modality M_i is independently missing with probability p_i. The probability of observing a specific combination g_k is given by:

$$\Pr(g_k) = \prod_{M_i \in g_k} (1 - p_i) \prod_{M_j \notin g_k} p_j.$$

In practice, even slight deviations from uniform missingness probabilities (i.e., $p_i \neq 0.5$) result in the majority of samples clustering into a few dominant modality combination groups, leaving the rest as rare "tail groups". Our results in Sec. 5 empirically show that directly training models on such imbalanced dataset leads to inconsistent performance across different modality combinations. To understand this, recent studies [5, 7] suggest that the parameter updates in deep networks are largely determined by the gradient direction aligned with the principal eigenvector of the Neural Tangent Kernel (NTK) matrix, which captures the dominant gradient descent direction across the input space. Motivated by this, we introduce a gradient-based analysis to understand why the performance on tail modality combination groups is significantly lower. Specifically, for each group $g_k \in \mathcal{G}$, we randomly sample an equal number of data points X_k and obtain the averaged gradient direction $GD(X_k^\theta)$ for the fusion block parameterized by θ. Details of gradient direction calculation are attached in the Appendix. For the whole dataset, we similarly compute the gradient direction $GD(X^\theta)$. The gradient consistency score between group g_k and the entire dataset is defined as the similarity between their gradient directions:

$$sim_k = \frac{GD(X^\theta) \cdot GD(X_k^\theta)}{||GD(X^\theta)||\,||GD(X_k^\theta)||}. \quad (1)$$

As shown in Fig. 1(b), for existing methods, the gradient direction of the whole dataset aligns closely with that of majority groups, maintaining high consistency scores during training. In contrast, the gradient direction for tail groups increasingly diverges from that of the whole dataset as training progresses, a phenomenon that prior research has linked to suboptimal learning outcomes [34, 41]. There is a clear correlation between the prevalence of modality combination groups and their gradient consistency scores, explaining impaired performance on tail groups. This phenomenon motivates the development of methods that effectively leverage each group $g_k \in \mathcal{G}$ during training, ensuring adequate optimization for samples from underrepresented groups.

Due to the information asymmetry [2, 12] introduced by modality missingness and the unique cross-modal interactions associated with specific modality combinations [20], each modality-combination group needs a distinct and tailored fusion strategy based on the available modalities. This requires a unified model parameterized by θ that can dynamically instantiate group-specific fusion functions $f_{g_k}(\mathbf{X}g_k; \theta) \rightarrow Y$ for each modality combination g_k. It is challenging to learn these $|\mathcal{G}|$ distinct fusion functions within a single framework while maintaining parameter efficiency across the exponentially large combination space. This motivates us to employ a Mixture-of-Experts (MoE) architecture, naturally suited for such adaptive parameter sharing and dynamic fusion learning.

3.2. Overall Distributionally Robust Framework

In this section, we will present the overall design of our proposed approach for learning a robust multi-modal model on datasets with long-tailed modality combination distributions. As demonstrated in Fig. 2(a), our framework consists of a set of modality-specific feature encoders and a MoE-based fusion block for integrating multi-modal information. The feature encoders are first jointly optimized with the fusion block for several iterations to adapt to each modality, and then they are kept frozen in subsequent model training.

Given a multi-modal input data sampled from arbitrary modality combination $g_k \in \mathcal{G}, k \in \{1, ..., 2^m - 1\}$ ($|\mathcal{G}| = 2^m - 1$ is the total number of all potentially possible modality combination groups in $\mathcal{G}$), the model first extracts embeddings $Z_i \in \mathbb{R}^{l_i \times d}, i \in \{1, ..., m\}, M_i \in g_k$, for each available modality respectively, using the modality-specific encoders. Here, d is the dimension of the embedding space, and l_i is the token length of the embedding. For missing modalities, instead of zero-padding, we follow the processing in Yun et al. [42] to create a set of learnable embeddings $B_k \in \mathbb{R}^d, k \in \{1, ..., 2^m - 1\}$ for each modality combination group respectively. For all missing modalities $M_{i'} \notin g_k$, we assign $Z_{i'} \in \mathbb{R}^{l_i \times d}$ by broadcasting B_k to represent a group-specific yet modality-agnostic learnable embedding for the missing data within this group. These two sets of embeddings form a consistent input embedding space $Z_1, ..., Z_m$ shared by all the data samples. They are then concatenated and fed into the soft MoE block to fuse the multi-modal representations H, which will be projected by the task head to predict the final task output Y.

