Mitigating Pervasive Modality Absence Through Multimodal Generalization and Refinement

Wuliang Huang[1,2,4], **Yiqiang Chen**[1,2,3,4*], **Xinlong Jiang**[1,2,4*], **Chenlong Gao**[1,2,4], **Teng Zhang**[1,2,4], **Qian Chen**[1,2,4], **Yifan Wang**[5]

[1]Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]Peng Cheng Laboratory
[4]Beijing Key Laboratory of Mobile Computing and Pervasive Device
[5]Tsinghua Shenzhen International Graduate School, Tsinghua University
{huangwuliang19b, yqchen, jiangxinlong, gaochenlong, zhangteng19s, chenqian20b}@ict.ac.cn,
yifan-wa22@mails.tsinghua.edu.cn

Abstract

The performance of multimodal models often deteriorates when modality absence occurs. The absence disrupts the learned inter-modal correlations, resulting in biased multimodal representations. This challenge is especially pronounced when the absence is pervasive, affecting both the training and inference phases. Recent studies have attempted to reconstruct the missing information; however, most of them require complete supervision, which is seldom available in scenarios of pervasive absence. The quality of reconstruction remains a critical issue. Alternatively, others aim to learn robust representations from the available modalities but the substantial variations and biases are not fully addressed. This paper introduces the Multimodal Generalization and Refinement (MGR) framework to mitigate the issue of pervasive modality absence. MGR begins by acquiring generalized multimodal representations and iteratively refines them to recognize and calibrate the biased representations. Initially, multimodal samples with absence are embedded through foundation models, and MGR integrates independent unimodal features to further enhance generalization. Additionally, a novel mixed-context prompt is adopted to identify biases in both features and correlations. A redistribution operation can then refine these biases through graph pooling, culminating in robust and calibrated multimodal representations, which are suitable for downstream tasks. Comprehensive experiments on four benchmark datasets demonstrate that the proposed MGR framework outperforms state-of-the-art methods, effectively mitigating the impact of pervasive modality absence.

Introduction

Multi-modality fusion models exploit the complementary from various modalities to enhance the performance of downstream tasks (Fukui et al. 2016; Cai et al. 2020). Each modality conveys specific information, and their integration yields comprehensive and generalized representations (Baltrušaitis, Ahuja, and Morency 2019; Liang et al. 2022). Typically, multi-modality models work under the *assumption* that all modalities are complete (Lee et al. 2023). However, this assumption is frequently violated in real-world scenarios due to privacy concerns, noise, or data collection limitations (Zhang et al. 2022; Karimijafarbigloo et al. 2023; Yang et al. 2020), leading to performance degradation in the absence of certain modalities (Tran et al. 2017). Specifically, when the absence of modalities is pervasive, affecting both the training and inference phases, it becomes extremely difficult to learn unbiased and robust multimodal representations (Lee et al. 2023; Jang, Wang, and Kim 2024).

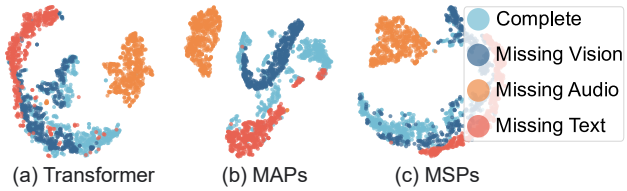


Figure 1: The distributions of multimodal representations derived from identical samples exhibit substantial variations depending on the type of modality absence. Here, MAPs (Lee et al. 2023) and MSPs (Jang, Wang, and Kim 2024) are two latest methods designed for mitigating absence.

Principal Challenges. As depicted in Figure 1, the distributions of multimodal embeddings derived from identical samples vary significantly when different modalities are missing. This observation proves that the absence of modalities can lead to biased representations. The biases brought by incomplete modalities increase the likelihood of suboptimal or erroneous predictions, obstructing the model from learning the intrinsic correlations between modalities. Therefore, it is essential to develop generalized multimodal representation, and refine biases among them to mitigate the impact of modality absence.

*Yiqiang Chen and Xinlong Jiang are corresponding authors.

Existing Approaches. To address the issue of absence, recent studies have attempted to compensate for missing modality through the remaining data (Tran et al. 2017). However, most of the reconstruction requires supervision (Dong et al. 2023; Wang, Li, and Cui 2023; Liu et al. 2024), which is difficult to adopt in scenarios with pervasive modality absence. Moreover, the quality of reconstruction determines the subsequent fusion process, leading to suboptimal performance when there is a significant gap between the missing and available modalities (Sun et al. 2024).

Alternatively, latest studies suggest learning robust representations from the remaining data through methods such as distillation (Li, Yang, and Zhang 2023), multi-task learning (Ma et al. 2022), and prompt-based approaches (Lee et al. 2023; Jang, Wang, and Kim 2024). Among these, prompt-based methods have demonstrated promising results and show compatibility with the recent advancements in Transformer-based multimodal foundation models (Vaswani et al. 2017; Kim, Son, and Kim 2021). Nonetheless, these approaches do not entirely address the substantial variations and biases. Such residual biases cumulatively impact the fusion process. Additionally, the prompts may not accurately fit the current complex contexts, thereby degrading performance and generalization.

Proposed Solution. Towards the challenge of pervasive modality absence, this paper introduces the Multimodal Generalization and Refinement (MGR) framework. Compared to existing methods, MGR can effectively recognize biases in multimodal representations and further calibrate them to yield robust and unbiased multimodal representations. Within MGR, an initial generalized multimodal representation is acquired, followed by iterative refinement processes that incorporate mixed-context prompts to identify biases in both features and correlations, and calibrate these biases via graph pooling-based redistribution.

Initially, MGR utilizes the multimodal foundation model (Kim, Son, and Kim 2021) to acquire multimodal representations. The missing modality is replaced with learnable embeddings, which can potentially be imputed through high-order inter-modal interactions. Given that the representation may be substantially weakened by absence, MGR integrates independent unimodal features to enhance generalization. To derive robust features, MGR employs mixed-context prompts, generated based on absence status and global context, to accurately estimate feature and correlation biases. Rather than simply discarding the biased tokens, which could result in information loss, MGR redistributes the reliable information to neighboring tokens, thereby preserving reliable information. The refinement process is iterative and culminates in robust and unbiased multimodal representations, optimally suited for downstream tasks.

