Decoupling and Reconstructing: a Multimodal Sentiment Analysis Framework Towards Robustness

Anonymous Author(s)

Abstract

Multimodal sentiment analysis (MSA) has shown promising results but often poses significant challenges in real-world applications due to its dependence on the complete and aligned multimodal sequences. While existing approaches attempt to address missing modalities through feature reconstruction, they often neglect the complex interplay between homogeneous and heterogeneous relationships in multimodal features. To address this problem, we propose ***Decoupled-Adaptive Reconstruction (DAR)***, a novel framework that explicitly addresses these limitations through two key components: (1) a mutual information-based decoupling module that decomposes features into common and independent representations, and (2) a reconstruction module that independently processes these decoupled features before fusion for downstream tasks. Extensive experiments on two benchmark datasets demonstrate that DAR significantly outperforms existing methods in both modality reconstruction and sentiment analysis tasks, particularly in scenarios with missing or unaligned modalities. Our results show improvements of 2.21% in bi-classification accuracy and 3.9% in regression error compared to state-of-the-art baselines on the MOSEI dataset.

1 Introduction

As an important research direction in artificial intelligence, multimodal emotion recognition aims to achieve more accurate and comprehensive emotional understanding through the integration and analysis of information from different modalities (such as speech, text, vision, etc.)[Liang *et al.*, 2021; Lv *et al.*, 2021a]. With the rapid development of deep learning technologies and the increasing abundance of multimodal data, significant progress has been made in this field.

Compared to laboratory environments where high-quality data samples can be artificially selected for training, data collected in real scenarios may face varying degrees of missing issues, leading to otherwise well-performing multimodal sentiment classification models to face severe performance loss when dealing with real-world incomplete data.

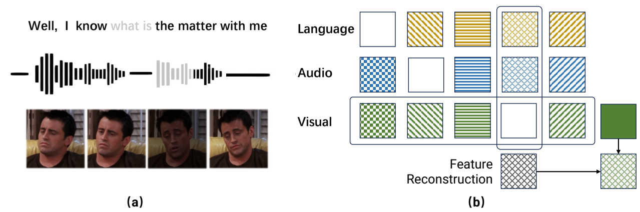

Figure 1: (a) shows an example of incomplete data entry, with the gray overlay indicating invisibility. (b) shows an illustration of feature reconstruction, where blank parts are missing features and colors represent modal-independent features, textures represent modal-common features.

Recently, research trends have shifted from laboratory conditions to modeling data from natural scenarios. This shift creates a wider application space for MSA in the real world, despite concerns due to issues such as sensor failure and automatic speech recognition (ASR), which lead to inconsistencies such as incomplete data in real-world deployments. Many influential solutions have been proposed to address the major problem of incomplete data in multimodal sentiment analysis. For example, [Yuan *et al.*, 2021] introduced a transformer-based feature reconstruction mechanism, TFR-Net, which aims to improve the robustness of the model in dealing with random deletions in unaligned multimodal sequences by reconstructing the missing data. Zhang introduced a model (LNLN)[Zhang *et al.*, 2024], the Language Dominated Noise Resistant Learning Network, to improve the robustness of MSA to incomplete data. It aims to enhance the completeness of linguistic mood features, which are considered dominant moods due to their richer emotional cues and supported by other auxiliary moods.

Previous methods have demonstrated that the use of missing data during training is helpful in improving the robustness of the model in an incomplete input scenario and have also verified that reconstructing complete data using missing data allows the model to learn more stable features. However, these methods have the following problems: the process of reconstructing complete inputs does not take into account the redundancy and complementarity that exists between different modal data, resulting in the model failing to achieve the desired reconstruction effect; at the same time, the inclusion of reconstruction loss may cause the model to pay too much attention to the consistency between the complete data and the missing data after feature extraction, resulting in the degradation of the encoder effect and the failure to effectively extract key features.

To solve the above problems, we propose a feature decoupling-reconstructing approach for multimodal feature fusion. As shown in Figure 1, we first decompose modal features into modal-independent and modal-common features by methods of mutual information-based approach. Then we reconstruct features corresponding to two complete inputs according to the respective properties of the two types of features. We also use a specialized neural network for the output from complete data to guide the supervised feature reconstruction of the model features for the downstream task. Finally, the loss is added to the overall loss to avoid the degradation problem caused by the feature encoder's tendency to favor the reconstruction effect. The contributions of this work can be summarized as:

- We propose a new approach that is suitable for feature reconstruction to decouple sequence features based on mutual information.
- We propose a missing feature reconstruction method based on decoupled features, which intuitively reflects the redundancy and complementary relationship between different modal data.
- We validate our approach on two widely used multimodal sentiment analysis datasets and compare it with other robust and non-robust fusion methods. The results demonstrate that our approach outperforms other existing models on several metrics and achieves the best overall performance.