As demonstrated in Sec. 3.1, modality combinations follow a long-tailed distribution and could cause inconsistent optimization across groups. To mitigate this imbalance issue, we build our approach upon the group distributionally robust optimization (DRO) framework [17, 29]. During training, it up-weights samples from under-optimized modality-combination groups, facilitating the model to achieve more accurate prediction especially on the worst-performing groups. Specifically, DRO assumes the overall data can be grouped by $\mathcal{G}$ and defines a set of distributions $\mathcal{Q}^{\mathcal{G}}$ based on $\mathcal{G}$. It evaluates the model on $\mathcal{Q}^{\mathcal{G}}$, so as to dynamically assign weights to sample from different modality combination groups (Fig. 2(a)). In our paper, $\mathcal{G}$ is the modality combination family. $\mathcal{Q}^{\mathcal{G}}$ is built on grouped distributions $D_k, k \in \{1, \ldots, 2^m - 1\}$ of each modality combination, and thus contains a set of potential testing domains $Q \in \mathcal{Q}^{\mathcal{G}}$. Formally, it is defined as f-divergence ball $\{Q : D_f(Q \,\|\, D^{\mathcal{G}}) \leq \rho\}$, where $D^{\mathcal{G}}$ is the distribution when the entire data distribution is grouped based on $\mathcal{G}$ [44]. Here the uncertainty radius is set to zero, because we do not consider out-of-distribution shifts in this paper. This set of potential test domains could be further written as a convex hull $\mathcal{Q}^{\mathcal{G}} = \left\{ \sum_{k=1}^{|\mathcal{G}|} \lambda_k D_k \;\middle|\; \lambda \in \Delta_{|\mathcal{G}|-1} \right\}$, where $\Delta_{|\mathcal{G}|-1}$ is the probability simplex.

DRO aims to consistently optimize the model's performance on each test domain $Q \in \mathcal{Q}^{\mathcal{G}}$, which is obtained by guaranteeing the model has a lower loss even in the worst-performing case. Motivated by Sagawa et al. [31], Zhang et al. [44], we adopt the simplex-constrained formulation to optimize on the worst-case distribution:

$$\min_{\theta} \max_{\lambda \in \Delta_{|\mathcal{G}|-1}} \sum_{k=1}^{|\mathcal{G}|} \lambda_k \, R_k(\theta), \quad R_k(\theta) = \mathbb{E}_{(x,y)\sim D_k}\big[\ell(f_\theta(x), y)\big], \tag{2}$$

where θ is the model parameter, $R_k(\theta)$ is the loss on the grouped distribution D_k. During training, we alternate between optimizing model parameters and group weights:

$$\theta^{(t+1)} = \theta^{(t)} - \eta \sum_k \lambda_k^{(t)} \nabla_\theta R_k\big(\theta^{(t)}\big), \tag{3}$$

$$\lambda_k^{(t+1)} = \frac{\lambda_k^{(t)} \exp\big(\gamma \, R_k(\theta^{(t)})\big)}{\sum_j \lambda_j^{(t)} \exp\big(\gamma \, R_j(\theta^{(t)})\big)}, \tag{4}$$

where η is the learning rate and $\gamma > 0$ is the hyper-parameter that controls the sharpness. The exponentiated update smoothly approximates $\max_k R_k(\theta)$ while keeping non-zero gradients for all groups [32]. We refresh λ every N training steps in practice.

