Residual-Guided Expert Specialization for Incomplete Multimodal Learning

Seunghun Baek, Jihwan Park, Jaeyoon Sim, Minjae Jeong, Hoseok Lee, and Won Hwa Kim

Pohang University of Science and Technology (POSTECH), South Korea
Code: https://github.com/seunghub/MARS

Abstract. As real-world prediction systems often face missing modalities at inference, incomplete multimodal learning (IML) remains a practical challenge. While prior methods aim to learn representations robust to missing inputs, representations from incomplete modalities inevitably deviate from their full-modality counterparts due to missing evidence. To explicitly leverage these deviations, we propose **MARS** (**M**issingness-**A**ware **R**esidual-guided **S**pecialization), a mixture-of-experts framework that guides expert specialization based on how representations are reshaped by missingness. By contrasting task representations derived from incomplete inputs with their complete counterparts during training, we derive a privileged residual signal that captures this representational gap. The residual signal guides **a residual router** to assign samples to the experts specialized for the corresponding deviation patterns. In parallel, **a feature router** learns to imitate this routing behavior using only incomplete inputs, enabling deployment without access to full modalities. To mitigate this train–test router gap, we develop a discrepancy-aware noise regularization that adaptively perturbs the residual router's decisions when the feature router deviates, enhancing the expert robustness under imperfect imitation. Experiments on multimodal classification (CASIA-SURF, CREMA-D, UPMC Food-101) and segmentation (MCubeS) under missing scenarios show that MARS consistently surpasses baselines, while remaining efficient and extensible to diverse backbones and tasks.

Keywords: Incomplete Multimodal Learning · Mixture-of-Experts

1 Introduction

Leveraging information from multiple sources (*i.e.*, modalities) has become a mainstream topic in computer vision across diverse tasks (*e.g.*, classification [20, 29, 32], segmentation [14, 31]) and domains (*e.g.*, medical imaging [1, 2, 13], remote sensing [9, 16]). By integrating complementary cues from diverse sources, multimodal learning enables a deeper understanding of the target task. However, utilizing all modalities, which were available during training, may be infeasible during inference due to practical reasons such as sensor malfunction or acquisition cost. To make multimodal systems practically reliable, it is crucial to make prediction models robust to test-time missingness. This *incomplete multimodal*

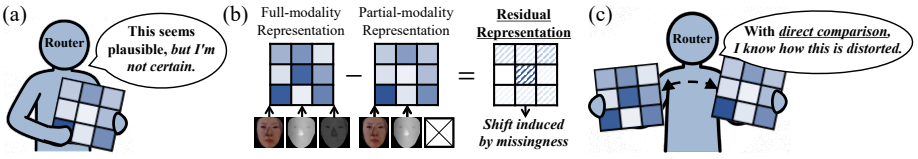


Fig. 1: Motivation of our residual-guided routing strategy. (a) In conventional Mixture-of-Experts (MoE) frameworks, the router bases its decision on the same input representation as the experts, which may lack sufficient task evidence and be distorted under missing modalities. (b) By contrasting the partial-modality representation with its full-modality counterpart, the residual captures how missingness reshapes the task representation. (c) Our router conditions expert specialization on this residual, enabling routing that accounts for such deviations and yields more reliable predictions.

learning (IML) setting is distinct from scenarios [15,28] that assume missingness during training, whose objective is to fully leverage incomplete multimodal data.

Under the IML scenario, various approaches have been proposed to handle potential modality missingness. Imputation-based methods [3,6] aim to reconstruct the missing modalities from the observed ones. However, these approaches often suffer from data hallucination [21,22] and heavy computational overhead [26], which limit their practicality. Recent efforts have therefore shifted toward learning representations that are inherently robust to potential modality missingness. Early approaches, such as LCR [33] and RFNet [7], focus on learning modality-invariant features, but they unavoidably discard useful complementary cues from individual modalities. Subsequent studies have attempted to retain the complementary information via feature-level imputation [23], uncertainty modeling [25], or expert-based architectures [12,27]. But still, they rely solely on simulated input with missingness without explicitly exploring the *direct effect of the absence*, which causes systematic representation deviation.

Key insight. Unlike previous works, we examine how modality removal reshapes the task representation. The representation computed from all modalities serves as the most task-sufficient reference, and when some modalities are absent, the resulting representation is inevitably altered as supportive evidence is removed. Although MMANet [24] attempts to align representations from partial inputs with their full-modality counterparts via knowledge distillation, the partial-modality representations still remain unreliable due to inevitable information loss. Rather than relying solely on this altered representation, we consider *the difference (i.e., residual)* between representations obtained with and without missing modalities. This residual directly reflects how the task representation shift due to modality removal, as illustrated in Fig. 1. By leveraging such deviation patterns captured by the residual, the model can better account for deviations in the input representation instead of making blind decisions.

In this regime, we introduce **MARS** (**M**issingness-**A**ware **R**esidual-guided **S**pecialization), a mixture-of-experts (MoE) framework [19] equipped with *a residual router* that inputs the residual. Exposing such residual during training allows the router to directly assign samples to experts specialized for the corresponding deviation pattern (Sec. 3.2). Notice that this residual router cannot

operate without access to complete information. Therefore, for test-time deployment with missing modalities, *a feature router* is introduced to approximate the residual router's behavior without relying on residual signals (Sec. 3.3). This *dual-router scheme* thus enables a similar routing strategy at inference.

Moreover, to mitigate the inherently inevitable discrepancy between the residual and feature routers, we introduce a discrepancy-aware noise regularization (Sec. 3.4) that adaptively increases the residual router's stochasticity for experts with larger discrepancies between the feature and residual routers. This mechanism enhances expert robustness by encouraging exploration of alternative experts that may be preferable at test time under *imperfect imitation*. In addition, a discrepancy-guided sampling (Sec. 3.5) dynamically prioritizes modality combinations where the routers disagree the most, facilitating balanced learning under incomplete conditions. Together, these components yield robust expert behaviors and consistent performance gains across diverse domains and tasks.

Overall, our contributions are summarized as threefold:

- **Residual perspective.** We reformulate IML through a characterization of representation deviations in incomplete inputs, where routing is conditioned on such deviations while experts operate on incomplete features. This separation enables explicit expert specialization to missing-modality patterns rather than relying solely on unreliable incomplete representation.
- **Dual-router design.** We introduce a dual-router framework that disentangles expert specialization and deployment. A deployable feature router learns to replicate the privileged residual router's behavior at inference. Beyond knowledge distillation, we further guide the routing noise to mitigate train–test routing discrepancy and promote robust expert specialization.
- **Efficiency and generality.** Through extensive experiments on four datasets across diverse domains and tasks, MARS consistently improves performance across modality combinations while remaining computationally efficient.