Contributions. The primary contributions are as follows:

(1) We address the prevalent issue of modality absence and propose a Multimodal Generalization and Refinement (MGR) framework. This framework initially derives generalized representations and subsequently refines them to calibrate biases, thereby yielding robust and unbiased multimodal representations.

(2) We develop a novel mixed-context prompt that integrates multiple contexts to precisely estimate multimodal features and correlation biases. Additionally, we introduced a graph-based redistribution operation to refine biases while preserving reliable information.

(3) Extensive experiments conducted on four benchmark datasets demonstrate that the MGR framework significantly mitigates the impact of pervasive modality absence.

Related Work

Recent studies towards absence can be broadly categorized into imputation-based and joint learning-based strategies.

Modality Imputation Strategies for Absence

This category aims to reconstruct missing information by utilizing statistical properties derived from the remaining data (Tran et al. 2017; Suo et al. 2019; Zhao et al. 2024). Generative methods such as generative adversarial networks (Shang et al. 2017; Qin et al. 2022; Cao et al. 2020), variational autoencoders (Xu et al. 2024), and diffusion models (Wang, Li, and Cui 2023), have been employed for this purpose. Recent research specifically focuses on learning the translation function between missing and available modalities. Dong et al. (2023) propose to regularize and generate the missing modality via translation. Similarly, Liu et al. (2024) approximate the missing modality. Among these methods, most recovery operations require supervision, such as the ground truth of the missing modality. However, in scenarios where modality absence is pervasive, such supervision is rarely available. Additionally, the synthesis process may introduce noise or non-converging characteristics (Karimijafarbigloo et al. 2023). To better accommodate pervasive absence, we leverage multimodal Transformers (Kim, Son, and Kim 2021) to implicitly impute the missing data through high-order interactions. A further refinement process is then introduced to calibrate the biases if the imputation is unreliable, thereby ensuring the final robust and unbiased representations for downstream tasks.

Joint Learning Strategies for Absence

Joint learning-based strategies exploit the interaction between modalities (Akbari et al. 2021), learning robust representations from the available data. Ma et al. (2022) are pioneers in investigating the robustness of Transformers against incomplete modalities, proposing multi-task optimization to enhance robustness. Yang et al. (2024) present a two-step training procedure aimed at improving robustness by addressing the modality preference issue. Recently, prompt-based methods (Lee et al. 2023; Jang, Wang, and Kim 2024) have also been proposed, since they are compatible with the latest advanced multimodal foundation models (Vaswani et al. 2017; Kim, Son, and Kim 2021). Lee et al. (2023) design modality-missing-aware prompts and plug them into Transformers to handle missing cases. Jang, Wang, and Kim (2024) optimize these prompts further by leveraging orthogonality to promote diverse representations. However, one remaining challenge is that these methods do not fully address the significant variations between multimodal tokens. These

residual variations and biases cumulatively impact the subsequent fusion process, degrading performance and generalization. To mitigate this issue, we propose to identify and explicitly refine the biases, via prompts and pooling-based redistribution operation, thereby enhancing the robustness and generalization of multimodal representations.

Methodology of The MGR Framework

To tackle the challenge of pervasive modality absence during both training and inference phases, this paper introduces the Multimodal Generalization and Refinement (MGR) framework. This section provides a comprehensive introduction, following the illustration in Figure 2.

Preliminaries

We denote a multimodal sample as $(\mathcal{M}, \mathcal{B})$, where $\mathcal{M}$ represents the available modalities, and $\mathcal{B}$ denotes the absent modalities. $\mathcal{B} = \varnothing$ signifies that all modalities are available. In this paper, we specifically consider three modalities: $\{V, A, T\}$, representing vision, audio, and text, respectively. The raw input data of each modality is denoted as $\mathbf{X}^i$, where $i \in \{V, A, T\}$. The input shapes may vary due to the intrinsic nature of each modality. The goal of MGR is to learn robust multimodal representations $\mathbf{z} = M_{\text{MGR}}(\mathcal{M}, \mathcal{B})$ for each sample, which can be effectively utilized for downstream tasks under pervasive modality absence scenarios.

Multimodal Representation Generalization

The MGR framework initially devises to learn comprehensive multimodal representations, enhancing the generalization of the model to mitigate the impact of modality absence. The process is primarily accomplished by multimodal Transformers foundation (Kim, Son, and Kim 2021). We embed the available modalities through tokenization and leverage multimodal Transformers to capture the multimodal features. Subsequently, we integrate unimodal features to enhance the robustness of the representations.

Uncertainty-Aware Multimodal Embedding. The available raw inputs are first embedded by modality-specific tokenizers f_ϑ^i to generate tokens $\mathbf{x}^i \in \mathbb{R}^{N_i \times d}$, whereas the absent modalities are substituted with learnable embeddings $\mathbf{E}_{\text{miss}}^i \in \mathbb{R}^{N_i \times d}$. Here, d denotes the predefined hidden size, N_i represents the number of tokens of the i-th modality, and $N = \sum_{i \in \{V,A,T\}} N_i$ is the total number of tokens.

For $i \in \{V, A, T\}$, the tokenization process is defined as:

$$\mathbf{x}^i = \begin{cases} f_\vartheta^i(\mathbf{X}^i) + \mathbf{E}_{\text{avail}}^i, & \text{if } i \in \mathcal{M} \quad \text{(Availability)}, \\ \mathbf{E}_{\text{miss}}^i, & \text{if } i \in \mathcal{B} \quad \text{(Absence)}, \end{cases} \tag{1}$$

where $\mathbf{E}_{\text{avail}}^i \in \mathbb{R}^{N_i \times d}$ serves as an indicator to differentiate available modalities. A multimodal Transformer model, instantiated by ViLT (Kim, Son, and Kim 2021), is then employed to capture the multimodal features $\mathbf{h} \in \mathbb{R}^{N \times d}$:

$$\mathbf{h} = \text{Transformer}(\mathbf{x}), \quad \mathbf{x} = \left[\mathbf{x}^V \middle\| \mathbf{x}^A \middle\| \mathbf{x}^T\right]_{\text{dim}=0}. \tag{2}$$

The operation $[\cdot\|\cdot]_{\text{dim}}$ here indicates concatenation along the specified dimension. The two learnable embeddings, $\mathbf{E}_{\text{miss}}^i$ and $\mathbf{E}_{\text{avail}}^i$, are capable of effectively perceiving the uncertainty associated with the absence of modalities. The representations of each modality, denoted as $\mathbf{h}^V, \mathbf{h}^A, \mathbf{h}^T$, can subsequently be derived by appropriately slicing $\mathbf{h}$.