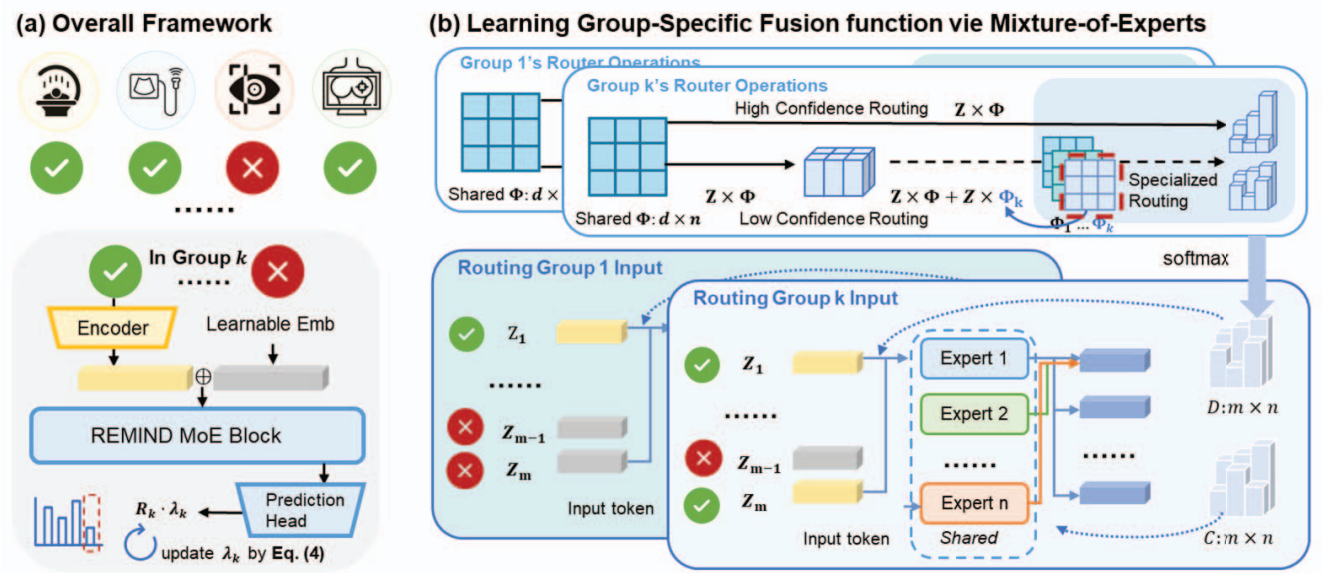

Figure 2. **Overview of REMIND.** We model high-modality learning under missing modalities.

3.3. Tackling Concept Shift with Soft MoE

As aforementioned, each modality combination group requires its own dynamic fusion function due to the inherent concept shift across different modality availability patterns. Mixture-of-Experts (MoE) architectures provide a natural foundation to satisfy this need, as they can dynamically create input-dependent routing pathways that have the potential to enable learning group-specific fusion strategies.

Soft MoE-based Multi-modal Fusion Architecture. We adopt a Soft MoE [26] model architecture consisting of three primary components: a self-attention module, an expert layer comprising n expert networks, and a learnable routing matrix $\mathbf{\Phi}$. Given the modality-specific embeddings $\{Z_1, \ldots, Z_m\} \in \mathbb{R}^{l \times d}$ with l token length for each modality in latent dimension d, they are first concatenated and projected through the self-attention module to produce $\mathbf{Z} \in \mathbb{R}^{M \times d}$, where $M = l\,m$.

The router matrix $\mathbf{\Phi} \in \mathbb{R}^{d \times n}$ generates sample-adaptive routing strategies through a softmax-based assignment, providing dispatch weights $\mathbf{D} \in \mathbb{R}^{M \times n}$ and combine weights $\mathbf{C} \in \mathbb{R}^{M \times n}$. Dispatch weights $\mathbf{D}$ are computed by column-wise softmax on $\mathbf{Z\Phi}$, specifically:

$$D_{ij} = \frac{\exp((\mathbf{Z\Phi})_{ij})}{\sum_{i'=1}^{M} \exp((\mathbf{Z\Phi})_{i'j})}, \tag{5}$$

This dispatch weights $\mathbf{D}$ determine the modality-to-experts routing function to derive expert-specific input representations $\tilde{\mathbf{Z}} = \mathbf{D}^T\mathbf{Z}$. Each expert j processes its input through a neural network f_j, yielding outputs $\tilde{\mathbf{Y}}_{.j} = f_j(\tilde{\mathbf{Z}}_{.j})$. Similarly, combine weights $\mathbf{C}$ are computed via a row-wise softmax:

$$C_{ij} = \frac{\exp((\mathbf{Z\Phi})_{ij})}{\sum_{j'=1}^{n} \exp((\mathbf{Z\Phi})_{ij'})}, \tag{6}$$

which indicates how to combine the outputs from all experts to obtain the final fused output representation for prediction head: $\mathbf{Y} = \mathbf{C}\tilde{\mathbf{Y}}$.

Group-Specific Adaptive Routing via Residual Matrices. While Soft MoE can dynamically create fusion functions, the learned routing matrix $\mathbf{\Phi}_{\text{shared}}$ shared across all groups limits the model's ability to provide group-adaptive routing strategies, particularly for tail modality combination groups that are under-optimized during the training.

To address this limitation, instead of training separate experts per group, we propose a group-specific adaptive routing approach by introducing additional residual matrices $\mathbf{\Phi}_k$. It enables scalable adaptivity to a wide range of modality combinations, by facilitating both knowledge sharing across groups through shared routing matrix $\mathbf{\Phi}_{\text{shared}}$, and the extraction of fine-grained differences via residual matrix $\mathbf{\Phi}_k$. Specifically, for modality combination group g_k, we define the final group-specific adaptive routing function as:

$$\mathbf{\Phi} = \mathbf{\Phi}_{\text{shared}} + \mathbf{\Phi}_k. \tag{7}$$

During training, the residual design of the zero-initialized group-specific routing matrices $\mathbf{\Phi}_k$ allows them to incrementally refine rather than override the shared knowledge encoded in $\mathbf{\Phi}$. This ensures controlled specialization of the routing function.