2 Related Work

Existing IML methods for test-time modality missingness can be broadly categorized into imputation-based, representation-based, and expert-based approaches.

Imputation-based methods. Early works attempt to impute missing inputs from existing modalities. Generative models [3,6] have been employed to reconstruct the absent modalities. However, retrieving signals from partial observations often leads to quality issues [21,22], while the computational overhead and reliance on generative models at inference make real-time applications infeasible.

Representation learning under incompleteness. A subsequent line of work aims to learn representations that are robust to simulated modality missingness. HeMIS [10], LCR [33] and RFNet [7] learn modality-invariant representations, whereas mmFormer [30] and ShaSpec [23] retain modality-specific cues. While MMANet [24] distills model knowledge from a full-modality teacher to a partial-modality student, DMRNet [25] enhances robustness through probabilistic uncertainty modeling. However, the estimated uncertainty itself remains unreliable, as it is inferred solely from partial observations without explicit supervision.

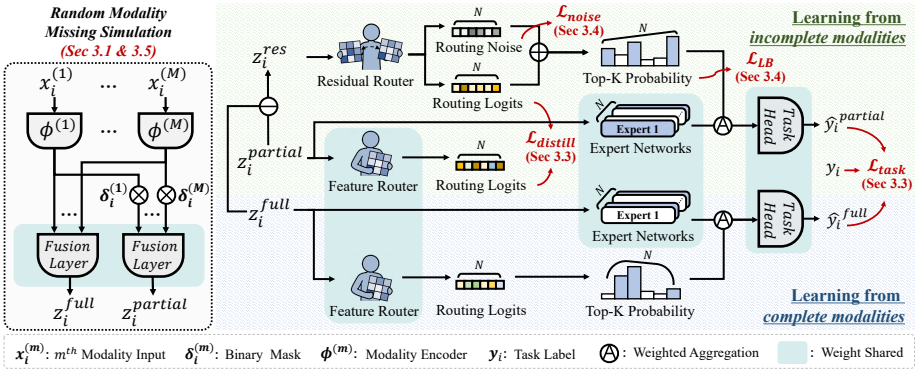


Fig. 2: Overview of MARS. During training, complete and incomplete features (*i.e.*, z_i^{full} and $z_i^{partial}$) are used to compute residuals that guide the residual router to specialize experts in distinct representation deviation patterns. The feature router learns to imitate this routing, enabling deployment without access to residuals. Discrepancy-aware noise and load-balancing further enhance expert robustness and diversity.

Expert-based specialization. MoMKE [27] first introduced modality-specific experts, trained on unimodal data, and aggregated them through a soft router during a second training stage. However, all available inputs are fed into every expert during aggregation, which leads to unnatural cross-feeding (e.g., using RGB input in an IR-trained expert) and high computational overhead. SimMLM [12] alleviated this by routing each modality only to its corresponding expert, thereby achieving more efficient inference. Nevertheless, its aggregation operates at the logit level, which limits fine-grained specialization and increases computational cost. Moreover, both methods require two-stage training and impose a structural constraint that ties the number of experts to the number of modalities.

In contrast, Flex-MoE [28] defines experts at the modality-combination level, assigning a dedicated expert to each combination while allowing additional experts to be selected per sample. However, this design still requires at least 2^M-1 experts for M modalities. Moreover, defining experts strictly by modality combinations restricts specialization to predefined modality combinations. In practice, the task-specific contribution of each modality may vary across samples. Samples within the same combination may require distinct expertise, while different combinations often exhibit similar representation deviation patterns.

Our perspective. Our model neither *i)* imposes structural constraints or predefined roles on experts, nor *ii)* relies solely on unreliable embeddings with missing modalities. By comparing task embeddings from partial observations with their complete counterparts, the router conditions expert specialization on representation deviation patterns, enabling adaptive handling of incomplete inputs.

3 Method

The goal of MARS is *i)* to train experts to specialize in task-specific deviation caused by modality missingness, and *ii)* to ensure consistent utilization of this

specialization during inference. Fig. 2 illustrates the overview, and following sections describe how its components and training procedure are designed.

3.1 Problem Definition

Following recent works [12,25,27], we consider the general setting of IML, where some modalities may be missing at inference while remaining fully available during training. Given a dataset, each sample is denoted as $x_i = \{x_i^{(m)}\}_{m=1}^{M}$, where M indicates the number of modalities. Each sample with the m-th modality $x_i^{(m)}$ is encoded by a modality-specific encoder $\phi^{(m)}$, which yields $e_i^{(m)}{=}\phi^{(m)}(x_i^{(m)})$. To simulate missing-modality scenarios during training, each modality is stochastically deactivated using a binary indicator $\delta_i^{(m)} \in \{0, 1\}$, where $\delta_i^{(m)} = 0$ masks its corresponding embedding $e_i^{(m)}$. Masked embeddings are concatenated and fused by a shared fusion layer f_{fuse} to obtain a unified multimodal representation as

$$z_i = f_{\text{fuse}}\Big(\text{Concat}\big(\{\delta_i^{(m)}\, e_i^{(m)}\}_{m=1}^{M}\big)\Big). \tag{1}$$

This task representation z_i is later used to compute task-specific predictions.

In previous methods, models process only one masked input (i.e., a single modality combination) per sample, from which a single task representation z_i is extracted and used for prediction. In contrast, MARS jointly operates on two fused features of the same sample, one obtained from *the complete input* and the other from *the masked input*. By directly comparing these features, MARS learns an explicit notion of how representations change under missing modalities.

Preliminaries. A sparse Mixture-of-Experts model [19] replicates some parts of the network as multiple expert modules. A router assigns scores to experts based on the input representation and aggregates the top-scoring expert outputs through weighted average. The router's decision encourages experts to specialize across diverse input patterns with a modest increase in computational cost.

