# A Unified Self-Distillation Framework for Multimodal Sentiment Analysis with Uncertain Missing Modalities

**Mingcheng Li**[1,2*], **Dingkang Yang**[1,2*], **Yuxuan Lei**[1], **Shunli Wang**[1], **Shuaibing Wang**[1], **Liuzhen Su**[1], **Kun Yang**[1], **Yuzheng Wang**[1], **Mingyang Sun**[1], **Lihua Zhang**[1,2,3†]

[1]Academy for Engineering and Technology, Fudan University
[2]Cognition and Intelligent Technology Laboratory (CIT Lab)
[3]Engineering Research Center of AI and Robotics, Ministry of Education, Shanghai, China
mingchengli21@m.fudan.edu.cn, {dkyang20, lihuazhang}@fudan.edu.cn

## Abstract

Multimodal Sentiment Analysis (MSA) has attracted widespread research attention recently. Most MSA studies are based on the assumption of modality completeness. However, many inevitable factors in real-world scenarios lead to uncertain missing modalities, which invalidate the fixed multimodal fusion approaches. To this end, we propose a **U**nified multimodal **M**issing modality self-**D**istillation **F**ramework (UMDF) to handle the problem of uncertain missing modalities in MSA. Specifically, a unified self-distillation mechanism in UMDF drives a single network to automatically learn robust inherent representations from the consistent distribution of multimodal data. Moreover, we present a multi-grained crossmodal interaction module to deeply mine the complementary semantics among modalities through coarse- and fine-grained crossmodal attention. Eventually, a dynamic feature integration module is introduced to enhance the beneficial semantics in incomplete modalities while filtering the redundant information therein to obtain a refined and robust multimodal representation. Comprehensive experiments on three datasets demonstrate that our framework significantly improves MSA performance under both uncertain missing-modality and complete-modality testing conditions.

## Introduction

As an important part of human-computer interaction, Multimodal Sentiment Analysis (MSA) is becoming a hot research area, which aims to understand and interpret human sentiments through multiple forms of human expressions (*e.g.*, language content, voice tone, and facial behavior). Previous studies have shown that more effective and valuable joint multimodal representations can be obtained by combining complementary features in different modalities (Shraga et al. 2020; Springstein, Müller-Budack, and Ewerth 2021), benefiting from the evolution of learning-based techniques (Yang et al. 2023c; Chen et al. 2024; Li, Yang, and Zhang 2023; Yang et al. 2023d). Most MSA works (Hazarika, Zimmermann, and Poria 2020; Yu et al. 2021; Yang et al. 2022a,d, 2023b, 2022b; Li, Wang, and Cui 2023) are based on the assumptions that all modalities are available during the training and testing phases. In real applications, the assumption will not hold due to many inevitable factors, such as privacy, device, or security constraints, resulting in significant degradation of model performance.

Recently, several efforts focus on solving the problem of uncertain missing modalities in MSA, which can be broadly divided into two groups: **(i)** joint learning methods (Pham et al. 2019; Wang et al. 2020; Zhao, Li, and Jin 2021; Zeng, Liu, and Zhou 2022; Liu et al. 2024), which try to learn integrated representations based on the relations among different modalities; **(ii)** generative methods (Du et al. 2018; Ma et al. 2021; Luo, Xu, and Lai 2023; Lian et al. 2023), which reconstruct missing modalities utilizing available modalities. However, these methods suffer from the following limitations: 1) only perform interactions between fixed modality missing cases and fail to address stochastic real-world scenarios; 2) focus only on coarse-grained and localized interactions in missing modalities, leading to non-robust joint representations and invalid elemental correlations; 3) ignore redundant semantics in multimodal representations, resulting in performance bottlenecks.

To address the above issues, we propose a **U**nified multimodal **M**issing modality self-**D**istillation **F**ramework (UMDF) for the MSA task under uncertain missing modalities. UMDF has the following three core contributions. (**i**) We design a unified self-distillation mechanism in UMDF to automatically learn robust inherent representations from the consistent distribution of multimodal data representations by bidirectional knowledge transfer within a single network. The bidirectional knowledge transfer pathway can supervise the model to maintain similar feature distributions and logits distributions between heterogeneous modality missing cases. This effective pathway inhibits the unidirectional reliance on the learned features (Morcos et al. 2018) and is beneficial in two ways: the knowledge transfer from more to fewer modalities facilitates the recovery of lost information of the missing modalities, while in the opposite direction, enhances modality-specific features. (**ii**) We propose a multi-grained crossmodal interaction module that progressively performs coarse- and fine-grained crossmodal attention on missing modalities. It can hierarchically capture the inter-modality interactions and intra-modality dynamics to complement and reproduce the semantics of missing ele-

*These authors contributed equally.
†Corresponding author.

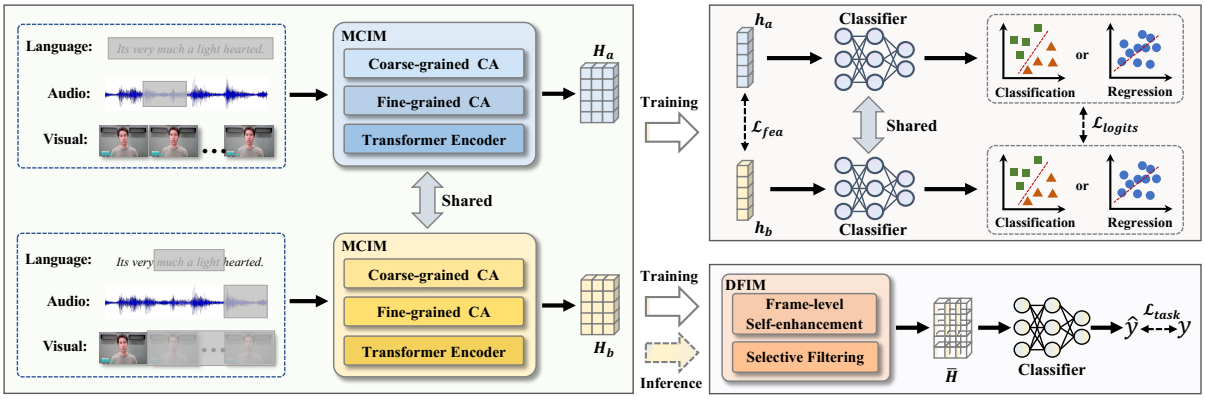


Figure 1: The overall framework of the proposed UMDF, consists of core components: the unified self-distillation mechanism, multi-grained crossmodal interaction module, and dynamic feature integration module.

ments of modalities. (**iii**) We introduce a dynamic feature integration module to further enhance the beneficial semantics and filter the redundant features through frame-level self-enhancement and selective filtering strategies to yield more refined representations. Based on these components, UMDF significantly improves the MSA performance under uncertain missing-modality and complete-modality testing conditions on three multimodal benchmarks.