Additionally, to determine when these group-specific residual adjustments should be applied, we employ an uncertainty-based gating strategy. We use the entropy

of routing logits $\mathbf{Z}\mathbf{\Phi}_{\text{shared}}$ as one such uncertainty metric, though it can also be substituted by other measures (*e.g.*, KL-divergence). During the model inference, we use shared matrix Φ_{shared} alone, if the entropy $H(\mathbf{Z}\mathbf{\Phi}_{\text{shared}})$ is below a predefined threshold ξ, which indicates confident routing.

$$H(\mathbf{Z}\mathbf{\Phi}) = -\sum_{i,j} p_{ij} \log p_{ij}, \quad p_{ij} = \frac{\exp\left[(\mathbf{Z}\mathbf{\Phi})_{ij}\right]}{\sum_{i',j'} \exp\left[(\mathbf{Z}\mathbf{\Phi})_{i'j'}\right]}. \tag{8}$$

When the uncertainty exceeds the threshold, implying uncertain modality-to-expert assignment or near-uniform routing, which are found to be suboptimal in previous MoE literature [25, 27], the model will activate the group-specific routing adjustment Φ_k to learn a group-adaptive function. This mechanism preserves the model's confident routing patterns, while improving performance in cases where the shared routing matrix struggles.

Dispatch and combine matrices are then recomputed according to the learned adapted logits $\mathbf{Z}(\mathbf{\Phi}_{\text{shared}} + \mathbf{\Phi}_k)$. Importantly, this design is highly scalable to large numbers of modalities. Unlike previous MoE designs [9, 39, 42] that require additional modality-specific or MC-specific experts when scaling, the experts in our approach are shared across all modality combinations, and the parameter overhead is merely confined to lightweight residual routing matrices of MC groups. The choice of the threshold value and computational costs analysis are included in the Appendix.

4. Experiments

Datasets. To verify the generalizability of REMIND across high-modality multi-modal learning applications, we conduct extensive experiments over three publicly available medical multi-modal datasets for different tasks. Each dataset consists of at least three modalities with significant missing modality issues. Specifically, we use:

- EMBED (EMory BrEast Imaging Dataset) [13] contains 4 modalities, including M1: C-View synthetic Cranio-Caudal view (CC), M2: C-View synthetic 2-D Medio-Lateral Oblique view (MLO), M3: full-field digital mammography (FFDM) CC, and M4: FFDM MLO, which are used to train a breast density prediction model.
- MIMIC-IV (Medical Information Mart for Intensive Care IV Dataset) [14]. Following [9], we extract 3 modalities of M1: ICD-9 codes, M2: clinical text, and M3: labs' vital values, and use them to perform the 48-hour mortality prediction task.
- FPRM (Multi-modal Eye Imaging, Retina Characteristics, and Psychological Assessment Dataset) [46], where we utilize 4 modalities, including M1: 2D fundus photos, M2: 3D videos of retinal blood flow, M3: 2D oxygen saturation, and M4: tabular data for health deterioration grade classification. For evaluation, we report the accuracy and F1-score of the compared methods on each classification task.

In all experiments, we run each method with 3 different random seeds and report the mean and standard deviation of the performance. Groups with approximately less than 15% frequency are denoted 'Tail groups'. Full results are in the Appendix.

Baselines. We compare REMIND with the state-of-the-art methods in tackling missingness in high-modality learning, including Soft MoE [27, 39], FuseMoE [9], FlexMoE [42]. Besides, we also compare with other long-tailed learning approaches, by adapting them to multi-modal learning models. Specifically, we combine multi-modal MoE with group robustness strategies including Distribution Adjustment [24], GroupDRO [33], FairBatch [30], and FairMixup [4].

Implementation Details. We use ViT-Base [6] as the feature encoder for image modalities in EMBED, and the T5 encoder [28] for text modalities in the MIMIC-IV dataset. For the FPRM dataset, we use a 3D-ResNet [37] for 3D modalities, 2D-ResNet [11] encoders to encode 2D image modalities, and a patch embedder to encode tabular modalities. In main experiments, we set the batch size to 32, and use the Adam optimizer [15] with a weight decay of $5e-5$. For EMBED and FPRM datasets, we use the learning rate $5e-5$, and $1e-4$ learning rate for the MIMIC dataset. More implementation details and hyperparameter ablation analysis are included in the Appendix.

5. Results

5.1. Main Results

Our method better handles high-modality learning under missingness. As shown in Tables 1, 2, and 3, our method demonstrates consistent superior performance across various datasets, under different missing modality scenarios. Either simply applying existing group robustness methods to multi-modal learning, or directly fusing multi-modalities, fail to address the core challenges posed by long-tailed modality distributions.

Our method excels on tail groups, where existing state-of-the-art methods such as FuseMoE, FlexMoE, and Soft MoE struggle. Our approach achieves superior accuracy regardless of whether tail groups consist of single modalities (MIMIC) or complex/full modality combinations (FPRM).