3.2 MoE with Privileged Residual Router

When all modalities are available, we refer to resulting task representation z_i as a complete feature z_i^{full}, otherwise as an incomplete feature z_i^{partial}:

$$z_i = \begin{cases} z_i^{\text{full}}, & \text{if } \forall m,\ \delta_i^{(m)} = 1 \\ z_i^{\text{partial}}, & \text{otherwise.} \end{cases} \tag{2}$$

The MoE module is trained under a privileged setting where both z_i^{full} and z_i^{partial} are provided. Their difference z_i^{res} provides a residual signal that explicitly characterizes missingness-induced representation deviation patterns:

$$z_i^{\text{res}} = z_i^{\text{full}} - z_i^{\text{partial}}. \tag{3}$$

This z_i^{res} is fed into a *residual router* $\mathcal{R}_{\text{res}}$, which guides each expert to specialize in how the routed feature z_i^{partial} deviates. The $\mathcal{R}_{\text{res}}$ yields a clean logit vector

$l_i^{\text{res}} \in \mathbb{R}^N$ and an estimated noise standard deviation $\sigma_i \in \mathbb{R}^N$, where N denotes the number of experts. To encourage diverse expert exploration, we adopt *noisy top-K routing* [19] where Gaussian noise ϵ is injected to make logits l_i^{res} noisy as

$$\tilde{l}_i^{\text{res}} = l_i^{\text{res}} + \epsilon\sigma_i, \quad \epsilon \sim \mathcal{N}(0, I). \tag{4}$$

The top-K experts with the highest noisy logits $\tilde{l}_i^{\text{res}}$ are selected to promote specialization while keeping the computation efficient. Their routing probabilities p_i^{res} are then computed using top-K softmax operator $\text{Softmax}_K(\cdot)$, which applies the softmax over the selected top-K logits while assigning zero to the rest:

$$p_i^{\text{res}} = \text{Softmax}_K(\tilde{l}_i^{\text{res}}) \in \mathbb{R}^N. \tag{5}$$

Let $(p_i^{\text{res}})_j$ denote the routing probability of expert j for sample i. Each expert $E_j(\cdot)$ processes z_i^{partial}, and its output is weighted by $(p_i^{\text{res}})_j$. The aggregated output is then passed to a task head $h(\cdot)$, which produces a prediction $\hat{y}_i^{\text{partial}}$:

$$\hat{y}_i^{\text{partial}} = h\left(\sum_{j=1}^{N} (p_i^{\text{res}})_j \, E_j(z_i^{\text{partial}})\right). \tag{6}$$

This residual routing directly supervises deviation-aware expert specialization.

Remarks. The formulation above faces two practical issues: 1) z_i^{res} is unavailable at inference when complete modalities are missing. 2) When all modalities are present, z_i^{res} becomes zero, resulting in identical routing across samples. We therefore introduce a *feature router* that operates without privileged information.

3.3 Preparing for Unprivileged Inference

To enable practical deployment without residual signal, we introduce a feature router $\mathcal{R}_{\text{fea}}$ that performs routing using only the available embedding $z_i \in \{z_i^{\text{full}}, z_i^{\text{partial}}\}$. It is designed as the deployable router at inference, capable of handling both complete and incomplete modality conditions.

Under full modality condition, residual z_i^{res} inherently becomes zero. Routing is thus entirely handled by $\mathcal{R}_{\text{fea}}$ as $p_i^{\text{fea}} = \text{Softmax}_K(\mathcal{R}_{\text{fea}}(z_i^{\text{full}}))$. The task prediction $\hat{y}_i^{\text{full}}$ follows the same aggregation process in Eq. (6) as $\hat{y}_i^{\text{full}} = h\left(\sum_{j=1}^{N} (p_i^{\text{fea}})_j E_j(z_i^{\text{full}})\right)$. With a task-specific objective ℓ (e.g., cross-entropy) and target label y_i, the task-loss $\mathcal{L}_{task}$ combines both full and partial cases as

$$\mathcal{L}_{\text{task}} = \ell(\hat{y}_i^{\text{full}}, y_i) + \ell(\hat{y}_i^{\text{partial}}, y_i). \tag{7}$$

For incomplete inputs, $\mathcal{R}_{\text{res}}$ performs privileged routing using z_i^{res} to obtain $\hat{y}_i^{\text{partial}}$. In parallel, $\mathcal{R}_{\text{fea}}$ processes z_i^{partial} but is optimized only through a distillation loss $\mathcal{L}_{\text{distill}}$ that aligns its logits l_i^{fea} to those of $\mathcal{R}_{\text{res}}$ as

$$\mathcal{L}_{\text{distill}} = D_{\text{KL}}\left(\text{Softmax}(\text{GradStop}(l_i^{\text{res}})) \parallel \text{Softmax}(l_i^{\text{fea}})\right), \tag{8}$$

where D_{KL} denotes the Kullback–Leibler divergence, and $\text{GradStop}(\cdot)$ detaches gradients to prevent updates. This enables $\mathcal{R}_{\text{fea}}$ to learn from the privileged $\mathcal{R}_{\text{res}}$, making it the sole deployable router with similar expert selection at inference.

3.4 Discrepancy-Aware Noise Regularization

With noisy routing (Sec. 3.2), MoE frameworks typically incorporate a load–importance balancing loss $\mathcal{L}_{\text{LB}}$ [19] for balanced expert utilization. Let B denote the mini-batch size, and define the importance and load of expert j as

$$\text{importance}_j = \sum_{i=1}^{B} (p_i^{\text{res}})_j, \quad \text{load}_j = \sum_{i=1}^{B} \mathbb{I}[(p_i^{\text{res}})_j > 0], \tag{9}$$

where $\mathbb{I}(\cdot)$ is an indicator function that equals 1 if the condition holds, and 0 otherwise. The load-importance balancing loss $\mathcal{L}_{\text{LB}}$ is then computed as

$$\mathcal{L}_{\text{LB}} = \text{CV}^2\Big(\{\text{importance}_j\}_{j=1}^{N}\Big) + \text{CV}^2\Big(\{\text{load}_j\}_{j=1}^{N}\Big), \tag{10}$$

where squared coefficient of variation $\text{CV}^2(x) = (\sigma(x)/\mu(x))^2$ measures the uniformity across experts. This $\mathcal{L}_{\text{LB}}$ prevents expert collapse and ensures balanced participation, while the injected routing noise $\epsilon\sigma_i$ (Eq. (4)) promotes stochastic expert exploration. Beyond this conventional role, we further leverage the noise to mitigate the unavoidable routing discrepancy between training and inference.

The residual router $\mathcal{R}_{\text{res}}$ provides an optimal supervision by leveraging both z_i^{partial} and z_i^{full}. However, despite the distillation loss $\mathcal{L}_{\text{distill}}$, $\mathcal{R}_{\text{fea}}$ cannot fully reproduce the behavior of $\mathcal{R}_{\text{res}}$ since it operates only with z_i^{partial}. This inevitably widens the train–test routing gap which causes inconsistent expert assignments.