## Related Work

### Multimodal Sentiment Analysis

MSA is a multimodal task that perceives and processes heterogeneous data, such as language, audio, and visual, to understand and analyze human sentiments (Yang et al. 2022c, 2023a; Lei et al. 2023). Mainstream works (Zadeh et al. 2017, 2018a; Tsai et al. 2019; Hazarika, Zimmermann, and Poria 2020; Han, Chen, and Poria 2021) enhanced the MSA performance by designing complex structures, interaction mechanisms, or fusion paradigms. For instance, MMIM (Han, Chen, and Poria 2021) improved multimodal fusion efficiency by hierarchically maximizing mutual information in unimodal input pairs. However, these methods are based on the assumption of complete data and cannot be applied to missing modality scenarios. Recently, several works (Tran et al. 2017; Pham et al. 2019; Zhao, Li, and Jin 2021; Zeng, Liu, and Zhou 2022; Zeng, Zhou, and Liu 2022) have focused on solving the missing modality problem in MSA. For instance, TATE (Zeng, Liu, and Zhou 2022) presented a tag encoding module to guide the network to focus on missing modalities. However, the modality missing samples during training in the above methods are fixed and cannot be generalized to complex situations in real-world applications. In contrast, we randomly generate two heterogeneous modality missing versions for each sample in the training process.

### Knowledge Distillation

Knowledge distillation utilizes additional supervised information from pre-trained teacher models to assist in training student models (Hinton, Vinyals, and Dean 2015). For multimodal tasks with missing modalities, many studies transfer drak knowledge from the complete-modality teacher network to the missing-modality student network through co-training (Cho et al. 2021; Hu et al. 2020; Wang et al. 2021). Despite the promising results achieved by these methods, some limitations remain: 1) during co-training, the teacher network incurs a non-negligible memory overhead; 2) there is only fixed unidirectional supervision from the complete modalities to the missing modalities, failing to exploit and transfer the beneficial common semantics shared by the different missing-modality situations. To this end, we propose a unified self-distillation mechanism that drives a single network to learn more valuable joint multimodal representations bidirectionally from heterogeneous modality missing versions of the samples with low overhead.

## Methodology

### Framework Overview

Given a multimodal video segment with three modalities as $\boldsymbol{S} = [\boldsymbol{X}_L, \boldsymbol{X}_A, \boldsymbol{X}_V]$, where $\boldsymbol{X}_L \in \mathbb{R}^{T_L \times d_L}$, $\boldsymbol{X}_A \in \mathbb{R}^{T_A \times d_A}$, and $\boldsymbol{X}_V \in \mathbb{R}^{T_V \times d_V}$ denote language, audio, and visual modalities, respectively. $T_m$ is the sequence length and $d_m$ is the embedding dimension, where $m \in \{L, A, V\}$. The incomplete modality is denoted as $\boldsymbol{X}'_m$. We define two missing modality cases to simulate the holistic challenges in real-world scenarios: 1) intra-modality missingness, which indicates that some frame-level features in the modality sequence are missing; 2) inter-modality missingness, which denotes some modalities are entirely missing. Our goal is to recognize the utterance-level sentiments by utilizing the multimodal data with missing modalities.

As shown in Figure 1, the main workflow of the proposed UMDF is as follows: given a video segment sample $\boldsymbol{S} = [\boldsymbol{X}_L, \boldsymbol{X}_A, \boldsymbol{X}_V]$, two heterogeneous modality missing versions $\boldsymbol{S}_a$ and $\boldsymbol{S}_b$ are generated. $\boldsymbol{S}_a$ and $\boldsymbol{S}_b$ are fed into the multi-grained crossmodal interaction module to obtain the joint multimodal representations $\boldsymbol{H}_a$ and $\boldsymbol{H}_b$. Then, these two multimodal representations go through two branches: 1) achieving consistent supervision at the feature-level and logits-level through a self-distillation mechanism to adequately learn robust inherent representations among modal-

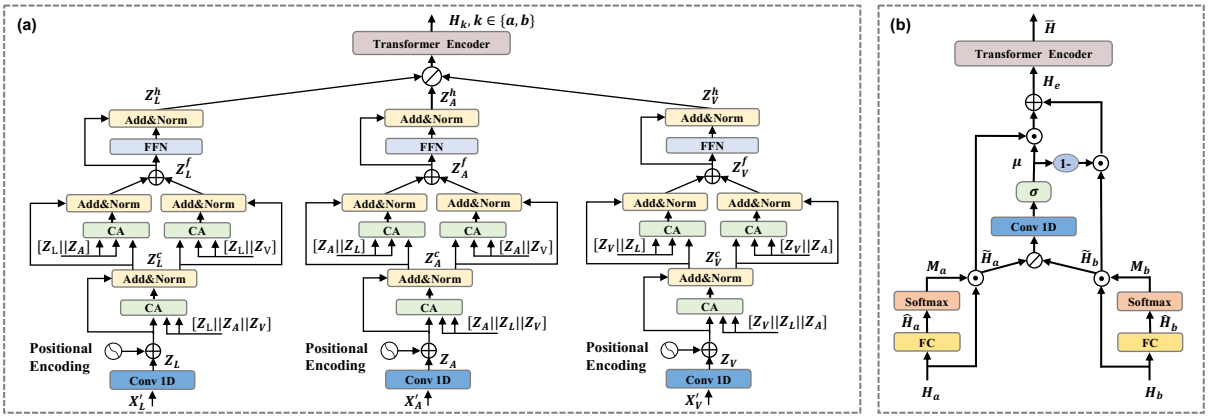


Figure 2: The architecture of the proposed (a) MCIM and (b) DFIM. FFN represents the feed-forward layer. $\odot$ denotes the Hadamard product. $\oplus$ denotes the element-wise sum. $\oslash$ denotes the concatenation operation. $\sigma$ denotes the sigmoid activation.

ities; 2) using a dynamic feature integration module to enhance and filter features to obtain a refined representation $\overline{\boldsymbol{H}}$. Ultimately, $\overline{\boldsymbol{H}}$ is fed into the task-specific fully connected layers to implement the sentiment prediction. In the inference phase, we clone a copy of the testing sample and use them together as a two-stream input to the model for multimodal sentiment analysis.

## Unified Self-distillation Mechanism

Traditional knowledge distillation approaches for modality missing aim to supervise the learning of missing-modality student networks with complete-modality teacher networks. It suffers from multiple limitations, such as high-performance requirements of the teacher network, expensive training costs, and fixed information transfer direction (Anil et al. 2018; Cho et al. 2021; Hu et al. 2020; Wang et al. 2021). Self-distillation is a fully-supervised pattern that exploits the potential capability of a single network from labeled data only without auxiliary models. The proposed unified self-distillation mechanism transfers knowledge bidirectionally between heterogeneous modality-missing versions of samples within a single network via soft labels. The soft labels contain more complete structured information than the ground-truth labels. The mutual knowledge transfer guides the model to enhance available semantics and restore missing semantics to yield more valuable inherent representations. In practice, for each mini-batch, we randomly generate two heterogeneous modality-missing versions (including the modality-complete case) based on each sample therein and encourage them to obtain consistent semantic features (*i.e.*, feature distribution consistency and logits distribution consistency) through the shared network.