Unimodal Generalization Enhancement. The output of the multimodal Transformer model, $\mathbf{h}$, is intended to serve as a comprehensive multimodal representation. Nonetheless, it may exhibit inherent biases. For example, the imputation of absent modalities might lack precision, or the available modalities could be compromised by potential unreliability. To mitigate these issues and improve generalization, MGR integrates additional unimodal features $\mathbf{h}_u$ for available modalities:

$$\mathbf{h}_e^i = \begin{cases} \mathbf{h}^i + \alpha_u \mathbf{h}_u^i, & \text{if } i \in \mathcal{M} \quad \text{(Availability)}, \\ \mathbf{h}^i, & \text{if } i \in \mathcal{B} \quad \text{(Absence)}. \end{cases} \tag{3}$$

The factor α_u is a hyperparameter. In the above equation, $\mathbf{h}_u^i$ is derived from modality-specific unimodal Transformers, which share a similar architecture as Equation (2). As the inputs to unimodal Transformers are exclusively unimodal features, with no absences involved, the unimodal features, $\mathbf{h}_u^i$, are anticipated to be more reliable and robust. The integration further leads to enhancing the representations to $\mathbf{h}_e^i$.

As discussed in previous studies (Albuquerque et al. 2019; Lu et al. 2023; Huang et al. 2024a), the diversity between representations is anticipated to be significant for generalization. Therefore, we introduce distance correlation (Székely, Rizzo, and Bakirov 2007; Zhen et al. 2022; Huang et al. 2024b) to quantify the dependencies between multimodal and unimodal features without assuming linearity or normality:

$$\mathcal{L}_{\text{div}} = \sum_{i \in \mathcal{M}} \frac{\mathcal{V}_{n_i}^2\left(\mathbf{h}_u^i, \mathbf{h}^i\right)}{\sqrt{\mathcal{V}_{n_i}^2\left(\mathbf{h}_u^i, \mathbf{h}_u^i\right) \mathcal{V}_{n_i}^2\left(\mathbf{h}^i, \mathbf{h}^i\right)} + \epsilon}, \tag{4}$$

where $\mathcal{V}_n^2(\mathbf{h}_u, \mathbf{h})$ is the empirical distance covariance, while $\mathcal{V}_n^2(\mathbf{h}_u, \mathbf{h}_u)$, $\mathcal{V}_n^2(\mathbf{h}, \mathbf{h})$ denote the empirical variances. $\mathcal{L}_{\text{div}}$ is computed for available modalities exclusively.

Potential Biases Identification and Refinement

After acquiring generalized multimodal representations $\mathbf{h}$, MGR iteratively refines these representations to recalibrate biases, thereby preserving only the reliable and robust information against modality absence. This refinement initiates by precisely perceiving the current contexts, thus can recognize biases within features or correlations. The identified biases are refined through a redistribution operation.

Contexts Perception with Mixed Prompts. To comprehensively estimate biases in multimodal representations, MGR employs prompts as conditional guidance. Recent advancements leverage prompts to navigate Transformer models in circumstances of missing modalities (Lee et al. 2023; Jang, Wang, and Kim 2024). Nonetheless, the prompts employed in these approaches focus predominantly on the absence status, potentially failing to adequately align with the existing contexts, such as feature or correlation dynamics.

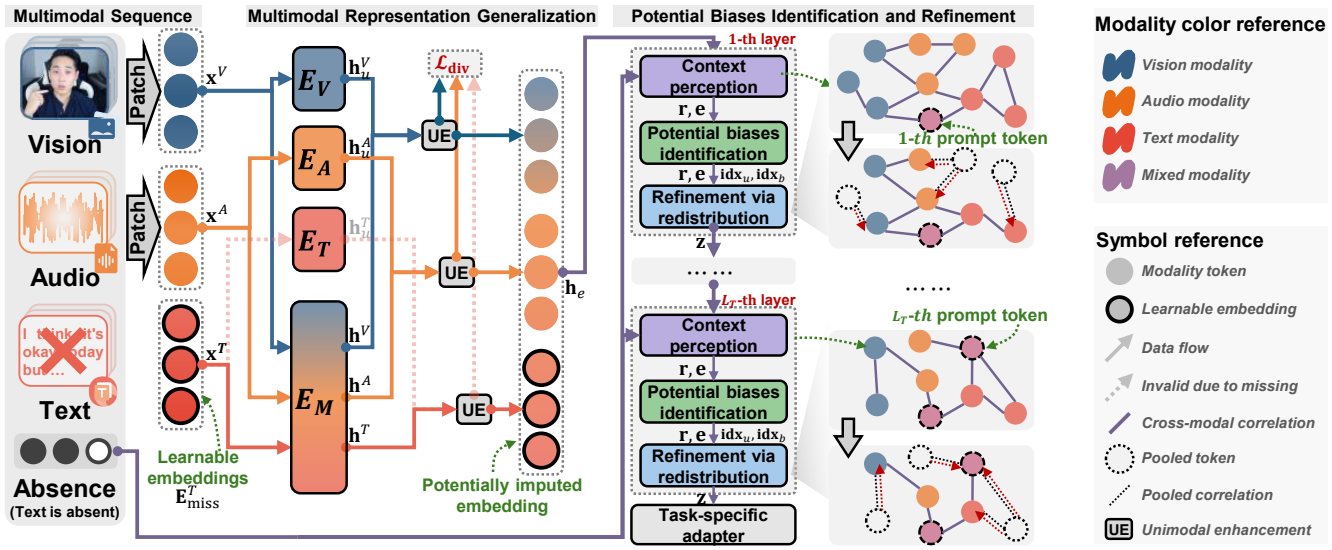


Figure 2: Illustration of the proposed Multimodal Generalization and Refinement (MGR) framework.