To mitigate this gap, we introduce a discrepancy-aware noise regularization $\mathcal{L}_{\text{noise}}$, which scales the noise variance $(\sigma_i^2)_j$ with the routing discrepancy of expert j. Higher noise variance induces the stochasticity of expert selection, allowing experts to remain robust even when the $\mathcal{R}_{\text{fea}}$ replaces $\mathcal{R}_{\text{res}}$ at inference.

Let l_i^{res} and l_i^{fea} denote the clean logits from $\mathcal{R}_{\text{res}}$ and $\mathcal{R}_{\text{fea}}$, respectively. We extract the top-K expert indices $\mathcal{T}_i^{\text{res}}, \mathcal{T}_i^{\text{fea}} \in \mathbb{R}^K$ from each router as

$$\mathcal{T}_i^{\text{res}} = \text{TopK}(l_i^{\text{res}}, K), \quad \mathcal{T}_i^{\text{fea}} = \text{TopK}(l_i^{\text{fea}}, K), \tag{11}$$

and define their union $\mathcal{U}_i = \mathcal{T}_i^{\text{res}} \cup \mathcal{T}_i^{\text{fea}}$. For each $j \in \mathcal{U}_i$, we measure the routing discrepancy $(m_i)_j$ as the absolute difference between the normalized clean logits:

$$(m_i)_j = \left|\text{Softmax}(l_i^{\text{res}})_j - \text{Softmax}(l_i^{\text{fea}})_j\right|, \quad j \in \mathcal{U}_i. \tag{12}$$

We then sort experts in $\mathcal{U}_i$ in descending order of $(m_i)_j$, which yields a permutation π_i such that $(m_i)_{\pi_i(1)} \geq \ldots \geq (m_i)_{\pi_i(|\mathcal{U}_i|)}$. To promote larger noise for experts with greater discrepancy, we penalize cases where the ordering of $(\sigma_i^2)_j$ contradicts that of $(m_i)_j$. Thus, the noise regularization $\mathcal{L}_{\text{noise}}$ is formulated as

$$\mathcal{L}_{\text{noise}} = \frac{1}{|\mathcal{U}_i| - 1} \sum_{j=1}^{|\mathcal{U}_i|-1} \text{Softplus}\Big(\log(\sigma_i)^2_{\pi_i(j+1)} - \log(\sigma_i)^2_{\pi_i(j)}\Big), \tag{13}$$

which softly enforces $(\sigma_i^2)_{\pi_i(1)} \geq \cdots \geq (\sigma_i^2)_{\pi_i(|\mathcal{U}_i|)}$. As experts with greater discrepancy $(m_i)_j$ are guided to exhibit higher routing noise $(\sigma_i^2)_j$, this encourages

exploration of alternative experts where the routers disagree. By being stochastically activated, these experts become more robust to test-time routing mismatch.

Finally, the overall training objective $\mathcal{L}_{\text{total}}$ is defined as

$$\mathcal{L}_{\text{total}} = \lambda_{\text{task}}\mathcal{L}_{\text{task}} + \lambda_{\text{LB}}\mathcal{L}_{\text{LB}} + \lambda_{\text{distill}}\mathcal{L}_{\text{distill}} + \lambda_{\text{noise}}\mathcal{L}_{\text{noise}}, \tag{14}$$

where each λ controls the balance between individual loss terms $\mathcal{L}$.

3.5 Top-K Discrepancy-Guided Modality Sampling

We further introduce an adaptive sampling strategy to balance the learning progress across different incomplete modality configurations. At the beginning of training, configurations are sampled uniformly, while the full-modality case is always included to provide complete supervision. After several warm-up epochs, the sampling probabilities are updated according to the routing discrepancy.

Let $\mathcal{C} = \{c_1, \ldots, c_{|\mathcal{C}|}\}$ denote all feasible incomplete modality combinations, where each $c_{j\in\{1,\ldots,|\mathcal{C}|\}}$ is represented by a binary vector composed of the masking indicators $\{\delta^{(m)} \in \{0,1\}\}_{m=1}^{M}$. We denote q_j as the sampling probability of configuration c_j satisfying $\sum_{j=1}^{|C|} q_j = 1$. For each training epoch, we measure the average top-K routing disagreement d_j between the routers over all samples whose masked configuration corresponds to c_j as

$$d_j = 1 - \frac{1}{|\mathcal{D}_{c_j}|} \sum_{i\in\mathcal{D}_{c_j}} \frac{|\mathcal{T}_i^{\text{res}} \cap \mathcal{T}_i^{\text{fea}}|}{K}, \tag{15}$$

where $\mathcal{D}_{c_j}$ denotes the set of training samples assigned to c_j throughout the current epoch. At the end of the t-th iteration, discrepancies are normalized by a softmax to update the next sampling probabilities $q_j^{(t+1)}$ as

$$q_j^{(t+1)} = \frac{\exp(d_j/\tau)}{\sum_{k=1}^{|\mathcal{C}|}\exp(d_k/\tau)}, \tag{16}$$

where the temperature τ adjusts the smoothness. Combinations showing higher router disagreement d_j are sampled more frequently in subsequent epochs, allowing the model to focus on modality settings that are harder to align.

4 Experiments

We evaluate our method on 1) multimodal spoof classification, 2) material segmentation, 3) emotion recognition, and 4) food classification using CASIA-SURF [29], MCubeS [14], CREMA-D [4], and UPMC Food-101 [8]. For all datasets, missing-modality conditions follow the common protocols (Sec. 3.1) used in previous IML works [12, 24, 25, 27]. Due to space limitations, detailed dataset descriptions and result analyses for CREMA-D [4] (Tab. 3) and UPMC Food-101 [8] (Tab. 4) are provided in Sec. B of the supplementary material.