**Feature Distillation.** We adopt Maximum Mean Discrepancy (MMD) (Gretton et al. 2012) as a nonparametric metric to measure the discrepancy between two feature distributions, *i.e.*, $\boldsymbol{H}_a$ and $\boldsymbol{H}_b$. MMD has been widely used in domain adaptation to estimate the discrepancy between two domains (Long et al. 2015), and it has good robustness and efficiency in computation and optimization. MMD is a kernel two-sample test that accepts or rejects the null hypothesis $\boldsymbol{p} = \boldsymbol{q}$ depending on the observed samples. Formally, the MMD defines the following discrepancy measures:

$$\mathcal{D}_{\mathcal{H}}(\boldsymbol{p}, \boldsymbol{q}) \triangleq \left\| \mathbf{E}_p \left[ \phi \left( \boldsymbol{S}_a \right) \right] - \mathbf{E}_q \left[ \phi \left( \boldsymbol{S}_b \right) \right] \right\|_{\mathcal{H}}^2, \tag{1}$$

where $\mathcal{H}$ is the Reproducing Kernel Hilbert Space (RKHS) with characteristic kernel $k$. The kernel $k$ is represented as $k(\boldsymbol{S}_a, \boldsymbol{S}_b) = \langle \phi(\boldsymbol{S}_a), \phi(\boldsymbol{S}_b) \rangle$, where $\phi(\cdot)$ denotes some feature mapping that maps the original samples to RKHS. The core theory of MMD is that hypothesis $\boldsymbol{p} = \boldsymbol{q}$ holds when and only when $\mathcal{D}_{\mathcal{H}}(\boldsymbol{p}, \boldsymbol{q}) = 0$. In practice, the unbiased estimator of MMD can be computed as:

$$\hat{\mathcal{D}}_{\mathcal{H}}(\boldsymbol{p}, \boldsymbol{q}) = \left\| \frac{1}{n} \sum_{i=1}^{n} \phi \left( \boldsymbol{S}_a \right) - \frac{1}{n} \sum_{i=1}^{n} \phi \left( \boldsymbol{S}_b \right) \right\|_{\mathcal{H}}^2, \tag{2}$$

where $n$ is the number of samples in a mini-batch. The feature distillation loss is represented as:

$$\mathcal{L}_{fea} = \frac{1}{n} \sum_{i=1}^{n} \hat{\mathcal{D}}_{\mathcal{H}}(\boldsymbol{h}_a, \boldsymbol{h}_b), \tag{3}$$

where $\boldsymbol{h}_a$ and $\boldsymbol{h}_b$ denote the last elements of joint multimodal representations $\boldsymbol{H}_a$ and $\boldsymbol{H}_b$.

**Logits Distillation.** To minimize the discrepancy in the distribution between both logits, we construct soft labels to supervise the learning. Notably, our proposed unified self-distillation mechanism can be applied to both regression and classification tasks with high scalability.

For the classification task, we use the Jensen-Shanno (JS) divergence as a discrepancy measure, which solves the problem of Kullback-Leible (KL) divergence asymmetry and can better measure the discrepancy between distributions. The KL divergence is represented as:

$$\mathcal{D}_{KL} \left( \boldsymbol{p}_b \| \boldsymbol{p}_a \right) = -\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{p} \left( \boldsymbol{x}_b \right) \log \frac{\boldsymbol{p} \left( \boldsymbol{x}_a \right)}{\boldsymbol{p} \left( \boldsymbol{x}_b \right)}, \tag{4}$$

where $\boldsymbol{p}_b$ is the target probabilities as soft labels to supervise the learning of the predicted probabilities $\boldsymbol{p}_a$. $f_y$ is a

fully-connected layer for computing the logits. The logits distillation loss is denoted as:

$$\begin{aligned}\mathcal{L}_{logits} &= \mathcal{D}_{JS}(f_y(\boldsymbol{h}_a)||f_y(\boldsymbol{h}_b)),\\ &= \frac{1}{2}(\mathcal{D}_{KL}(f_y(\boldsymbol{h}_a)||\boldsymbol{M})) + \mathcal{D}_{KL}(f_y(\boldsymbol{h}_b)||\boldsymbol{M})),\end{aligned} \tag{5}$$

where $\boldsymbol{M}$ is the average distribution of $f_y(\boldsymbol{h}_a)$ and $f_y(\boldsymbol{h}_b)$.

For the regression task, the Mean Square Error (MSE) is applied to estimate the discrepancy between both logits, which facilitates network convergence and penalizes large prediction errors. The logits distillation loss is expressed as:

$$\mathcal{L}_{logits} = \mathcal{D}_{MSE} = \frac{1}{n}\sum_{i=1}^{n}(f_y(\boldsymbol{h}_a) - f_y(\boldsymbol{h}_b))^2. \tag{6}$$

## Multi-grained Crossmodal Interaction Module

The modality heterogeneity normally leads to the distribution gap and information redundancy in multimodal fusion, resulting in task-irrelevant semantic and ambiguous multimodal joint representations. Additionally, the modality missing poses a greater challenge to the fusion and modeling of multimodal sequences. Although previous studies have made some advances in MSA with missing modalities (Pham et al. 2019; Wang et al. 2020), they only consider pairwise directional interactions among independent modalities, leading to non-robust joint representations and invalid elemental correlations. To tackle this problem, we propose a Multi-grained Crossmodal Interaction Module (MCIM) to thoroughly explore natural correlations among elements of missing modalities by simultaneously modeling inter-modality interactions and intra-modality dynamics. Specifically, MCIM sequentially performs coarse- and fine-grained crossmodal interactions. This hierarchical interaction paradigm stimulates the potential of incomplete modalities to reconstruct missing semantics progressively.

Figure 2(a) illustrates the architecture of MCIM, consisting of several muti-head crossmodal attention layers, Layer-Normalization (LN) layers, and Feed-Forward Networks (FFN). We define the source modality as $\boldsymbol{X}_s$ with $s \in \{L, A, V\}$ and the target modality as $\boldsymbol{X}_t$ with $t \in \{L, A, V\}$. We embed the target modality as $\boldsymbol{Q}_t = \boldsymbol{X}_t\boldsymbol{W}_{\boldsymbol{Q}_t}$ with $\boldsymbol{W}_{\boldsymbol{Q}_t} \in \mathbb{R}^{d_t \times d_k}$, and the source modality as $\boldsymbol{K}_s = \boldsymbol{X}_s\boldsymbol{W}_{\boldsymbol{K}_s}$ with $\boldsymbol{W}_{\boldsymbol{K}_s} \in \mathbb{R}^{d_s \times d_k}$ and $\boldsymbol{V}_s = \boldsymbol{X}_s\boldsymbol{W}_{\boldsymbol{V}_s}$ with $\boldsymbol{W}_{\boldsymbol{V}_s} \in \mathbb{R}^{d_s \times d_v}$. The latent adaptation from $\boldsymbol{X}_s$ to $\boldsymbol{X}_t$ is presented as the Crossmodal Attention (CA):

$$\boldsymbol{X}_{s\to t} = \mathrm{CA}(\boldsymbol{X_s}, \boldsymbol{X_t}) = \mathrm{softmax}(\frac{\boldsymbol{X_t}\boldsymbol{W_{Q_t}}\boldsymbol{W}_{\boldsymbol{K}_s}^{\top}\boldsymbol{X}_s^{\top}}{\sqrt{d_k}})\boldsymbol{X}_s\boldsymbol{W}_{\boldsymbol{V}_s}, \tag{7}$$

where scaled softmax function computes the score matrix. Subsequently, the process of the forward computation is represented as $\boldsymbol{X}_t = \mathrm{LN}(\boldsymbol{X}_t + \boldsymbol{X}_{s\to t})$.