5.2. Analysis

Q1: Has gradient inconsistency been alleviated? Our earlier analysis identified a significant correlation between gradient consistency and the frequency of modality combinations encountered in datasets. Specifically, rare modality combinations tend to exhibit substantial gradient divergence, severely limiting model learning. Figure 3 demon-

Modality Availability				Tail	Multi-modal Learning Methods			Long-Tailed Learning Methods				Ours
M_1	M_2	M_3	M_4		Trans	FuseMoE	FlexMoE	FairMixup	Reweigh	GroupDRO	FairBatch	REMIND
		✓		✓	74.1	73.9	73.1	72.8	69.8	72.0	**75.8**	74.6
✓		✓		✓	78.1	75.3	79.2	**84.2**	74.9	77.0	82.8	79.9
			✓	✓	70.7	73.4	70.7	74.4	72.8	69.9	74.6	**74.6**
		✓	✓		77.2	76.9	77.5	78.7	77.7	78.0	78.0	**79.5**
✓		✓	✓	✓	73.3	67.3	70.9	71.5	67.3	69.7	64.9	**74.0**
✓	✓	✓	✓		81.8	81.4	81.9	83.2	81.7	81.8	83.1	**84.0**
Entire Dataset					78.5	78.2	78.6	79.9	78.9	78.9	79.7	**80.7**

Table 1. **Performance comparison (ACC$^{\uparrow}$) on EMBED dataset.** Across different missing-modality scenarios, REMIND achieves the overall best performance.

Modality Availability			Tail	Multi-modal Learning Methods			Long-Tailed Learning Methods				Ours
M_1	M_2	M_3		Trans	FuseMoE	FlexMoE	FairMixup	Reweigh	GroupDRO	FairBatch	REMIND
✓			✓	89.3	87.2	86.6	89.3	86.9	85.2	86.9	**90.8**
	✓		✓	80.1	82.8	**83.1**	82.3	76.3	77.7	67.8	78.5
✓	✓			84.8	83.4	84.2	86.6	85.7	85.2	85.2	**86.8**
		✓	✓	67.8	72.6	76.2	67.8	61.5	64.8	68.6	**76.4**
✓		✓		86.0	85.6	84.9	85.5	86.1	86.5	85.1	**88.1**
	✓	✓		78.8	79.8	75.6	**82.6**	71.8	81.3	76.4	82.3
✓	✓	✓		84.4	85.6	85.3	87.7	86.9	87.6	86.2	**88.0**
Entire Dataset				83.0	83.7	83.1	85.2	82.5	84.3	82.6	**86.1**

Table 2. **Performance comparison (ACC$^{\uparrow}$) on MIMIC-IV dataset.** Across different missing-modality scenarios, REMIND achieves the overall best performance.

Modality Availability				Tail	Multi-modal Learning Methods			Long-Tailed Learning Methods				Ours
M_1	M_2	M_3	M_4		Trans	FuseMoE	FlexMoE	FairMixup	Reweigh	GroupDRO	FairBatch	REMIND
✓				✓	82.1	84.0	84.3	83.6	83.2	84.8	84.3	**85.5**
✓	✓	✓		✓	89.0	85.2	74.8	76.8	86.1	84.2	82.9	**93.8**
✓			✓		81.4	81.5	82.3	78.8	81.4	82.2	82.0	**82.5**
✓	✓		✓	✓	66.7	66.7	75.0	78.3	66.7	83.3	75.0	**100**
✓	✓	✓	✓	✓	77.6	79.7	74.8	71.1	76.0	73.2	77.0	**86.2**
Entire Dataset					81.0	81.5	80.9	78.2	80.8	81.2	81.4	**83.8**

Table 3. **Performance comparison (ACC$^{\uparrow}$) on FPRM dataset.** Across different missing-modality scenarios, REMIND achieves the overall best performance.

strates that this gradient inconsistency problem worsens as training progresses. The head groups maintain high gradient consistency throughout training, creating a stark divergence that prevents effective learning on tail combinations. In contrast, our method demonstrates a more stable gradient consistency for both head and tail groups throughout the training. REMIND helps global updates become less head-dominated; head-vs-global consistency drops while the head–tail gap shrinks. Ablation results (Table 4) further verify that both components of DRO and expert specialization are crucial for alleviating this issue and improving model performance.

Q2: Have we achieved meaningful expert specialization? To investigate whether our group-adaptive MoE architecture learns meaningful specialization patterns, we visualize the heatmaps of expert assignment scores across different modality combination groups (Fig. 4). The visualization shows the frequency with which each expert falls into

M1	M2	M3	M4	W/o DRO	W/o Expert Spec.	Ours
		1		73.6	72.0	74.6
1		1		80.6	82.8	79.9
			1	75.1	68.8	74.6
	1	1		79.1	78.7	79.5
1		1	1	72.7	72.7	74.0
1	1	1	1	84.3	83.8	84.0
Entire Dataset				80.5	80.0	80.7