4.1 Datasets

CASIA-SURF [29] is a large-scale multimodal face anti-spoofing benchmark that contains RGB, Depth, and Infrared (IR) modalities. Each modality provides complementary cues for distinguishing live faces from spoofing attacks. The dataset offers binary labels indicating live or spoof, with N=29k, 1k, and 57k samples for training, validation, and testing data. Performance is evaluated using the Average Classification Error Rate (ACER, %) [29], which represents the percentage of misclassified samples. CASIA-SURF has been adopted as a standard benchmark for incomplete multimodal classification [24, 25].
MCubeS [14] is a multimodal material segmentation dataset composed of 500 samples captured from 42 diverse outdoor scenes. Each sample includes four modalities: RGB, Degree of Linear Polarization (DoLP), Angle of Linear Polarization (AoLP), and Near-Infrared (NIR) reflectance. Every pixel is annotated with one of 20 material categories such as *concrete*, *water*, and *metal*. The official split includes 302 training, 96 validation, and 102 test samples. Performance is measured by the mean Intersection-over-Union (mIoU), where higher values indicate better segmentation accuracy. We employ this recent dataset in IML scenarios to assess segmentation robustness under missing-modality conditions.
CREMA-D [4] and **UPMC Food-101** [8] are multimodal classification datasets comprising *audio–visual* and *image–text* modalities, respectively. They enable evaluation beyond vision-centric datasets. We follow the experimental setups of DMRNet [25] for CREMA-D and SimMLM [12] for UPMC Food-101.

4.2 Implementation Details

For CASIA-SURF [29] and MCubeS [14], we adopt a sparse MoE [19] framework with 16 experts, where the top five are activated per sample. The residual router $\mathcal{R}_{res}$ employs a single linear projection layer, and the feature router $\mathcal{R}_{fea}$ consists of a two-layer MLP. All experiments are run on a single NVIDIA RTX 6000 ADA.
CASIA-SURF. Following DMRNet [25], we employ a ResNet-18 [11] backbone with a binary indicator $\delta^{(m)}$ (Sec. 3.1). The final convolution block is used as the expert set. Training is performed for 100 epochs using SGD [17] with a learning rate of 5×10^{-4}, weight decay of 5×10^{-4}, and a batch size of 64. Loss weights are λ_{task}=1, λ_{LB}=0.05, λ_{distill}=1, and λ_{noise}=0.01. Sampling begins at epoch 20.
MCubeS. We build on DeepLab v3+ [5] with the same Bernoulli indicator scheme. The final layer before the task head serves as the expert set. Models are trained for 500 epochs using SGD with a learning rate of 5×10^{-3}, weight decay of 5×10^{-4}, and a batch size of 8. Loss weights are set to λ_{task}=1, λ_{LB}=0.01, λ_{distill}=0.05, and λ_{noise}=0.1, and our sampling strategy starts from epoch 30.
CREMA-D and **UPMC Food-101.** Refer to Sec. B (supplementary material).

4.3 Baselines

We compare our method with recent state-of-the-art approaches. For classification, we report the DMRNet [25] results from the original paper and reimplement Flex-MoE [28], MoMKE [27], and SimMLM [12] using a ResNet-

Table 1: Classification results (ACER ↓) on CASIA-SURF [29] under all modality combinations. MARS consistently outperforms baselines and achieves the lowest error rate across all settings, demonstrating strong robustness to missing modalities. The best and second-best results are highlighted with **bold** and underlined, respectively. (●: Presence, ○: Absence)

Modality			Methods											
RGB	Depth	IR	ResNet-18	HeMIS	LCR	RFNet	mmFormer	ShaSpec	MMANet	DMRNet	Flex-MoE	MoMKE	SimMLM	MARS
●	○	○	11.75	14.36	13.44	12.43	11.15	11.57	8.57	8.23	9.10	10.04	7.82	**6.92**
○	●	○	5.87	4.70	4.40	4.17	3.67	6.25	2.27	2.01	2.63	1.95	2.42	**1.87**
○	○	●	16.62	16.21	15.26	14.69	13.99	10.71	10.04	8.98	7.25	12.57	9.05	**3.96**
●	●	○	4.61	3.23	3.32	2.23	1.93	3.11	1.61	1.21	1.27	1.20	1.13	**1.06**
●	○	●	6.68	6.27	5.16	4.27	4.77	4.23	3.01	3.00	3.58	5.09	3.38	**2.09**
○	●	●	4.95	3.68	3.53	3.22	3.10	2.52	1.18	0.80	1.19	1.06	1.15	**0.66**
●	●	●	2.21	1.97	1.88	1.18	1.94	1.79	0.87	0.66	0.84	0.79	0.63	**0.45**
Average			7.52	7.18	6.71	6.02	5.93	5.59	3.94	3.58	3.69	4.67	3.66	**2.43**

Table 2: Segmentation results (mIoU ↑) on MCubeS [14] with all RGB combinations.

RGB	AoLP	DoLP	NIR	DeepLab v3+	DMRNet	Flex-MoE	MoMKE	SimMLM	MARS
●	○	○	○	0.4249	0.4647	0.4490	0.4150	0.4462	**0.4730**
●	●	○	○	0.4238	0.4682	0.4528	0.4241	0.4473	**0.4767**
●	○	●	○	0.4249	0.4670	0.4529	0.4326	0.4420	**0.4753**
●	○	○	●	0.4247	0.4672	0.4515	0.4008	0.4355	**0.4769**
●	●	●	○	0.4269	0.4694	0.4559	0.4173	0.4427	**0.4777**
●	●	○	●	0.4269	0.4701	0.4554	0.4325	0.4286	**0.4808**
●	○	●	●	0.4261	0.4691	0.4558	0.4238	0.4224	**0.4776**
●	●	●	●	0.4271	0.4712	0.4586	0.4376	0.4220	**0.4808**
Average				0.4257	0.4683	0.4540	0.4273	0.4367	**0.4773**

Table 3: Results on CREMA-D [4].

Audio	Visual	ShaSpec	MMANet	DMRNet	MARS
●	○	59.86	58.89	58.87	**61.04**
○	●	47.17	47.31	55.10	**59.43**
●	●	66.80	67.87	70.10	**76.10**
Average		57.94	57.66	61.35	**65.52**

Table 4: Results on Food-101 [8].

Image	Text	ShaSpec	MoMKE	SimMLM	MARS
●	○	69.22	70.46	72.20	**91.86**
○	●	86.55	86.59	87.20	**90.18**
●	●	92.73	92.71	**94.99**	92.72
Average		82.83	83.25	84.81	**91.59**

18 [11] backbone for fair comparison. Since the original Flex-MoE assumes incomplete modalities during training, we adapt it by providing both complete and randomly dropped modalities via $\delta^{(m)}$. For segmentation, all baselines are re-implemented on DeepLab v3+ [5]. All hyperparameters are tuned for optimal performance. Further experimental details are given in Sec. D of the supplementary material.