We introduce MCIM with the example of using language modality as the target modality. Firstly, we input the incomplete modality $\boldsymbol{X}'_m \in \mathbb{R}^{T_m \times d_m}$ with $m \in \{L, A, V\}$ into a 1D temporal convolutional layer with kernel size $3 \times 3$ to make them project to the same dimension, denoted as $\hat{\boldsymbol{X}}'_m = \boldsymbol{W}_{3\times3}(\boldsymbol{X}'_m)$. Then, we augment the positional embedding (Vaswani et al. 2017) to $\hat{\boldsymbol{X}}'_m$ to obtain the low-level representations $\boldsymbol{Z}_m \in \mathbb{R}^{T_m \times d}$. The representations $\boldsymbol{Z}_m$ are concatenated to obtain a coarse-grained representation $\boldsymbol{Z}_{LAV} = [\boldsymbol{Z}_L, \boldsymbol{Z}_A, \boldsymbol{Z}_V] \in \mathbb{R}^{(T_L+T_A+T_V)\times d}$, and two fine-grained representations as $\boldsymbol{Z}_{LA} = [\boldsymbol{Z}_L, \boldsymbol{Z}_A] \in \mathbb{R}^{(T_L+T_A)\times d}$ and $\boldsymbol{Z}_{LV} = [\boldsymbol{Z}_L, \boldsymbol{Z}_V] \in \mathbb{R}^{(T_L+T_V)\times d}$. The coarse- and fine-grained crossmodal interactions are denoted as:

$$\boldsymbol{Z}_L^c = \mathrm{LN}(\mathrm{CA}(\boldsymbol{Z}_{LAV}, \boldsymbol{Z}_L) + \boldsymbol{Z}_L), \tag{8}$$

$$\boldsymbol{Z}_L^f = \mathrm{LN}(\mathrm{CA}(\boldsymbol{Z}_{LA}, \boldsymbol{Z}_L^c) + \boldsymbol{Z}_L^c) + \mathrm{LN}(\mathrm{CA}(\boldsymbol{Z}_{LV}, \boldsymbol{Z}_L^c) + \boldsymbol{Z}_L^c). \tag{9}$$

The $\boldsymbol{K}_s$ and $\boldsymbol{V}_s$ of CA are splices of two modalities or three modalities, neither of which is a zero vector. Ultimately, $\boldsymbol{Z}_L^f$ is fed into a FFN $\mathcal{F}_\theta(\cdot)$ with the LN layer to obtain $\boldsymbol{Z}_L^h$, which is represented as $\boldsymbol{Z}_L^h = \mathrm{LN}(\mathcal{F}_\theta(\boldsymbol{Z}_L^f) + \boldsymbol{Z}_L^f$. In practice, we stack $D$-layers of MCIM to gradually complement and enrich the sentiment semantics of modal representations. Ultimately, $\boldsymbol{Z}_L^h$, $\boldsymbol{Z}_A^h$, and $\boldsymbol{Z}_V^h$ are concatenated and feed into a transformer encoder to achieve further interactions, yielding $\boldsymbol{H}_k \in \mathbb{R}^{T_m \times 3d}$ with $k \in \{a, b\}$.

## Dynamic Feature Integration Module

The modality missing blurs the valuable sentiment semantics of the samples, leading to redundant joint multimodal representations. To this end, we propose a Dynamic Feature Integration Module (DFIM) to achieve adaptive information integration of two heterogeneous missing modality representations. The core philosophy is retaining and enhancing the beneficial semantics in the incomplete modalities and filtering the redundant information.

From Figure 2(b), DFIM receives joint multimodal representations $\boldsymbol{H}_k \in \mathbb{R}^{T_m \times 3d}$ with $k \in \{a, b\}$ and $m \in \{L, A, V\}$ generated from heterogeneous modality missing versions of samples as inputs. Firstly, the frame-level self-enhancement strategy is utilized to enhance the joint multimodal representations, which includes the following steps: 1) dimension adjustment of $\boldsymbol{H}_k$ via a fully connected layer, denoted as $\hat{\boldsymbol{H}}_k = \mathcal{F}_k(\boldsymbol{H}_k; \boldsymbol{W}_\theta) \in \mathbb{R}^{T_m \times 1}$, where $\boldsymbol{W}_\theta$ are network parameters. 2) The softmax function is applied to $\hat{\boldsymbol{H}}_k$ to obtain the score matrix $\boldsymbol{M}_k$, denoted as $\boldsymbol{M}_k = \mathrm{softmax}(\hat{\boldsymbol{H}}_k)$, and its $i$-th entry denotes the importance of $i$-th frame of $\hat{\boldsymbol{H}}_k$. 3) Matrix multiplication of $\boldsymbol{M}_k$ with $\boldsymbol{H}_k$ yields the self-enhanced representation $\tilde{\boldsymbol{H}}_k$. Subsequently, a selective filtering mechanism is proposed to filter the redundant information in the joint multimodal representation while retaining the sentiment semantics, denoted as:

$$\mu = \sigma\left(\boldsymbol{W}_{3\times3}([\tilde{\boldsymbol{H}}_a; \tilde{\boldsymbol{H}}_b])\right), \tag{10}$$

$$\boldsymbol{H}_e = \mu \odot \tilde{\boldsymbol{H}}_a + (1-\mu) \odot \tilde{\boldsymbol{H}}_b, \tag{11}$$

where $\sigma$ is the sigmoid function. The integrated representation $\boldsymbol{H}_e$ feeds into a transformer encoder to achieve further interactions, yielding $\overline{\boldsymbol{H}}$. The last element of $\overline{\boldsymbol{H}}$ is used to make predictions to obtain $\hat{\boldsymbol{y}}$ through fully-connected layers.

**Training Objective** The overall training objective $\mathcal{L}_{total}$ is expressed as $\mathcal{L}_{total} = \mathcal{L}_{task} + \lambda_1\mathcal{L}_{fea} + \lambda_2\mathcal{L}_{logits}$, where $\mathcal{L}_{task}$ is the task loss, $\lambda_1$ and $\lambda_2$ are the corresponding weights. For the classification and regression tasks, we use cross-entropy and MSE loss as the task losses, respectively.