To mitigate these limitations, MGR enhances prompt quality by incorporating diverse contextual information, generating the mixed-context prompt $\mathbf{P} \in \mathbb{R}^d$:

$$\mathbf{P} = f_{\text{Mix}}\left(\left[\sum_{i\in\mathcal{M}} \mathbf{p}^i \middle\| f_{\mathbf{\Phi}}(\mathbf{h}_e)\right]_{\text{dim}=0}\right). \tag{5}$$

Here, $\mathbf{p}^i \in \mathbb{R}^d$ represents the learnable modality-specific prompts indicating absence, with $\sum_{i\in\mathcal{M}} \mathbf{p}^i$ consolidating the contextual information pertinent to the current absence status. Furthermore, $f_{\mathbf{\Phi}}(\mathbf{h}_e)$ denotes the context of features, where $f_{\mathbf{\Phi}}$ functions as a feature and correlation estimator implemented through average pooling. The function f_{Mix} is a learnable transformation for context mixing.

MGR can obtain conditionally enhanced information $\mathbf{r} = P(\mathbf{h}_e|\mathbf{P})$ via the mixed-context prompt $\mathbf{P}$. The prompt is integrated into $\mathbf{h}_e$ as a new prompt token:

$$\mathbf{h}_p = [\mathbf{h}_e \| \mathbf{P}]_{\text{dim}=0}. \tag{6}$$

Additional methods for integrating prompts into models, such as summation, input, and attention integration (Lee et al. 2023), are discussed in Experiments section.

Subsequently, the conditionally enhanced feature $\mathbf{r}$ is obtained through attention mechanism with n_h heads:

$$\mathbf{r} = [\mathbf{r}_1 \| \dots \| \mathbf{r}_n]_{\text{dim}=1}\mathbf{W}^O, \quad \mathbf{r}_i = \mathbf{a}_i\mathbf{h}_p\mathbf{W}_i^V,$$
$$\text{where } \mathbf{a}_i = \text{Softmax}\left(\frac{\mathbf{h}\mathbf{W}_i^Q(\mathbf{h}\mathbf{W}_i^K)^T}{\sqrt{d/n_h}}\right), \tag{7}$$

where $\mathbf{W}_i^Q$, $\mathbf{W}_i^K$, and $\mathbf{W}_i^V$ represent learnable weights for the i-th head. $\mathbf{W}^O$ is the output linear layer. Furthermore, the perceived correlation $\mathbf{e} \in \mathbb{R}^{N\times N}$ between tokens are derived from the attention weights $\mathbf{a}_i$:

$$\mathbf{e} = \frac{1}{n_h}\sum_{i=1}^{n_h} \mathbf{a}_i. \tag{8}$$

Potential Biases Identification. Based on the estimated enhanced representation $\mathbf{r}$ and correlations $\mathbf{e}$ within diverse contexts, we can identify potential biases through ranking:

$$\boldsymbol{\xi} = \alpha_f \boldsymbol{\xi}_f + (1-\alpha_f)\boldsymbol{\xi}_e,$$
$$\boldsymbol{\xi}_f = \text{Tanh}\left(f_{\mathbf{\Psi}}(\mathbf{r})\right), \quad \boldsymbol{\xi}_e = \text{Tanh}\left(\sum_{i=1}^{N}\mathbf{e}_i\right). \tag{9}$$

Here, $\boldsymbol{\xi} \in \mathbb{R}^N$ represents the ranking scores of tokens, where higher values indicate a lower likelihood of bias. The learnable projector $f_{\mathbf{\Psi}}$ realizes ranking at feature-level, $\boldsymbol{\xi}_f$, while the correlation-level ranking $\boldsymbol{\xi}_e$ is derived from the sum of context correlations $\mathbf{e}_i$, which reflect biases inherent in correlation dynamics. The parameter α_f is a weight factor that balances and aggregates the dual rankings.

The top-ρ tokens, which are considered highly likely to be unbiased, are retained, while the rest are discarded:

$$\text{idx}_u = \text{ArgSort}(\boldsymbol{\xi})_{1:\rho|\xi|}, \text{idx}_b = \text{ArgSort}(\boldsymbol{\xi})_{\rho|\xi|:N}, \tag{10}$$

where idx_u and idx_b represent the indices of unbiased and biased tokens, respectively.

Refinement through Redistribution. While it is common to discard biased tokens by simply omitting them, this approach can lead to information loss due to uncertainties in index splitting and the potential reliable information that biased tokens may contain. To address this issue, we propose a graph-based redistribution method to preserve the useful information embedded within biased tokens.

Considering a graph where each vertex represents a token, let $\mathbf{r}_i$ be the feature of the i-th vertex, the adjacency matrix $\mathbf{A}$, which governs the information passing flow between tokens, is derived from the correlation matrix $\mathbf{e}$ given in Equation 8:

$$\mathbf{A}_{ij} = \begin{cases} \mathbf{e}_{ij}, & \text{if } i \in \text{idx}_b \text{ and } j \in \text{idx}_u, \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