Table 4. **Ablation studies on EMBED dataset.**

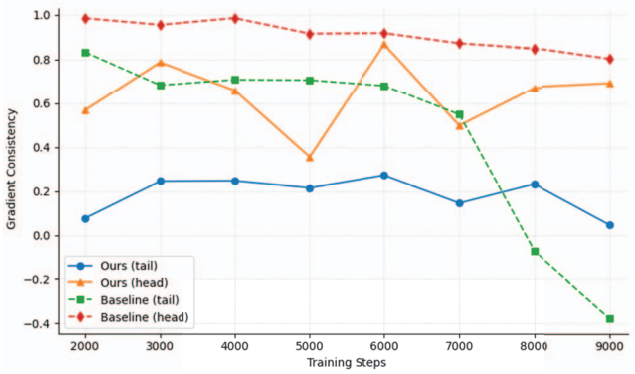

Figure 3. **Gradient inconsistencies across training steps.**

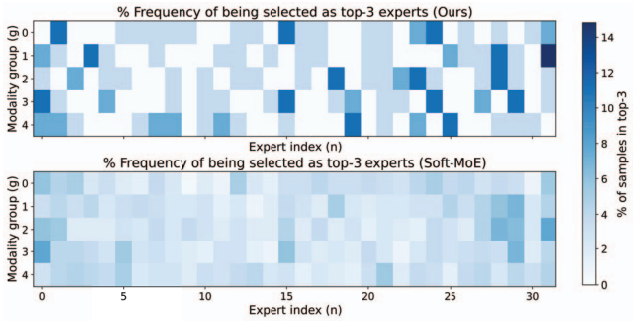

Figure 4. **Visualization of top experts patterns across modality combination groups on FPRM Dataset.**

the top-3 experts for different modality combinations. Our method exhibits significantly clearer and more distinct specialization patterns across modality combinations compared to the baseline Soft MoE, verifying REMIND has facilitated learning of group-specific fusion functions.

Q3: Can our method handle extreme missing modality scenarios? Real clinical scenarios often involve certain modalities with extremely high missing rates, which raises the question of whether incorporating such sparsely available modalities in multi-modal model development provides any benefit. To investigate whether our method can handle such scenarios, we manually created extreme missing scenarios for each modality in the FPRM dataset by simulating 80% missing rates. As shown in Table 5, for each modality M_i, we start from the subset of FPRM dataset where all modalities are available, and artificially remove a random 80% of samples from modality M_i, creating two distinct sample groups: M_i Available and M_i Absent. We then evaluate overall performance and performance on each group separately to assess whether methods can (1) effectively utilize sparse modalities when available and (2) maintain robustness when they are absent. Results demonstrate that even under extreme missingness, our method can still effectively incorporate these sparse modalities to achieve substantial performance improvements over baselines.

	Basic Soft MoE			Ours		
	Overall	M_i Available	M_i Absent	Overall	M_i Available	M_i Absent
M1	70.7	81.5	68.1	78.1 (+7.4%)	84.3 (+2.8%)	76.6 (+8.5%)
M2	73.9	71.5	74.6	84.1 (+10.2%)	84.2 (+12.7%)	84.1 (+9.5%)
M3	76.4	76.0	76.5	79.6 (+3.2%)	85.3 (+9.3%)	78.2 (+1.7%)
M4	79.4	79.4	79.4	87.3 (+7.9%)	87.8 (+8.4%)	87.2 (+7.8%)

Table 5. **Performance under 80% artificial missingness for each modality on FPRM dataset.** For each row, 80% of samples have modality M_i removed. REMIND brings large improvements

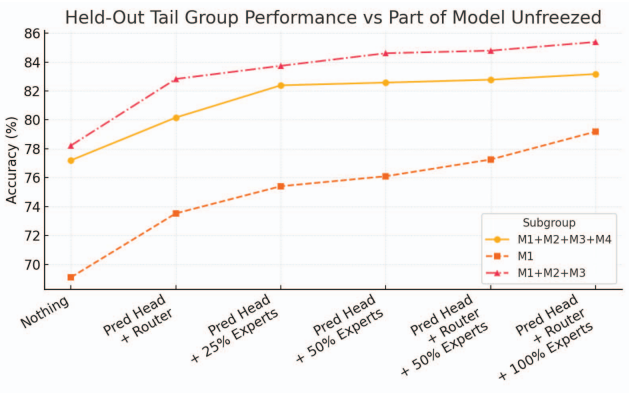

Figure 5. **Performance of held-out tail groups in the FPRM dataset.** We show performances when unfreezing and finetuning different model parts. X-Axis indicates the parts of model being finetuned: from zero-shot ('Nothing') to full finetuning ('Pred Head +Router + 100% experts').

Q4: Can our method adapt to unseen modality combinations? One interesting finding is that our framework can be easily adapted to modality combinations unseen during training by only finetuning the router matrix and prediction head. We simulated this scenario by completely excluding certain modality combinations from the training set, then evaluating different fine-tuning strategies on them. As shown in Fig. 5, our analysis reveals that unfreezing the prediction head and routing matrices could already achieve good performance with minimal parameter updates. Our group-adaptive mechanism is able to learn transferable representations and allows core expert knowledge to be shared across groups, thus requiring only router adaptation for novel modality combinations.