4.4 Spoof Classification Results

Tab. 1 summarizes the spoof classification results on the CASIA-SURF dataset. Among the baselines, DMRNet [25] achieves the lowest average ACER of 3.58, while MARS achieves 2.43 reducing it by 1.15. The model consistently attains the lowest ACER across all modality combinations, demonstrating strong robustness to missing modalities. The most challenging case occurs when only the IR is available. In this setting, MARS reduces the ACER from 8.98 (in DMRNet) to 3.96, effectively reducing the error by more than half. These results confirm that MARS robustly handles representation deviations induced by modality removal.

4.5 Material Segmentation Results

Tab. 2 reports the segmentation results on the MCubeS [14] dataset. Following the initial work [14], we focus on combinations that include the RGB, since non-RGB modalities, such as NIR, AoLP, or DoLP, alone lack sufficient radiometric

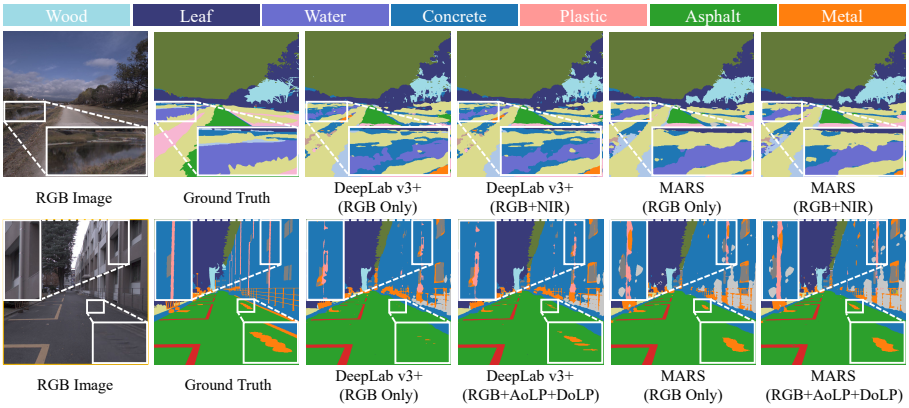


Fig. 3: Qualitative results on multimodal material segmentation (MCubeS [14]). Top: When near-infrared (NIR) information is missing, DeepLab v3+ [5] degrades in recognizing water and reflected objects, whereas MARS remains relatively consistent. Bottom: In the absence of polarization cues (AoLP/DoLP), MARS correctly separates metallic and dielectric materials even under incomplete inputs.

diversity to discern all classes. Consistent with the classification results, DMR-Net [25] achieves the best baseline performance with an average mIoU of 0.4683. MARS improves it to 0.4773, surpassing baselines across all combinations.

Fig. 3 presents qualitative comparisons on test samples from the MCubeS dataset. When compared to DeepLab v3+ [5], MARS provides clearer and more consistent segmentation results. When the NIR is missing, which helps identify water, DeepLab v3+ often misinterprets reflected surfaces over water as separate objects, while MARS remains robust to such misleading cues. Similarly, polarization cues such as AoLP and DoLP are important for distinguishing metallic and dielectric materials. Without them, the baseline predictions frequently confuse concrete, metal, and plastic, whereas MARS produces more stable segmentation results. These observations suggest that MARS better handles regions that are ambiguous under the given modalities, as *residual guidance encourages experts to focus on reliable cues* rather than misleading artifacts.

5 Model Behavior Analysis

We analyze the MARS framework to better understand the effect of its key components and the behavior of expert routing under missing-modality scenarios.

5.1 Ablation Study

Tab. 5 presents the ablation results on CASIA-SURF and MCubeS. Starting from each baseline, ResNet-18 and DeepLab v3+, the introduction of the MoE structure improves performance by diversifying strategies on inputs. Adding residual routing further enhances those MoE by leveraging the representational deviations for expert specialization. Variance ordering brings an additional gain by guiding the routing noise to account for train-test gap induced from router

Table 5: Ablation study. Performance is averaged over all modality configurations. "+" denotes cumulative inclusion.

Method	CASIA-SURF (ACER ↓)	MCubeS (mIoU ↑)
ResNet-18 / DeepLab v3+	7.52	0.4257
+ MoE	4.12	0.4327
+ Residual Routing	3.78	0.4661
+ Noise Regularization	2.72	0.4750
+ Discrepancy-guided sampling	2.43	**0.4773**
Oracle Routing	**0.94**	0.4768

Table 6: Error rate (ACER, %) under expert deactivation on CASIA-SURF. Changes in ().

Modality			Deactivated Expert			
RGB	Depth	IR	None	E12	E0	E5
●	○	○	6.92	**7.46 (0.54↑)**	7.54 (0.62↑)	6.92 (0.00–)
○	●	○	1.87	1.87 (0.00–)	1.74 (0.13↓)	1.87 (0.00–)
○	○	●	3.96	3.96 (0.00–)	**7.58 (3.62↑)**	3.96 (0.00–)
●	●	○	1.06	0.99 (0.07↓)	0.93 (0.13↓)	1.06 (0.00–)
●	○	●	2.09	1.52 (0.57↓)	1.73 (0.36↓)	2.09 (0.00–)
○	●	●	0.66	0.66 (0.00–)	0.48 (0.18↓)	0.66 (0.00–)
●	●	●	0.45	0.31 (0.14↓)	0.36 (0.09↓)	0.93 (0.48↑)
Average			2.43	2.40 (0.03↓)	2.91 (0.48↑)	2.50 (0.07↑)

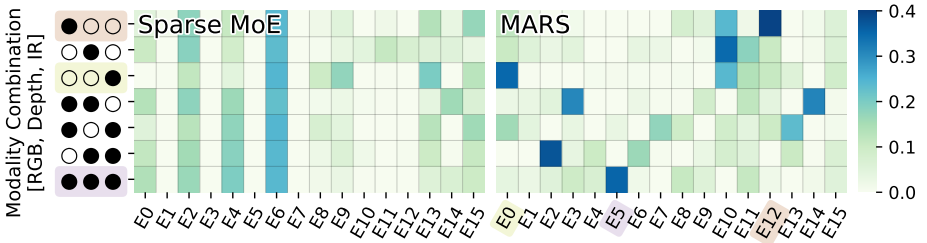


Fig. 4: Top-K routing probabilities of the feature router at inference on CASIA-SURF [29]. Probabilities are averaged over modality combination. Our method shows diverse expert utilization, unlike the baseline MoE [19] activating similar experts.

change. Finally, discrepancy sampling, which selectively emphasizes large gap between routers cases during training, yields the best overall performance.