Method	CMU–MOSI F1 Score	CMU–MOSI Accuracy	CMU–MOSEI F1 Score	CMU–MOSEI Accuracy	UR–FUNNY F1 Score	UR–FUNNY Accuracy	AV–MNIST F1 Score	Average (4 Datasets)
Baseline ICML'21	66.43 /2.26	67.36 /2.68	82.20 /0.85	73.97 /1.06	58.79 /3.50	61.42 /1.04	53.43 /5.11	66.23
MTL CVPR'22	66.45 /2.00	70.68 /0.95	82.64 /0.81	73.78 /1.63	60.65 /2.81	61.99 /1.18	56.20 /1.88	67.48
MaskMentor MM'24	67.16 /3.04	68.11 /6.77	82.97 /0.19	72.10 /1.03	59.09 /3.22	61.31 /1.86	56.25 /3.43	66.71
MaPLe CVPR'23	67.74 /2.05	68.33 /1.81	82.63 /0.76	74.47 /0.87	60.23 /2.84	63.21 /0.99	56.07 /2.75	67.53
MAPs CVPR'23	68.14 /3.12	70.15 /2.58	84.90 /0.77	76.85 /0.91	60.40 /1.69	62.93 /0.74	55.76 /4.87	68.45
MSPs ICASSP'24	69.05 /2.00	**71.55** /1.67	83.38 /1.63	75.22 /1.85	60.34 /2.50	62.84 /1.42	56.98 /1.45	68.48
MGR (Ours)	**69.23** /2.29	71.07 /1.84	**85.27** /0.45	**77.48** /0.51	**60.93** /1.92	**63.78** /0.69	**58.15** /0.37	**69.42**

Table 1: Comparison results under a absence prevalence ratio $\kappa = 0.7$.

Actually, $\mathbf{A}$ is a submatrix of $\mathbf{e}$, establishing bipartite connections between biased and unbiased tokens. The redistribution operation propagates reliable information as formulated by (Morris et al. 2019), utilizing high-order interactions to achieve the refined representation $\mathbf{z}$:

$$\mathbf{z}_i = \sigma\left(\mathbf{r}_i\mathbf{W}_{c1} + \sum_{j=1}^{N}\mathbf{r}_j\mathbf{W}_{c2}\mathbf{A}_{ji}\right), \tag{12}$$

where $\mathbf{W}_{c1}$ and $\mathbf{W}_{c2}$ are learnable transformations, and σ is activation. In contrast to mere omission, the redistribution operation ensures the preservation of valuable information from biased tokens. The entire refinement process, as described in Equations (5)–(12), is iteratively executed over L_T iterations, thereby ensuring smooth convergence. Consequently, only ρ^{L_T} percent tokens, conveying merged information, are preserved. This approach not only enhances robustness but also accelerates computational efficiency.

The refined representation $\mathbf{z}$ obtained for each sample is subsequently utilized for downstream tasks. In the context of classification or regression, the final prediction is derived as $\mathcal{Y}' = f_{\text{task}}(\mathbf{z})$, where f_{task} represents the task-specific classifier or regressor. The entire MGR framework is trained in an end-to-end manner using the loss function $\mathcal{L}$:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}}, \tag{13}$$

Here, $\mathcal{L}_{\text{task}}$ denotes the task-specific loss, such as mean absolute error for regression or cross-entropy for classification. The parameter λ_{div} is a hyperparameter that regularizes the diversity loss $\mathcal{L}_{\text{div}}$ as defined in Equation (4).

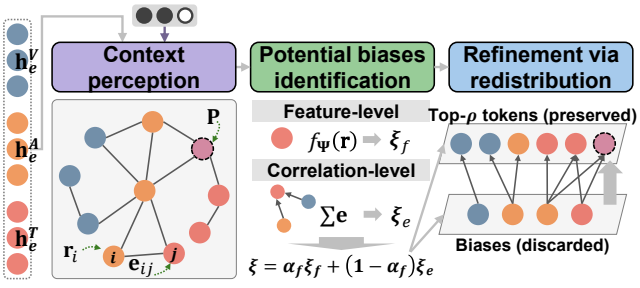


Figure 3: Illustration of the graph-based redistribution operation for biased token refinement.

Experiments

Experimental Setup

Benchmark Datasets. Four multimodal datasets are adopted to evaluate the MGR framework. CMU–MOSI (Zadeh et al. 2016), CMU–MOSEI (Zadeh et al. 2018; Liang et al. 2018), and UR–FUNNY (Hasan et al. 2019) are video-based datasets encompassing three modalities (V, A, T). CMU–MOSEI and UR–FUNNY are among the largest datasets within their respective domains. AV–MNIST (Liang et al. 2021) is a synthetic noisy dataset with two modalities (V, A).

Implementation Details. The data preprocessing and the partitioning of training, validation, and testing sets follows MultiBench (Liang et al. 2021). The framework is implemented using PyTorch (Paszke et al. 2019). The AdamW optimizer (Loshchilov and Hutter 2017) with a learning rate of $1e$-4 is utilized. Regarding modality absence, the model is evaluated at two prevalence ratios, $\kappa = 0.7$ and $\kappa = 1$, which indicate the percentage of samples with one modality absence in training, validation, and testing phases. $\kappa = 1$ signifies that all samples exhibit one modality absence. Each experiment is conducted using four random seeds for the absence schema and trained under four random seeds, resulting in each experiment being replicated 16 times to ensure reliable results. The average performance and standard deviation across these 16 runs are reported.

Baselines. The proposed MGR framework is compared with several state-of-the-art methods of top-tier publications. These methods include training strategies designed for absence: MTL (Ma et al. 2022) and MaskMentor (Zhao et al. 2024). Additionally, latest prompt-based methods, MaPLe (Khattak et al. 2023), MAPs (Lee et al. 2023), and MSPs (Jang, Wang, and Kim 2024), are also evaluated. The ViLT (Kim, Son, and Kim 2021) model serves as the baseline.

Comparison Results

We compare the proposed MGR framework to the aforementioned baselines. The results are presented in Table 1 and Table 2. The performance is reported as avg./std. over 16 runs.

The results demonstrate that the MGR framework consistently outperforms the baselines in both the absence preva-

Method	CMU–MOSI		CMU–MOSEI		UR–FUNNY		Average
	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	(3 Datasets)
Baseline ICML'21	65.46 /2.20	66.13 /2.45	81.92 /1.44	73.11 /1.52	58.15 /2.06	60.91 /1.02	67.61
MTL CVPR'22	63.16 /2.40	68.22 /1.19	82.10 /0.62	73.19 /1.37	59.09 /5.32	59.37 /1.11	67.52
MaskMentor MM'24	64.74 /2.87	67.19 /1.11	82.60 /0.10	70.84 /0.16	57.70 /3.35	61.73 /0.81	67.47
MaPLe CVPR'23	66.45 /1.75	65.74 /1.62	82.04 /0.99	73.49 /1.50	59.37 /3.88	62.45 /0.78	68.26
MAPs CVPR'23	65.52 /2.51	67.24 /1.72	84.34 /0.75	76.08 /0.82	60.43 /2.18	62.42 /1.31	69.34
MSPs ICASSP'24	66.73 /1.90	68.61 /1.62	82.46 /1.65	73.77 /2.07	60.07 /3.51	62.00 /1.43	68.94
MGR (Ours)	**67.46** /1.86	**68.99** /1.68	**85.02** /0.31	**76.53** /0.31	**60.68** /2.26	**62.89** /0.74	**70.26**

* Due to the fact that AV–MNIST only contains two modalities, results for the scenario where all samples have modality absence are not reported since multimodal fusion is not applicable.