6. Conclusion

In this paper, we introduce REMIND, a novel framework for high-modality learning under missingness. Our analysis links prior methods' underperformance on long-tailed modality combinations to gradient inconsistency and concept shifts. Extensive experiments demonstrate that REMIND consistently improves performance, particularly on tail modality-combination groups and under extreme missing-modality scenarios.

Acknowledgment

The author acknowledges funding support by National Science Foundation (NSF) via grant IIS-2435746, Defense Advanced Research Projects Agency (DARPA) under contract No. HR00112520042, as well as the University of Michigan MIDAS PODS Grant Award.

References

[1] Julián N Acosta, Guido J Falcone, Pranav Rajpurkar, and Eric J Topol. Multimodal biomedical ai. *Nature medicine*, 28(9):1773–1784, 2022. 1

[2] Mengxi Chen, Fei Zhang, Zihua Zhao, Jiangchao Yao, Ya Zhang, and Yanfeng Wang. Probabilistic conformal distillation for enhancing missing modality robustness. *Advances in Neural Information Processing Systems*, 37:36218–36242, 2024. 2, 3, 4

[3] Qianqian Chen, Jiadong Zhang, Runqi Meng, Lei Zhou, Zhenhui Li, Qianjin Feng, and Dinggang Shen. Modality-specific information disentanglement from multi-parametric mri for breast tumor segmentation and computer-aided diagnosis. *IEEE Transactions on Medical Imaging*, 43(5):1958–1971, 2024. 3

[4] Ching-Yao Chuang and Youssef Mroueh. Fair mixup: Fairness via interpolation. *arXiv preprint arXiv:2103.06503*, 2021. 3, 6

[5] Fang Dong, Mengyi Chen, Jixian Zhou, Yubin Shi, Yixuan Chen, Mingzhi Dong, Yujiang Wang, Dongsheng Li, Xiaochen Yang, Rui Zhu, et al. Once read is enough: Domain-specific pretraining-free language models with cluster-guided sparse experts for long-tail domain knowledge. *Advances in Neural Information Processing Systems*, 37:88956–88980, 2024. 3, 12

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6, 18

[7] Stanislav Fort, Gintare Karolina Dziugaite, Mansheej Paul, Sepideh Kharaghani, Daniel M Roy, and Surya Ganguli. Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. *Advances in Neural Information Processing Systems*, 33:5850–5861, 2020. 3, 12

[8] Mohammad Hamghalam, Alejandro F Frangi, Baiying Lei, and Amber L Simpson. Modality completion via gaussian process prior variational autoencoders for multi-modal glioma segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 442–452. Springer, 2021. 3

[9] Xing Han, Huy Nguyen, Carl Harris, Nhat Ho, and Suchi Saria. Fusemoe: Mixture-of-experts transformers for fleximodal fusion. *arXiv preprint arXiv:2402.03226*, 2024. 1, 2, 6

[10] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification with dynamic evidential fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(2):2551–2566, 2022. 3

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[12] Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34:10944–10956, 2021. 2, 4

[13] Jiwoong J Jeong, Brianna L Vey, Ananth Bhimireddy, Thomas Kim, Thiago Santos, Ramon Correa, Raman Dutt, Marina Mosunjac, Gabriela Oprea-Ilies, Geoffrey Smith, et al. The emory breast imaging dataset (embed): A racially diverse, granular dataset of 3.4 million screening and diagnostic mammographic images. *Radiology: Artificial Intelligence*, 5(1):e220047, 2023. 6

[14] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023. 6

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[16] Adrienne Kline, Hanyin Wang, Yikuan Li, Saya Dennis, Meghan Hutch, Zhenxing Xu, Fei Wang, Feixiong Cheng, and Yuan Luo. Multimodal machine learning in precision health: A scoping review. *NPJ digital medicine*, 5(1):171, 2022. 1

[17] Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33:8847–8860, 2020. 4

[18] Paul Pu Liang, Zhun Liu, Yao-Hung Hubert Tsai, Qibin Zhao, Ruslan Salakhutdinov, and Louis-Philippe Morency. Learning representations from imperfect time series data via tensor rank regularization. *arXiv preprint arXiv:1907.01011*, 2019. 3

[19] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Jeffrey Tsaw, Yudong Liu, Shentong Mo, Dani Yogatama, Louis-Philippe Morency, and Ruslan Salakhutdinov. High-modality multimodal transformer: Quantifying modality & interaction heterogeneity for high-modality representation learning. *arXiv preprint arXiv:2203.01311*, 2022. 1, 3

[20] Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, et al. Quantifying & modeling multimodal interactions: An information decomposition framework. *Advances in Neural Information Processing Systems*, 36:27351–27393, 2023. 2, 4