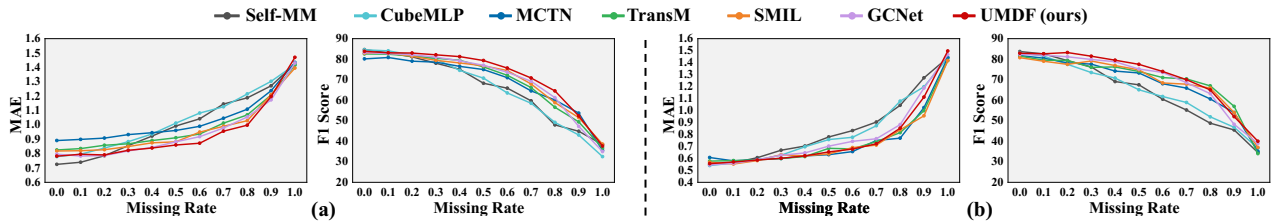


Figure 3: Comparison results of various missing rates on (a) MOSI and (b) MOSEI. We report the MAE and F1 score metrics.

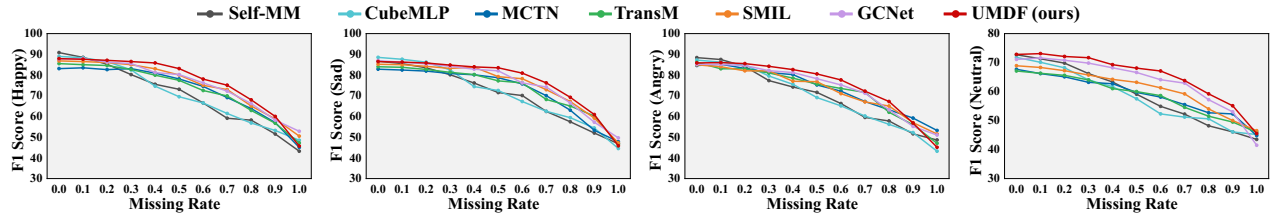


Figure 4: Comparison results of various missing rates on IEMOCAP. We comprehensively report the F1 score metrics for the happy, sad, angry, and neutral categories.

# Experiments

## Datasets and Evaluation Metrics

We conduct experiments on three standard MSA datasets, including MOSI (Zadeh et al. 2016), MOSEI (Zadeh et al. 2018b), and IEMOCAP (Busso et al. 2008). MOSI is a realistic dataset consisting of 2,199 opinion video clips. There are 1,284, 229, and 686 video clips in train, valid, and test data, respectively. MOSEI is a dataset made up of 22,856 movie review video clips, which has 16,326, 1,871, and 4,659 samples in train, valid, and test data. Each sample of MOSI and MOSEI is labeled by human annotators with a sentiment score of -3 (strongly negative) to +3 (strongly positive). On the MOSI and MOSEI datasets, we utilize two evaluation metrics, including the Mean Absolute Error (MAE) and F1 score computed for positive/negative classification results. The IEMOCAP dataset contains conversation videos. As recommended by (Wang et al. 2019), four emotions (*i.e.,* happy, sad, angry, and neutral) are selected for emotion recognition. The F1 score is used as the metric.

## Implementation Details

**Feature Extraction.** For the language modality, we convert the transcripts of the video into pre-trained Glove word embedding (Pennington, Socher, and Manning 2014) to obtain a 300-dimensional vector. For the audio modality, we employ the COVAREP toolkit (Degottex et al. 2014) to extract 74-dimensional low-level acoustic features, such as 12 Mel-frequency cepstral coefficients (MFCCs) and glottal source parameters. For the visual modality, we use the Facet (Baltrušaitis, Robinson, and Morency 2016) to indicate 35 facial action units, recording facial muscle movement to represent emotions.

**Experimental Setup.** All models are built on the Pytorch toolbox with NVIDIA Tesla V100 GPUs. The Adam optimizer (Kingma and Ba 2014) is employed for network optimization. For MOSI, MOSEI, and IEMOCAP, the detailed hyper-parameter settings are as follows: the batch sizes are $\{128, 16, 32\}$, the learning rates are $\{1e-3, 1e-3, 2e-3\}$, the epoch counts are $\{100, 30, 40\}$, and the attention heads are $\{10, 8, 10\}$. The feature size and MCIM layer count are 40 and 4 on all three datasets. The hyper-parameters are determined via the validation set. The raw features at the modality missing positions are replaced by zero vectors. For a fair comparison, we re-implement the SOTA methods based on the public codebase and combine them with our experimental paradigms. The results are the average of several experiments using five different random seeds.

## Comparison with State-of-the-Art Methods

We compare UMDF with six representative and reproducible state-of-the-art (SOTA) methods, including complete-modality methods: Self-MM (Yu et al. 2021) and CubeMLP (Sun et al. 2022), and missing-modality methods: 1) joint learning methods (*i.e.*, MCTN (Pham et al. 2019) and TransM (Wang et al. 2020)), and 2) generative methods (*i.e.*, SMIL (Ma et al. 2021) and GCNet (Lian et al. 2023)). The comprehensive experiments aim to thoroughly evaluate the robustness and effectiveness of UMDF in the cases of intra-modality and inter-modality missingness.

**Robustness to Intra-modality Missingness.** Here, we adopt a random missing strategy where frame-level features are randomly dropped with probability $p$ to implement the case of intra-modality missingness. Figures 3 and 4 show the model performance curves to visually evaluate the models' robustness. We have the following key observations. **(i)** The performance of all models decreases with increasing the missing rate $p$, which shows that intra-modality missingness loses numerous valuable sentiment semantics, leading to blurred joint multimodal representations. **(ii)** The superior performance of UMDF under complete-modality testing condition (*i.e.*, $p = 0$) is reflected in two aspects:

| Datasets | Models | Testing Conditions | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\{l\}$ | | $\{a\}$ | | $\{v\}$ | | $\{l,a\}$ | | $\{l,v\}$ | | $\{a,v\}$ | | Avg. | | $\{l,a,v\}$ | |
| | | MAE | F1 | MAE | F1 | MAE | F1 | MAE | F1 | MAE | F1 | MAE | F1 | MAE | F1 | MAE | F1 |
| MOSI | Self-MM (Yu et al. 2021) | 1.008 | 67.80 | 1.396 | 40.95 | 1.452 | 38.52 | 0.993 | 69.81 | 0.961 | 74.97 | 1.331 | 47.12 | 1.190 | 56.53 | **0.725** | **84.64** |
| | CubeMLP (Sun et al. 2022) | 1.036 | 64.15 | 1.421 | 38.91 | 1.396 | 43.24 | 1.079 | 63.76 | 1.042 | 65.12 | 1.350 | 47.92 | 1.221 | 53.85 | 0.779 | 84.57 |
| | MCTN (Pham et al. 2019) | 0.913 | 75.21 | 1.138 | 59.25 | 1.137 | 58.57 | 0.875 | 77.81 | 0.895 | 74.82 | 1.064 | 64.21 | 1.004 | 68.31 | 0.891 | 80.12 |
| | TransM (Wang et al. 2020) | 0.870 | 77.64 | 1.106 | 63.57 | 1.153 | 56.48 | 0.817 | 82.07 | 0.853 | 80.90 | 1.035 | 67.24 | 0.972 | 71.32 | 0.825 | 82.57 |
| | SMIL (Ma et al. 2021) | 0.894 | 78.26 | 1.067 | 67.69 | **1.112** | 59.67 | 0.866 | 79.82 | 0.859 | 79.15 | 1.019 | 71.24 | 0.970 | 72.64 | 0.818 | 82.85 |
| | GCNet (Lian et al. 2023) | 0.853 | 80.91 | 1.071 | 65.07 | 1.135 | 58.70 | 0.792 | 84.73 | **0.810** | 83.58 | 0.994 | 70.02 | 0.943 | 73.84 | 0.796 | 83.20 |
| | **UMDF (Ours)** | **0.832** | **82.92** | **1.056** | **67.80** | 1.117 | **59.92** | **0.775** | **85.63** | 0.816 | **84.09** | **0.973** | **72.98** | **0.928** | **75.56** | 0.782 | 83.36 |
| MOSEI | Self-MM (Yu et al. 2021) | 0.723 | 71.53 | 1.308 | 43.57 | 1.367 | 37.61 | 0.701 | 75.91 | 0.717 | 74.62 | 1.278 | 49.52 | 1.016 | 58.79 | 0.548 | **83.69** |
| | CubeMLP (Sun et al. 2022) | 0.768 | 67.52 | 1.353 | 39.54 | 1.428 | 32.58 | 0.725 | 71.69 | 0.750 | 70.06 | 1.301 | 48.54 | 1.054 | 54.99 | **0.540** | 83.17 |
| | MCTN (Pham et al. 2019) | 0.654 | 75.50 | 1.125 | 62.72 | 1.138 | 59.46 | 0.668 | 76.64 | 0.654 | 77.13 | 1.080 | 64.84 | 0.887 | 69.38 | 0.607 | 81.75 |
| | TransM (Wang et al. 2020) | 0.661 | 77.98 | 1.107 | 63.68 | 1.160 | 58.67 | 0.630 | 80.46 | 0.651 | 78.61 | 1.102 | 62.24 | 0.885 | 70.27 | 0.577 | 81.48 |
| | SMIL (Ma et al. 2021) | 0.627 | 76.57 | 1.089 | 65.96 | 1.122 | 60.57 | 0.606 | 77.68 | 0.617 | 76.24 | 1.063 | 66.87 | 0.854 | 70.65 | 0.568 | 80.74 |
| | GCNet (Lian et al. 2023) | 0.602 | 80.52 | 1.064 | 66.54 | **1.107** | **61.83** | 0.586 | 81.96 | 0.597 | 81.15 | **1.016** | 69.21 | 0.829 | 73.54 | 0.545 | 82.35 |
| | **UMDF (Ours)** | **0.582** | **81.57** | **1.050** | **67.42** | 1.112 | 61.57 | **0.564** | **83.25** | **0.573** | **82.14** | 1.023 | **69.48** | **0.817** | **74.24** | 0.556 | 82.16 |

Table 1: Performance comparison under different testing conditions of inter-modality missingness on MOSI and MOSEI.

| **Models** | Metrics | Testing Conditions | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $\{l\}$ | $\{a\}$ | $\{v\}$ | $\{l,a\}$ | $\{l,v\}$ | $\{a,v\}$ | Avg. | $\{l,a,v\}$ |
| Self-MM (Yu et al. 2021) | Happy | 66.9 | 52.2 | 50.1 | 69.9 | 68.3 | 56.3 | 60.6 | **90.8** |
| | Sad | 68.7 | 51.9 | 54.8 | 71.3 | 69.5 | 57.5 | 62.3 | 86.7 |
| | Angry | 65.4 | 53.0 | 51.9 | 69.5 | 67.7 | 56.6 | 60.7 | **88.4** |
| | Neutral | 55.8 | 48.2 | 50.4 | 58.1 | 56.5 | 52.8 | 53.6 | **72.7** |
| CubeMLP (Sun et al. 2022) | Happy | 68.9 | 54.3 | 51.4 | 72.1 | 69.8 | 60.6 | 62.9 | 89.0 |
| | Sad | 65.3 | 54.8 | 53.2 | 70.3 | 68.7 | 58.1 | 61.7 | **88.5** |
| | Angry | 65.8 | 53.1 | 50.4 | 69.5 | 69.0 | 54.8 | 60.4 | 87.2 |
| | Neutral | 53.5 | 50.8 | 48.7 | 57.3 | 54.5 | 51.8 | 52.8 | 71.8 |
| MCTN (Pham et al. 2019) | Happy | 76.9 | 63.4 | 60.8 | 79.6 | 77.6 | 66.9 | 70.9 | 83.1 |
| | Sad | 76.7 | 64.4 | 60.4 | 78.9 | 77.1 | 68.6 | 71.0 | 82.8 |
| | Angry | 77.1 | 61.0 | 56.7 | 81.6 | 80.4 | 58.9 | 69.3 | 84.6 |
| | Neutral | 60.1 | 51.9 | 50.4 | 64.7 | 62.4 | 54.9 | 57.4 | 67.7 |
| TransM (Wang et al. 2020) | Happy | 78.4 | 64.5 | 61.1 | 81.6 | 80.2 | 66.5 | 72.1 | 85.5 |
| | Sad | 79.5 | 63.2 | 58.9 | 82.4 | 80.5 | 64.4 | 71.5 | 84.0 |
| | Angry | 81.0 | 65.0 | 60.7 | 83.9 | 81.7 | 66.9 | 73.2 | 86.1 |
| | Neutral | 60.2 | 49.9 | 50.7 | 65.2 | 62.4 | 52.4 | 56.8 | 67.1 |
| SMIL (Ma et al. 2021) | Happy | 80.5 | 66.5 | 63.8 | 83.1 | 81.8 | 68.2 | 74.0 | 86.8 |
| | Sad | 78.9 | 65.2 | 62.2 | 82.4 | 79.6 | 68.2 | 72.8 | 85.2 |
| | Angry | 79.6 | **67.2** | 61.8 | 83.1 | 82.0 | 67.8 | 73.6 | 84.9 |
| | Neutral | 60.2 | 50.4 | 48.8 | 65.4 | 62.2 | 52.6 | 56.6 | 68.9 |
| GCNet (Lian et al. 2023) | Happy | 81.9 | 67.3 | 66.6 | 83.7 | 82.5 | **69.8** | 75.3 | 87.7 |
| | Sad | 80.5 | 69.4 | 66.1 | **83.8** | 81.9 | 70.4 | 75.4 | 86.9 |
| | Angry | 80.1 | 66.2 | 64.2 | 82.5 | 81.6 | **68.1** | 73.8 | 85.2 |
| | Neutral | 61.8 | 51.1 | 49.6 | 66.2 | 63.5 | 53.3 | 57.6 | 71.1 |
| **UMDF (Ours)** | Happy | **82.4** | **68.6** | **67.2** | **85.9** | **84.2** | 69.1 | **76.2** | 87.9 |
| | Sad | **81.2** | **70.7** | **67.1** | 83.6 | **82.2** | **71.9** | **76.1** | 86.5 |
| | Angry | **81.6** | 67.0 | **64.8** | **83.9** | **82.5** | 67.9 | **74.6** | 85.8 |
| | Neutral | **64.3** | **53.2** | **50.9** | **67.2** | **65.3** | **55.0** | **59.3** | 70.5 |

Table 2: Performance comparison under different testing conditions of inter-modality missingness on IEMOCAP.