Table 2: Comparison results under a absence prevalence ratio $\kappa = 1$.

lence ratio $\kappa = 0.7$ and $\kappa = 1$, underscoring the effectiveness of the proposed method in addressing the impact of pervasive modality absence. When the absence ratio κ increases from 0.7 to 1, the performance of all methods decreases. As depicted in the tables, prompt-based methods, including MaPLe, MAPs, and MSPs, leverage prompts to handle modality absence and achieve competitive performance compared to other designs. Among these prompt-based methods, the proposed MGR achieves the best performance, demonstrating the efficacy of the proposed mixed-context prompt and the graph-based refinement operation. The explicit refinement of biases and redistribution of information from biased tokens are crucial for enhancing the robustness of multimodal models under modality absence.

Ablation Studies and Sensitivity Analysis

Method	Absence Type (V / A / T)	$\kappa = 0.7$		$\kappa = 1$	
		F1	Acc.	F1	Acc.
MAPs	○/○/●	85.11	78.61	84.90	78.52
MSPs		84.63	77.93	83.94	77.44
Ours		**86.63**	**80.00**	**86.12**	**79.63**
MAPs	○/●/○	83.40	74.72	81.24	71.23
MSPs		82.78	74.16	80.48	71.05
Ours		**84.52**	**76.25**	**83.82**	**74.53**
MAPs	●/○/○	84.20	75.50	83.42	74.08
MSPs		83.95	75.34	83.18	73.66
Ours		**84.67**	**76.03**	**83.79**	**74.52**

* A hollow circle signifies the corresponding modality is absence.

Table 3: Comparison when two modalities exhibit absence.

Ablation Studies on Modality Absence Types. The absence is observed across all modalities in the previous experiments (Table 1 and 2). This section further evaluates the performance of MGR under various conditions of modality absence. Table 3 and Table 4 correspond to scenarios where

Method	Absence Type (V / A / T)	$\kappa = 0.7$	
		F1 Score	Accuracy
MAPs	○/●/●	83.00 /1.75	76.55 /1.72
MSPs		84.19 /1.00	77.70 /1.05
Ours		**86.22** /0.63	**79.48** /0.59
MAPs	●/○/●	86.23 /0.52	79.55 /0.57
MSPs		85.97 /0.55	79.32 /0.82
Ours		**86.63** /0.26	**80.03** /0.39
MAPs	●/●/○	82.62 /0.63	71.37 /0.79
MSPs		82.57 /0.63	71.22 /0.64
Ours		**82.62** /0.57	**72.46** /0.64

* The results are reported exclusively for the scenario where $\kappa = 0.7$, since the issue of modality absence is not relevant when $\kappa = 1$ and only one modality exhibits absent.

Table 4: Comparison when one modality exhibits absence.

the absence occurs only in one or two modalities. Performances are reported on the CMU-MOSEI dataset, the largest among the benchmark datasets. It is demonstrated that MGR outperforms baseline models across different conditions of modality absence, thereby attesting to its robustness and effectiveness in addressing this issue.

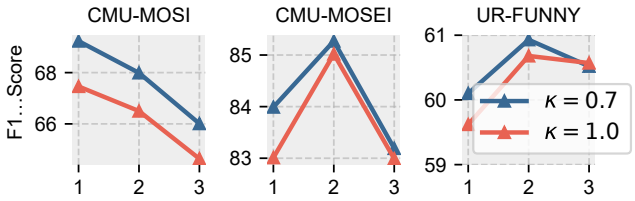
	Prompt type	MOSI	MOSEI	FUNNY	AV-M	Avg.
MGR +	MAPs/Input	67.86	85.10	**61.14**	57.80	67.98
	MAPs/Attn.	68.43	**85.50**	61.07	**58.15**	68.29
	Summation	67.91	85.22	60.65	58.04	67.96
	Token	**69.23**	85.27	60.93	**58.15**	**68.40**

Table 5: Ablation on prompt insertion types when $\kappa = 0.7$.

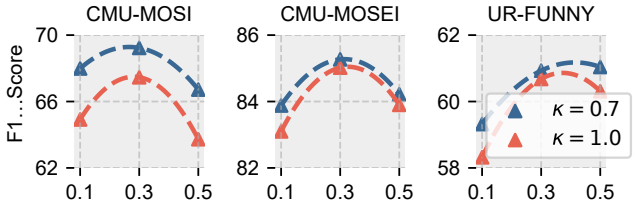
Ablation Studies on Prompt Insertion Types. As introduced in the methodology section, we inject the mixed-context prompt $\mathbf{P}$ into the multimodal Transformer model

	Prompt type	MOSI	MOSEI	FUNNY	Avg.
MGR +	MAPs/Input	65.18	84.29	60.32	69.93
	MAPs/Attn.	65.74	84.86	**61.21**	70.60
	Summation	66.01	**85.11**	60.56	70.56
	Token	**67.46**	85.02	60.68	**71.05**

Table 6: Ablation on prompt insertion types when $\kappa = 1$.



(a) Sensitivity of the number of refinement layers L_T.



(b) Sensitivity of the ρ in each refinement layer.

Figure 4: Sensitive analysis of hyperparameters.

at the token level by adding a prompt token. Here, we examine the impact of different types of prompt insertion, including input, attention, summation, and token-level insertion. The results are presented in Table 5 and Table 6. The input and attention-level prompt insertion types are derived from MAPs (Lee et al. 2023).The summation refers to the summing of the prompt with inputs. The adopted token-level prompt insertion has delivered the best performance on average and achieved most of the top-two performances across all datasets. It proves to be a general and effective choice.