We add an experiment to solidify our key idea that *residual signal promotes optimal expert specializations*. The Oracle routing experiment assumes access to full modalities even at test time, where the true residuals are fed to the residual router for inference. This setting achieves near-perfect classification accuracy (ACER 0.94), confirming that the experts themselves are already well specialized to modality missingness. Interestingly, in segmentation, our model slightly surpasses the Oracle (mIoU 0.4773 vs. 0.4768). We attribute this to the noise regularization, which prevents overreliance on the residual routing and encourages the experts to learn complementary decision boundaries for the feature routing.

5.2 Effectiveness of Residual Routing

To show the effectiveness of residual routing in expert specialization, we analyze the routing distribution (Fig. 4) and visualizations of feature attribution (Fig. 5).

Routing Distribution. Fig. 4 shows the routing probabilities of the feature router at test time on CASIA-SURF [29]. For each modality combination, we collect the top-K routing probabilities and average them across all samples to depict the expert activation distribution. To ensure a fair comparison, both the original MoE and our model share identical weight of hyperparameters including load-balancing weight λ_{LB}. In the baseline MoE, routing remains biased toward a few dominant experts irrespective of the modality combination. In contrast, our residual routing exhibits distinct yet structured expert utilization among different combinations. Most combinations activate at most two dominant experts,

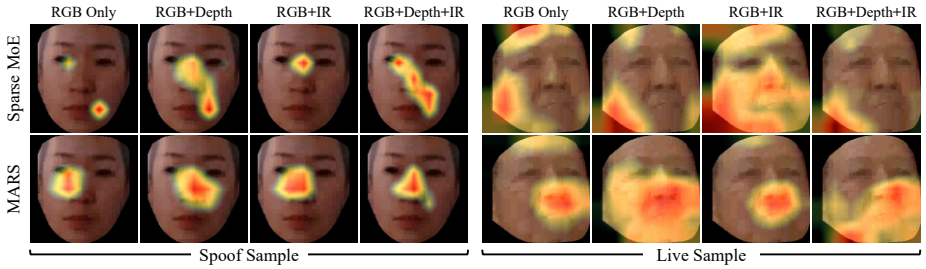
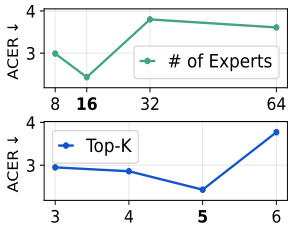


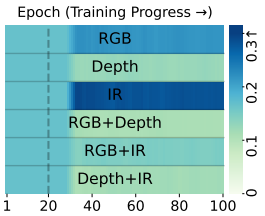
Fig. 5: Grad-CAM [18] visualization on CASIA-SURF [29]. Heatmaps are generated by backpropagating the predicted logit of the target class through z_i and then overlaid on the RGB image. Compared to the baseline MoE [19], which inconsistently attends to superficial artifacts (*e.g.*, around the mouth corner), MARS produces stable and semantically meaningful activations (*e.g.*, around the nose) across diverse combinations.



(a) Sensitivity analysis.

Method	#Params	FLOPs
DeepLab v3+	235M	639G
DMRNet	238M	639G
Flex-MoE	672M	716G
MoMKE	245M	1297G
SimMLM	247M	955G
MARS	244M	716G

(b) Complexity analysis.



(c) Sampling analysis.

Fig. 6: Comprehensive analyses of MARS. (a) Effect of the number of experts N and top-K on CASIA-SURF [29]. (b) Parameter and FLOPs comparison on MCubeS [14]. (c) Sampling probabilities across modality combinations on CASIA-SURF [29].

which rarely overlap to other combinations. Especially, we identify strong associations between specific modality combinations and experts (e.g., RGB–E12, IR–E0, Full–E5). This implies that the router effectively operates on missingness.

To further verify whether the experts act responsible for the combination, we deactivate certain experts at inference and summarize the corresponding performance drops for each combination in Tab. 6. The combinations most strongly associated with the expert consistently incur the largest error increase. This confirms that experts with strong affinity to particular modality combinations learn disentangled specializations aligned with distinct missing-modality patterns.

Feature Attribution. Fig. 5 visualizes Grad-CAM [18] results on CASIA-SURF [29]. For test samples x_i, we backpropagate the predicted target logit $\hat{y}_i$ through the fused embedding z_i, and overlay the activation map on its RGB image x_i^{RGB}. Each selected configuration includes one spoof and one live sample.

In the baseline MoE, attention regions vary largely across modality combinations, often highlighting superficial artifacts (e.g., mouth corner or forehead in RGB-only input). In contrast, our method consistently focuses on stable facial regions, particularly around the nose, which serves as a key discriminative cue for spoof detection. It also adapts its attention to secondary areas, such as the eyes or cheeks, depending on the available signals. This coherent attention pattern indicates that residual routing encourages reasoning based on reliable features rather than overfitting to spurious modality-specific artifacts.

5.3 Comprehensive Analyses

Sensitivity Study. We analyze the sensitivity on the CASIA-SURF dataset [29]. In Fig. 6a, the top plot varies the number of experts with top-K fixed to 5, and the bottom plot varies top-K with the number of experts fixed to 16. We observe that the performance degrades when either the number of experts or top-K increases beyond the optimal setting. This suggests that employing too many experts leads to redundancy among them, whereas activating more experts per sample could reduce their specialization.
Complexity Study. We evaluate model complexity on the MCubeS dataset [14] in terms of FLOPs and parameter count, as summarized in Fig. 6b. We use segmentation to assess scalability, since its backbone (DeepLab v3+ [5]) and resulting feature maps are heavier than those in classification tasks, providing a more realistic view of computational extensibility. While Flex-MoE [28] introduces substantial parameter overhead due to its modality bank, MARS remains compact by adding only a lightweight residual router. Our FLOPs exhibit a moderate increase compared to DeepLab v3+, which is expected from adopting the sparse MoE structure. In contrast, other MoE-based frameworks, *i.e.*, MoMKE [27] and SimMLM [12], show considerably higher computational costs, *i.e.*, FLOPs, due to cross-feeding encoders and logit-level aggregation, respectively.
Sampling Study. Fig. 6c visualizes the sampling probabilities q (in Eq. (16)) across various modality combinations on the CASIA-SURF dataset [29]. During the first 20 epochs, sampling is performed uniformly for all combinations. After activating the proposed discrepancy-aware sampling strategy (Sec. 3.5), the IR-only configuration exhibits the highest probability ($\sim$0.43), indicating a large discrepancy d between the feature and residual routers resulting in more emphasis during training. This aligns with the observation in Sec. 4.4 that prior methods performed poorly under the IR-only setting, while our model achieved significant improvement. The next highest probabilities are observed for the RGB-only and RGB+IR combinations, confirming that the sampling mechanism prioritizes modality configurations requiring further refinement.
Additional rationale for the residual is in Sec. A (supplementary material).