1) UMDF outperforms previous missing-modality methods (*i.e.*, MCTN, TransM, SMIL, and GCNet) in all datasets. 2) UMDF achieves competitive results with complete-modality methods (*i.e.*, Self-MM and CubeMLP). **(iii)** In the case of intra-modality missingness (*i.e.*, $0 < p < 1$), the missing-modality methods are superior to the complete-modality ones because their training paradigms focus on capturing valuable semantics and complementing multimodal repre-

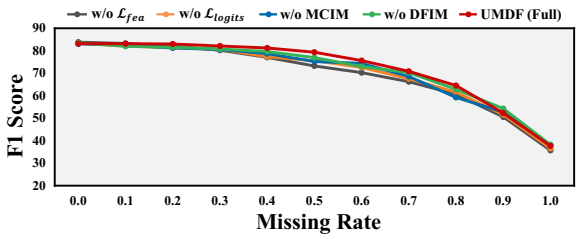


Figure 5: Ablation results of various missing rates on MOSI.

| Models | Testing Conditions | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $\{l\}$ | $\{a\}$ | $\{v\}$ | $\{l,a\}$ | $\{l,v\}$ | $\{a,v\}$ | Avg. | $\{l,a,v\}$ |
| UMDF | **82.92** | **67.80** | **59.92** | **85.63** | **84.09** | **72.98** | **75.56** | **83.36** |
| w/o $\mathcal{L}_{fea}$ | 80.84 | 65.75 | 58.08 | 83.79 | 82.02 | 70.54 | 73.50 | 82.76 |
| w/o $\mathcal{L}_{logits}$ | 81.05 | 66.04 | 58.63 | 83.98 | 82.47 | 71.08 | 73.88 | 82.53 |
| w/o MCIM | 81.76 | 66.53 | 59.02 | 84.24 | 83.01 | 71.79 | 74.39 | 83.09 |
| w/o DFIM | 82.03 | 67.12 | 59.18 | 84.85 | 83.86 | 72.16 | 74.87 | 82.73 |

Table 3: Ablation results for the testing conditions of inter-modality missingness on MOSI.

sentations from incomplete data. Benefiting from the proposed self-distillation mechanism, UMDF achieves missing features reconstruction and bidirectional knowledge transfer, resulting in the strongest robustness among all models.

**Robustness to Inter-modality Missingness.** Furthermore, we show the performance under the case of inter-modality missingness. In Tables 1&2, "$\{l\}$" means only the language modality is available, while audio and visual modalities are missing. "$\{l,a,v\}$" denotes the complete-modality testing condition where all modalities are available. "Avg." indicates the average performance in all six missing-modality testing conditions. We have the following key observations. **(i)** Firstly, the performance of UMDF in the case of inter-modality missingness is mostly worse than that of the full modality, illustrating that the sentiment semantics contained in the joint representation is enriched by combining complementary information from heterogeneous modalities. **(ii)** Among all models, UMDF works best, and its advantages derive from the multi-granularity fusion of cross-modality knowledge and the adaptive enhancement of modality-specific features. **(iii)** In the bimodal missing testing conditions, UMDF achieves comparable performance with the language modality as with the complete modalities. In the unimodal missing testing conditions, the combination of language and audio modalities performs best, even outperforming the complete modality input in individual metrics. These observations imply that the language modality contains the most informative sentiment clues and contributes indispensably to recovering missing semantics.

## Ablation Studies

To validate the necessity of the different components, we conduct comprehensive ablation studies under two missing modality cases on the MOSI dataset. The results are presented in Figure 5 and Table 3. **(i)** Firstly, $\mathcal{L}_{fea}$ and $\mathcal{L}_{logits}$ are removed from the self-distillation mechanism, and the decreased performance indicates that the bidirectional knowledge transfer plays a crucial role in the recovery of the missing element semantics. **(ii)** Moreover, we substitute self-attention layers for all crossmodal attention layers in MICM. The degraded performance indicates that it is imperative to hierarchically model inter-modal interactions and intra-modal dynamics to reproduce the missing semantics progressively. **(iii)** Eventually, DFIM is replaced by a simple concatenation operation. The worse performance revealing that adaptively enhancing and filtering the semantics in distinct modalities benefits the model performance decently.

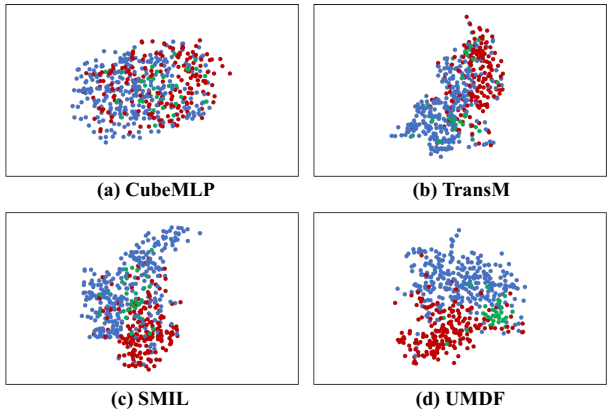


Figure 6: Visualization of the joint multimodal representations. Red, blue, and green markers indicate negative, positive, and neutral sentiment samples, respectively.

## Qualitative Analysis

To intuitively show the robustness of UMDF against modality missingness, we visualize the distribution of joint representations from UMDF and other methods on the MOSI dataset. The testing condition is set to only language modality available and $p = 0.5$. In Figure 6(a), joint representations with different emotion categories are heavily confounded, causing unsatisfactory performance. This finding implies that the complete-modality CubeMLP fails to tackle the missing modality challenge. In Figure 6(b)&(c), The missing modality methods TransM and SMIL mitigate indistinguishable sentiment semantics, which appropriately separates positive and negative categories. In contrast, UMDF has the most robust discrimination, as it effectively decouples the distinct sentiment representations in Figure 6(d).

## Conclusion

In this paper, we propose the UMDF framework to tackle the missing modality dilemma in the MSA task. UMDF yields robust joint multimodal representations through distillation-based distribution supervision and attention-based multi-grained interactions. Numerous experiments demonstrate the effectiveness of our framework under uncertain missing-modality and complete-modality testing conditions.

## Acknowledgements

This work is supported in part by the National Key R&D Program of China under Grant 2021ZD0113503.

## References

Anil, R.; Pereyra, G.; Passos, A.; Ormandi, R.; Dahl, G. E.; and Hinton, G. E. 2018. Large scale distributed neural network training through online distillation. *arXiv preprint arXiv:1804.03235*.

Baltrušaitis, T.; Robinson, P.; and Morency, L.-P. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)*, 1–10. IEEE.

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42: 335–359.