Method	MOSI	MOSEI	FUNNY	AV-M	Avg.
Ours	**69.23**	**85.27**	**60.93**	**58.15**	**68.40**
w/o mixed-context	66.92	83.44	60.56	57.71	67.01
w/o $\mathcal{L}_{\text{div}}$	67.09	83.53	60.47	55.37	66.62
w/o $\mathbf{h}_u$	66.86	83.73	59.76	53.73	66.02

Table 7: Ablation on different components when $\kappa = 0.7$.

Ablation Studies on Components. We conducted ablation studies to evaluate the impact of various components within MGR in Table 7 and Table 8. Initially, we replaced the mixed-context prompt $\mathbf{P}$ with the prompts proposed in MAPs (Lee et al. 2023). The absence of context led to a marked decline in performance, indicating that the mixed-context prompt effectively captures the current dynamics. Furthermore, the removal of $\mathcal{L}_{\text{div}}$ and unbiased and stable unimodal features $\mathbf{h}_u$ result in performance drop.

Method	MOSI	MOSEI	FUNNY	Avg.
Ours	**67.46**	**85.02**	**60.68**	**71.05**
w/o mixed-context	66.34	82.87	60.43	69.88
w/o $\mathcal{L}_{\text{div}}$	65.32	82.69	60.17	69.39
w/o $\mathbf{h}_u$	65.42	82.61	59.57	69.20

Table 8: Ablation on different components when $\kappa = 1$.

Sensitivity Analysis. For several critical hyperparameters, specifically the number of refinement layers L_T and the setting of ρ in each layer, we conducted a sensitivity analysis to evaluate their impact on performance. The results are presented in Figure 4. Based on these findings, we can determine the final choices for the hyperparameters.

Qualitative Analysis

Figure 5 presents the distribution of multi-modality token using t-SNE (Van der Maaten and Hinton 2008). In these plots, missing modalities are imputed by the multimodal Transformer. The visualization reveals that our proposed MGR effectively learns coherent multimodal representations, exhibiting a lower inter-modality heterogeneity gap compared to MSPs. Specifically, in scenarios where the audio modality is absent, our MGR achieves a more compact multimodal representation than MSPs. Additionally, in the absence of the vision modality, the imputed visual tokens and the crossmodal correlations learned by MGR closely approximate those in a complete scenario.

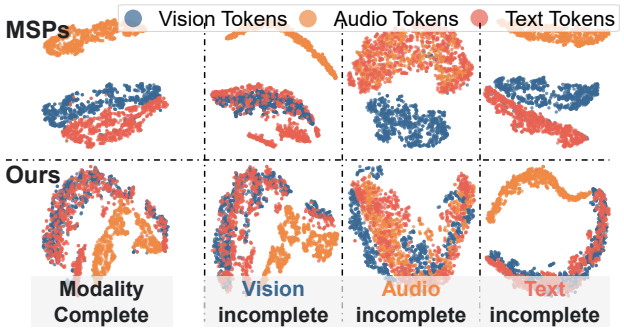


Figure 5: Visualization of multimodal token distributions.

Conclusion

We address the primary challenge of pervasive modality absence by introducing a Multimodal Generalization and Refinement (MGR) framework. A novel mixed-context prompt is introduced to estimate multimodal correlations, and redistribution operations are employed to refine biased tokens. Comprehensive experiments conducted on four benchmark datasets demonstrate that the proposed framework surpasses state-of-the-art methods. In future work, we will explore the potential applications of MGR in other multimodal models.

Acknowledgments

This work was supported by the National Key Research and Development Plan of China (No. 2023YFC3604805), the National Natural Science Foundation of China (No. 62472406), the Youth Innovation Promotion Association CAS, the Innovation Funding of ICT, CAS (No. E463050), the Postdoctoral Fellowship Program of CPSF (No. GZC20232737), and the Science and Technology Innovation Program of Hunan Province (No. 2022RC4006).

References

Akbari, H.; Yuan, L.; Qian, R.; Chuang, W.-H.; Chang, S.-F.; Cui, Y.; and Gong, B. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34: 24206–24221.

Albuquerque, I.; Monteiro, J.; Falk, T. H.; and Mitliagkas, I. 2019. Adversarial target-invariant representation learning for domain generalization. *arXiv preprint arXiv:1911.00804*, 8(2): 13.

Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443.

Cai, H.; Qu, Z.; Li, Z.; Zhang, Y.; Hu, X.; and Hu, B. 2020. Feature-Level Fusion Approaches Based on Multimodal EEG Data for Depression Recognition. *Information Fusion*, 59: 127–138.

Cao, B.; Zhang, H.; Wang, N.; Gao, X.; and Shen, D. 2020. Auto-GAN: self-supervised collaborative learning for medical image synthesis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 10486–10493.

Dong, H.; Nejjar, I.; Sun, H.; Chatzi, E.; and Fink, O. 2023. SimMMDG: A Simple and Effective Framework for Multi-Modal Domain Generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Fukui, A.; Park, D. H.; Yang, D.; Rohrbach, A.; Darrell, T.; and Rohrbach, M. 2016. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 457–468. Austin, Texas: Association for Computational Linguistics.

Hasan, M. K.; Rahman, W.; Bagher Zadeh, A.; Zhong, J.; Tanveer, M. I.; Morency, L.-P.; and Hoque, M. E. 2019. UR-FUNNY: A Multimodal Language Dataset for Understanding Humor. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2046–2056. Hong Kong, China: Association for Computational Linguistics.

Huang, W.; Chen, Y.; Jiang, X.; Gao, C.; Chen, Q.; Zhang, T.; Yan, B.; Wang, Y.; and Yang, J. 2024a. Correlation-Driven Multi-Modality Graph Decomposition for Cross-Subject Emotion Recognition. In *ACM Multimedia*.

Huang, W.; Chen, Y.; Jiang, X.; Gao, C.; Chen, Q.; Zhang, T.; Yan, B.; Wang, Y.; and Yang, J. 2024b. Correlation-Driven Multi-Modality Graph Decomposition for Cross-Subject Emotion Recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, 2272–2281. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-0686-8.