6 Conclusion

In this paper, we present MARS, a MoE framework that specializes experts by modeling missingness-induced representation deviations. By comparing full- and partial-modality representations, the residual router exploits residual cues that reflect how missingness affects task reasoning, specializing experts for the resulting deviations. The dual-router design and noise regularization together bridge the train–test routing gap, enabling benefits of residual-inspired routing even at inference. Furthermore, the proposed sampling balances training across modality configurations. Extensive incomplete multimodal experiments across domains and tasks demonstrate consistent superiority with minimal computational overhead. Thus, jointly leveraging observation and missingness offers a principled path toward robust multimodal systems under test-time incompleteness.

Acknowledgements

This research was supported by RS-2025-02216257 (65%), RS-2022-II220290 (30%), and RS-2019-II1091906 (AI Graduate Program at POSTECH, 5%).

References

1. Baek, S., Choi, I., et al.: Learning covariance-based multi-scale representation of neuroimaging measures for alzheimer classification. In: ISBI. pp. 1–5. IEEE (2023)
2. Baek, S., Sim, J., Wu, G., Kim, W.H.: Ocl: Ordinal contrastive learning for imputating features with progressive labels. In: MICCAI. pp. 334–344. Springer (2024)
3. Cai, L., Wang, Z., Gao, H., et al.: Deep adversarial learning for multi-modality missing data completion. In: ACM SIGKDD. pp. 1158–1166 (2018)
4. Cao, H., Cooper, D.G., Keutmann, M.K., et al.: Crema-d: Crowd-sourced emotional multimodal actors dataset. IEEE transactions on affective computing **5**(4), 377–390 (2014)
5. Chen, L.C., Zhu, Y., Papandreou, G., et al.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV. pp. 801–818 (2018)
6. Chen, Z., Li, H., Wang, F., et al.: Rethinking the diffusion models for missing data imputation: A gradient flow perspective. NeurIPS **37**, 112050–112103 (2024)
7. Ding, Y., Yu, X., Yang, Y.: Rfnet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation. In: ICCV. pp. 3975–3984 (2021)
8. Gallo, I., Ria, G., Landro, N., La Grassa, R.: Image and text fusion for upmc food-101 using bert and cnns. In: 2020 35th International conference on image and vision computing New Zealand (IVCNZ). pp. 1–6. IEEE (2020)
9. Gómez-Chova, L., Tuia, D., Moser, G., et al.: Multimodal classification of remote sensing images: A review and future directions. Proceedings of the IEEE **103**(9), 1560–1584 (2015)
10. Havaei, M., Guizard, N., Chapados, N., Bengio, Y.: Hemis: Hetero-modal image segmentation. In: MICCAI. pp. 469–477. Springer (2016)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
12. Li, S., Chen, C., Han, J.: Simmlm: A simple framework for multi-modal learning with missing modality. In: ICCV (2025)
13. Li, Y., Daho, M.E.H., Conze, P.H., Zeghlache, R., Le Boité, H., Tadayoni, R., Cochener, B., Lamard, M., Quellec, G.: A review of deep learning-based information fusion techniques for multimodal medical image classification. Computers in Biology and Medicine **177**, 108635 (2024)
14. Liang, Y., Wakaki, R., Nobuhara, S., et al.: Multimodal material segmentation. In: CVPR. pp. 19800–19808 (2022)
15. Ma, M., Ren, J., Zhao, L., et al.: Are multimodal transformers robust to missing modality? In: CVPR. pp. 18177–18186 (2022)
16. Ma, X., Zhang, X., Pun, M.O., et al.: A multilevel multimodal fusion transformer for remote sensing semantic segmentation. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–15 (2024)
17. Ruder, S.: An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747 (2016)
18. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., et al.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV. pp. 618–626 (2017)

19. Shazeer, N., Mirhoseini, A., Maziarz, K., et al.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. ICLR (2017)
20. Sleeman IV, W.C., Kapoor, R., Ghosh, P.: Multimodal classification: Current landscape, taxonomy and future directions. ACM Computing Surveys **55**(7), 1–31 (2022)
21. Sun, Y., Sheng, D., Zhou, Z., et al.: Ai hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content. Humanities and Social Sciences Communications **11**(1), 1–14 (2024)
22. Tivnan, M., Yoon, S., Chen, Z., et al.: Hallucination index: An image quality metric for generative reconstruction models. In: MICCAI. pp. 449–458. Springer (2024)
23. Wang, H., Chen, Y., Ma, C., et al.: Multi-modal learning with missing modality via shared-specific feature modelling. In: CVPR. pp. 15878–15887 (2023)
24. Wei, S., Luo, C., Luo, Y.: Mmanet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning. In: CVPR. pp. 20039–20049 (2023)
25. Wei, S., Luo, Y., Wang, Y., et al.: Robust multimodal learning via representation decoupling. In: ECCV. pp. 38–54. Springer (2024)
26. Wu, R., Wang, H., Chen, H.T., et al.: Deep multimodal learning with missing modality: A survey. ACM Computing Surveys (2024)
27. Xu, W., Jiang, H., Liang, X.: Leveraging knowledge of modality experts for incomplete multimodal learning. In: ACM MM. pp. 438–446 (2024)
28. Yun, S., Choi, I., Peng, J., et al.: Flex-moe: Modeling arbitrary modality combination via the flexible mixture-of-experts. NeurIPS **37**, 98782–98805 (2024)
29. Zhang, S., Wang, X., Liu, A., et al.: A dataset and benchmark for large-scale multi-modal face anti-spoofing. In: CVPR. pp. 919–928 (2019)
30. Zhang, Y., He, N., Yang, J., et al.: mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In: MICCAI. pp. 107–117. Springer (2022)
31. Zhang, Y., Sidibé, D., Morel, O., et al.: Deep multimodal fusion for semantic image segmentation: A survey. Image and Vision Computing **105**, 104042 (2021)
32. Zheng, X., Tang, C., Wan, Z., et al.: Multi-level confidence learning for trustworthy multimodal classification. In: AAAI. vol. 37, pp. 11381–11389 (2023)
33. Zhou, T., Canu, S., Vera, P., et al.: Brain tumor segmentation with missing modalities via latent multi-source correlation representation. In: MICCAI. pp. 533–541. Springer (2020)