Chen, J.; Yang, D.; Jiang, Y.; Lei, Y.; and Zhang, L. 2024. MISS: A Generative Pretraining and Finetuning Approach for Med-VQA. *arXiv preprint arXiv:2401.05163*.

Cho, J. W.; Kim, D.-J.; Choi, J.; Jung, Y.; and Kweon, I. S. 2021. Dealing with missing modalities in the visual question answer-difference prediction task through knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1592–1601.

Degottex, G.; Kane, J.; Drugman, T.; Raitio, T.; and Scherer, S. 2014. COVAREP—A collaborative voice analysis repository for speech technologies. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 960–964. IEEE.

Du, C.; Du, C.; Wang, H.; Li, J.; Zheng, W.-L.; Lu, B.-L.; and He, H. 2018. Semi-supervised deep generative modelling of incomplete multi-modality emotional data. In *Proceedings of the 26th ACM International Conference on Multimedia (ACM MM)*, 108–116.

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; and Smola, A. 2012. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1): 723–773.

Han, W.; Chen, H.; and Poria, S. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *arXiv preprint arXiv:2109.00412*.

Hazarika, D.; Zimmermann, R.; and Poria, S. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*, 1122–1131.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hu, M.; Maillard, M.; Zhang, Y.; Ciceri, T.; La Barbera, G.; Bloch, I.; and Gori, P. 2020. Knowledge distillation from multi-modal to mono-modal segmentation networks. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 772–781. Springer.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lei, Y.; Yang, D.; Li, M.; Wang, S.; Chen, J.; and Zhang, L. 2023. Text-oriented Modality Reinforcement Network for Multimodal Sentiment Analysis from Unaligned Multimodal Sequences. *arXiv preprint arXiv:2307.13205*.

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

Li, Y.; Wang, Y.; and Cui, Z. 2023. Decoupled Multimodal Distilling for Emotion Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6631–6640.

Lian, Z.; Chen, L.; Sun, L.; Liu, B.; and Tao, J. 2023. GCNet: graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. 2015. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning (ICML)*, 97–105. PMLR.

Luo, W.; Xu, M.; and Lai, H. 2023. Multimodal Reconstruct and Align Net for Missing Modality Problem in Sentiment Analysis. In *Proceedings of the 31th ACM International Conference on Multimedia (ACM MM)*, 411–422. Springer.

Ma, M.; Ren, J.; Zhao, L.; Tulyakov, S.; Wu, C.; and Peng, X. 2021. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 3, 2302–2310.

Morcos, A. S.; Barrett, D. G.; Rabinowitz, N. C.; and Botvinick, M. 2018. On the importance of single directions for generalization. *arXiv preprint arXiv:1803.06959*.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Pham, H.; Liang, P. P.; Manzini, T.; Morency, L.-P.; and Póczos, B. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 01, 6892–6899.

Shraga, R.; Roitman, H.; Feigenblat, G.; and Cannim, M. 2020. Web table retrieval using multimodal deep learning. In *Proceedings of the 43rd international ACM SIGIR Conference on Research and Development in Information Retrieval*, 1399–1408.

Springstein, M.; Müller-Budack, E.; and Ewerth, R. 2021. QuTI! quantifying text-image consistency in multimodal documents. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2575–2579.

Sun, H.; Wang, H.; Liu, J.; Chen, Y.-W.; and Lin, L. 2022. CubeMLP: An MLP-based model for multimodal sentiment analysis and depression estimation. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, 3722–3729.

Tran, L.; Liu, X.; Zhou, J.; and Jin, R. 2017. Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1405–1414.

Tsai, Y.-H. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.-P.; and Salakhutdinov, R. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the Conference Association for Computational Linguistics Meeting (ACL)*, volume 2019, 6558.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30.

Wang, Y.; Shen, Y.; Liu, Z.; Liang, P. P.; Zadeh, A.; and Morency, L.-P. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 01, 7216–7223.

Wang, Y.; Zhang, Y.; Liu, Y.; Lin, Z.; Tian, J.; Zhong, C.; Shi, Z.; Fan, J.; and He, Z. 2021. ACN: adversarial co-training network for brain tumor segmentation with missing modalities. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 410–420. Springer.

Wang, Z.; Wan, Z.; Wan, X.; and Wan, X. 2020. Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis. In *Proceedings of The Web Conference 2020*, 2514–2520.

Yang, D.; Chen, Z.; Wang, Y.; Wang, S.; Li, M.; Liu, S.; Zhao, X.; Huang, S.; Dong, Z.; Zhai, P.; and Zhang, L. 2023a. Context De-Confounded Emotion Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19005–19015.

Yang, D.; Huang, S.; Kuang, H.; Du, Y.; and Zhang, L. 2022a. Disentangled Representation Learning for Multimodal Emotion Recognition. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, 1642–1651.

Yang, D.; Huang, S.; Liu, Y.; and Zhang, L. 2022b. Contextual and Cross-Modal Interaction for Multi-Modal Speech Emotion Recognition. *IEEE Signal Processing Letters*, 29: 2093–2097.

Yang, D.; Huang, S.; Wang, S.; Liu, Y.; Zhai, P.; Su, L.; Li, M.; and Zhang, L. 2022c. Emotion Recognition for Multiple Context Awareness. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 13697, 144–162.

Yang, D.; Kuang, H.; Huang, S.; and Zhang, L. 2022d. Learning Modality-Specific and -Agnostic Representations for Asynchronous Multimodal Language Sequences. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, 1708–1717.

Yang, D.; Liu, Y.; Huang, C.; Li, M.; Zhao, X.; Wang, Y.; Yang, K.; Wang, Y.; Zhai, P.; and Zhang, L. 2023b. Target and source modality co-reinforcement for emotion understanding from asynchronous multimodal sequences. *Knowledge-Based Systems*, 265: 110370.

Yang, D.; Yang, K.; Wang, Y.; Liu, J.; Xu, Z.; Yin, R.; Zhai, P.; and Zhang, L. 2023c. How2comm: Communication-efficient and collaboration-pragmatic multi-agent perception. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*.

Yang, K.; Yang, D.; Zhang, J.; Wang, H.; Sun, P.; and Song, L. 2023d. What2comm: Towards Communication-Efficient Collaborative Perception via Feature Decoupling. In *Proceedings of the 31th ACM International Conference on Multimedia (ACM MM)*, 7686–7695.

Yu, W.; Xu, H.; Yuan, Z.; and Wu, J. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, 12, 10790–10797.

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

Zadeh, A.; Liang, P. P.; Mazumder, N.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018a. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, 1, 5634–5641.

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. 2018b. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the Conference Association for Computational Linguistics Meeting (ACL)*, 2236–2246.

Zeng, J.; Liu, T.; and Zhou, J. 2022. Tag-assisted Multimodal Sentiment Analysis under Uncertain Missing Modalities. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1545–1554.

Zeng, J.; Zhou, J.; and Liu, T. 2022. Mitigating Inconsistencies in Multimodal Sentiment Analysis under Uncertain Missing Modalities. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2924–2934.

Zhao, J.; Li, R.; and Jin, Q. 2021. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the Conference Association for Computational Linguistics Meeting (ACL)*, 2608–2618.