Jang, J.; Wang, Y.; and Kim, C. 2024. Towards Robust Multimodal Prompting With Missing Modalities. ICASSP 2024:arXiv:2312.15890.

Karimijafarbigloo, S.; Azad, R.; Kazerouni, A.; Ebadollahi, S.; and Merhof, D. 2023. MMCFormer: Missing Modality Compensation Transformer for Brain Tumor Segmentation. In *Medical Imaging with Deep Learning*.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023. MaPLe: Multi-Modal Prompt Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19113–19122.

Kim, W.; Son, B.; and Kim, I. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 5583–5594. PMLR.

Lee, Y.-L.; Tsai, Y.-H.; Chiu, W.-C.; and Lee, C.-Y. 2023. Multimodal Prompting with Missing Modalities for Visual Recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14943–14952. Vancouver, BC, Canada: IEEE. ISBN 9798350301298.

Li, M.; Yang, D.; and Zhang, L. 2023. Towards Robust Multimodal Sentiment Analysis Under Uncertain Signal Missing. *IEEE Signal Processing Letters*, 30: 1497–1501.

Liang, P. P.; Liu, Z.; Bagher Zadeh, A.; and Morency, L.-P. 2018. Multimodal Language Analysis with Recurrent Multistage Fusion. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 150–161. Brussels, Belgium: Association for Computational Linguistics.

Liang, P. P.; Lyu, Y.; Fan, X.; Wu, Z.; Cheng, Y.; Wu, J.; Chen, L. Y.; Wu, P.; Lee, M. A.; Zhu, Y.; et al. 2021. MultiBench: Multiscale Benchmarks for Multimodal Representation Learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Liang, X.; Qian, Y.; Guo, Q.; Cheng, H.; and Liang, J. 2022. AF: An Association-Based Fusion Method for Multi-Modal Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 9236–9254.

Liu, Z.; Zhou, B.; Chu, D.; Sun, Y.; and Meng, L. 2024. Modality Translation-Based Multimodal Sentiment Analysis under Uncertain Missing Modalities. *Information Fusion*, 101: 101973.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lu, W.; Wang, W.; Yidong, J.; and Xie, X. 2023. Towards optimization and model selection for domain generalization: A mixup-guided solution. In *The KDD'23 Workshop on Causal Discovery, Prediction and Decision*, 75–97. PMLR.

Ma, M.; Ren, J.; Zhao, L.; Testuggine, D.; and Peng, X. 2022. Are Multimodal Transformers Robust to Missing Modality? In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18156–18165.

Morris, C.; Ritzert, M.; Fey, M.; Hamilton, W. L.; Lenssen, J. E.; Rattan, G.; and Grohe, M. 2019. Weisfeiler and Leman Go Neural: Higher-Order Graph Neural Networks. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19, 4602–4609. Honolulu, Hawaii, USA: AAAI Press. ISBN 978-1-57735-809-1.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Qin, Z.; Liu, Z.; Zhu, P.; and Ling, W. 2022. Style transfer in conditional GANs for cross-modality synthesis of brain magnetic resonance images. *Computers in Biology and Medicine*, 148: 105928.

Shang, C.; Palmer, A.; Sun, J.; Chen, K.-S.; Lu, J.; and Bi, J. 2017. VIGAN: Missing View Imputation with Generative Adversarial Networks. In *2017 IEEE International Conference on Big Data (Big Data)*, 766–775.

Sun, Y.; Liu, Z.; Sheng, Q. Z.; Chu, D.; Yu, J.; and Sun, H. 2024. Similar Modality Completion-Based Multimodal Sentiment Analysis under Uncertain Missing Modalities. *Information Fusion*, 110: 102454.

Suo, Q.; Zhong, W.; Ma, F.; Yuan, Y.; Gao, J.; and Zhang, A. 2019. Metric learning on healthcare data with incomplete modalities. In *IJCAI*, volume 3534, 3540.

Székely, G. J.; Rizzo, M. L.; and Bakirov, N. K. 2007. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6): 2769 – 2794.

Tran, L.; Liu, X.; Zhou, J.; and Jin, R. 2017. Missing Modalities Imputation via Cascaded Residual Autoencoder. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4971–4980.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention Is All You Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30*, 5998–6008. Curran Associates, Inc.

Wang, Y.; Li, Y.; and Cui, Z. 2023. Incomplete Multimodality-Diffused Emotion Recognition. *Advances in Neural Information Processing Systems*, 36: 17117–17128.

Xu, Z.; Wang, T.; Liu, D.; Hu, D.; Zeng, H.; and Cao, J. 2024. Audio-Visual Cross-Modal Generation with Multimodal Variational Generative Model. In *2024 IEEE International Symposium on Circuits and Systems (ISCAS)*, 1–5.

Yang, X.; Chen, Y.; Yu, H.; Zhang, Y.; Lu, W.; and Sun, R. 2020. Instance-Wise Dynamic Sensor Selection for Human Activity Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01): 1104–1111.

Yang, Z.; Wei, Y.; Liang, C.; and Hu, D. 2024. Quantifying and Enhancing Multi-modal Robustness with Modality Preference. In *The Twelfth International Conference on Learning Representations*. ICLR 2024.

Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L.-P. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

Zadeh, A. B.; Liang, P. P.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2236–2246.

Zhang, C.; Chu, X.; Ma, L.; Zhu, Y.; Wang, Y.; Wang, J.; and Zhao, J. 2022. M3Care: Learning with Missing Modalities in Multimodal Healthcare Data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, 2418–2428. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9385-0.

Zhao, Z.; Li, J.; Wang, L.; Wang, Y.; and Lu, H. 2024. MaskMentor: Unlocking the Potential of Masked Self-Teaching for Missing Modality RGB-D Semantic Segmentation. In *ACM Multimedia 2024*.

Zhen, X.; Meng, Z.; Chakraborty, R.; and Singh, V. 2022. On the versatile uses of partial distance correlation in deep learning. In *European Conference on Computer Vision*, 327–346. Springer.