[21] Hong Liu, Dong Wei, Donghuan Lu, Jinghan Sun, Liansheng Wang, and Yefeng Zheng. M3ae: Multimodal representation learning for brain tumor segmentation with missing modalities. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1657–1665, 2023. 3

[22] Yanbei Liu, Lianxi Fan, Changqing Zhang, Tao Zhou, Zhitao Xiao, Lei Geng, and Dinggang Shen. Incomplete multimodal representation learning for alzheimer's disease diagnosis. *Medical Image Analysis*, 69:101953, 2021. 3

[23] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2537–2546, 2019. 3

[24] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020. 3, 6

[25] Mohammed Muqeeth, Haokun Liu, and Colin Raffel. Soft merging of experts with adaptive routing. *arXiv preprint arXiv:2306.03745*, 2023. 6

[26] Joan Puigcerver, Carlos Riquelme, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. *arXiv preprint arXiv:2308.00951*, 2023. 5

[27] Joan Puigcerver, Carlos Riquelme, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. *arXiv preprint arXiv:2308.00951*, 2023. 6

[28] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 6, 18

[29] Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019. 4

[30] Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. Fairbatch: Batch selection for model fairness. In *9th International Conference on Learning Representations*, 2021. 2, 3, 6

[31] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. 4

[32] Dvir Samuel and Gal Chechik. Distributional robustness loss for long-tail learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9495–9504, 2021. 4

[33] Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. Evaluating machine accuracy on imagenet. In *International Conference on Machine Learning*, pages 8634–8644. PMLR, 2020. 3, 6

[34] Guangyuan Shi, Qimai Li, Wenlong Zhang, Jiaxin Chen, and Xiao-Ming Wu. Recon: Reducing conflicting gradients from the root for multi-task learning, 2023. *URL https://arxiv.org/abs/2302.11289*, 2023. 4

[35] Jiang-Xin Shi, Tong Wei, Yuke Xiang, and Yu-Feng Li. How re-sampling helps for long-tail learning? *Advances in Neural Information Processing Systems*, 36:75669–75687, 2023. 2, 3

[36] Phan Nguyen Minh Thao, Cong-Tinh Dao, Chenwei Wu, Jian-Zhe Wang, Shun Liu, Jun-En Ding, David Restrepo, Feng Liu, Fang-Ming Hung, and Wen-Chih Peng. Medfuse: Multimodal ehr data fusion with masked lab-test modeling and large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3974–3978, 2024. 18

[37] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 6

[38] Shicai Wei, Chunbo Luo, and Yang Luo. Mmanet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20039–20049, 2023. 1, 2

[39] Chenwei Wu, Zitao Shuai, Zhengxu Tang, Luning Wang, and Liyue Shen. Dynamic modeling of patients, modalities and tasks via multi-modal multi-task mixture of experts. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 3, 6, 18

[40] Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *arXiv preprint arXiv:2409.07825*, 2024. 3

[41] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33:5824–5836, 2020. 4

[42] Sukwon Yun, Inyoung Choi, Jie Peng, Yangfan Wu, Jingxuan Bao, Qiyiwen Zhang, Jiayi Xin, Qi Long, and Tianlong Chen. Flex-moe: Modeling arbitrary modality combination via the flexible mixture-of-experts. *arXiv preprint arXiv:2410.08245*, 2024. 1, 2, 3, 4, 6, 18

[43] Chongsheng Zhang, George Almpanidis, Gaojuan Fan, Binquan Deng, Yanbo Zhang, Ji Liu, Aouaidjia Kamel, Paolo Soda, and João Gama. A systematic review on long-tailed learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2025. 2, 3

[44] Fengda Zhang, Zitao Shuai, Kun Kuang, Fei Wu, Yueting Zhuang, and Jun Xiao. Unified fair federated learning for digital healthcare. *Patterns*, 5(1), 2024. 4

[45] Guanran Zhang, Yanlin Qu, Yanping Zhang, Jiayi Tang, Chunyan Wang, Haogui Yin, Xiaoping Yao, Gengshi Liang, Ting Shen, Qiushi Ren, et al. Multimodal eye imaging, retina characteristics, and psychological assessment dataset. *Scientific Data*, 11(1):836, 2024. 1

[46] Qingyang Zhang, Yake Wei, Zongbo Han, Huazhu Fu, Xi Peng, Cheng Deng, Qinghua Hu, Cai Xu, Jie Wen, Di Hu, et al. Multimodal fusion on low-quality data: A comprehensive survey. *arXiv preprint arXiv:2404.18947*, 2024. 1, 3, 6

[47] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(9):10795–10816, 2023. 3

[48] Fei Zhao, Chengcui Zhang, and Baocheng Geng. Deep multimodal data fusion. *ACM computing surveys*, 56(9):1–36, 2024. 3

[49] Ye Zhu, Yu Wu, Nicu Sebe, and Yan Yan. Vision+ x: A survey on multimodal learning in the light of data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9102–9122, 2024. 1