2 Related Work

2.1 Robust Representation Learning in MSA

Multimodal Sentiment Analysis (MSA) methods can be categorized into Context-based MSA and Noise-aware MSA, depending on the modeling approach[Zhang *et al.*, 2024]. Most of previous works ([Zadeh *et al.*, 2017]; [Tsai *et al.*, 2019]; [Mai *et al.*, 2020]; [Hazarika *et al.*, 2020]; [Liang *et al.*, 2020]; [Rahman *et al.*, 2020]; [Yu *et al.*, 2021]; [Han *et al.*, 2021]; [Lv *et al.*, 2021b]; [Yang *et al.*, 2022]; [Guo *et al.*, 2022]; [Zhang *et al.*, 2023]) can be classified to Context-based MSA. This line of work primarily focuses on learning unified multimodal representations by analyzing contextual relationships within or between modalities. For example, [Zadeh *et al.*, 2017] explore computing the relationships between different modalities using the Cartesian product. [Tsai *et al.*, 2019] utilize pairs of Transformers to model long dependencies between different modalities. [Yu *et al.*, 2021] propose generating pseudo-labels for each modality to further mine the information of consistency and discrepancy between different modalities. Despite these advances, context-based methods are usually suboptimal under varying levels of noise effects (e.g. random data missing). Several recent works ([Mittal *et al.*, 2020];[Yuan *et al.*, 2021];[Yuan *et al.*, 2024];[Li *et al.*, 2025]) have been proposed to tackle this issue.

In concrete terms, [Hazarika *et al.*, 2020] and [Yang *et al.*, 2022] apply feature disentanglement to each modality, modeling multimodal representations from multiple feature subspaces and perspectives. [Yu *et al.*, 2021] and [Liang *et al.*, 2021] explore self-supervised learning and semi-supervised learning to enhance multimodal representations, respectively. [Tsai *et al.*, 2019] and [Rahman *et al.*, 2020] introduce Transformer to learn the long dependencies of modalities. [Zhang *et al.*, 2023] devise a language-guided learning mechanism that uses modalities with more intensive sentiment cues to guide the learning of other modalities. Noise-aware MSA focuses more on perceiving and eliminating the noise present in the data. For example, [Mittal *et al.*, 2020] design a modality check module based on metric learning and Canonical Correlation Analysis (CCA) to identify the modality with greater noise. [Yuan *et al.*, 2021] design a feature reconstruction network to predict the location of missing information in sequences and reconstruct it. [Yuan *et al.*, 2024] introduce adversarial learning to perceive and generate cleaner representations. [Zhang *et al.*, 2024] proposed LNLN, explored the capability of language-guided mechanisms in resisting noise and provide new. perspectives for the study of MSA in noisy scenarios.

2.2 Multimodal feature decoupling

One of the more important features of multimodal tasks, compared to unimodal tasks, is the redundancy and complementarity of the modal information prior([Zhao *et al.*, 2024]). A lot of work has been done to explore the decoupling of modal features into irrelevant classifications and apply them to downstream tasks, starting from the commonalities and differences of information between different modalities. Currently, multimodal feature decoupling can be categorized into two kinds: spatial-based and mutual information-based, among which the spatial-based work is [Hazarika *et al.*, 2020] and [Li *et al.*, 2023], The degree of similarity and dissimilarity of features is measured using the vanilla cosine distances between feature vectors, respectively. And the mutual information-based approach is [Yang *et al.*, 2023] and [Xia *et al.*, 2024]. The former defines similar and dissimilar features by constructing positive and negative examples, and the latter optimizes the loss of mutual information by constructing time-series versions of the upper and lower bounds on the use of mutual information approximations.

Inspired by works on mutual information-based feature decomposition([Yang *et al.*, 2023];[Xia *et al.*, 2024]), the sequence feature decoupling module proposed in this paper employs a similarity measure based on both mutual information and spatial properties, which assumes that similar features have high mutual information between them, while mutual information between dissimilar features should be minimized.

3 The DAR Model

3.1 Task Setup

In this paper, we consider three modalities, i.e., language (l), visual(v), acoustic (a). These modalities are represented as $\mathbf{U}_l \in \mathbb{R}^{T_l \times d_l}$, $\mathbf{U}_v \in \mathbb{R}^{T_v \times d_v}$, and $\mathbf{U}_a \in \mathbb{R}^{T_a \times d_a}$ respectively. Here T_m denotes the length of the utterance, such as

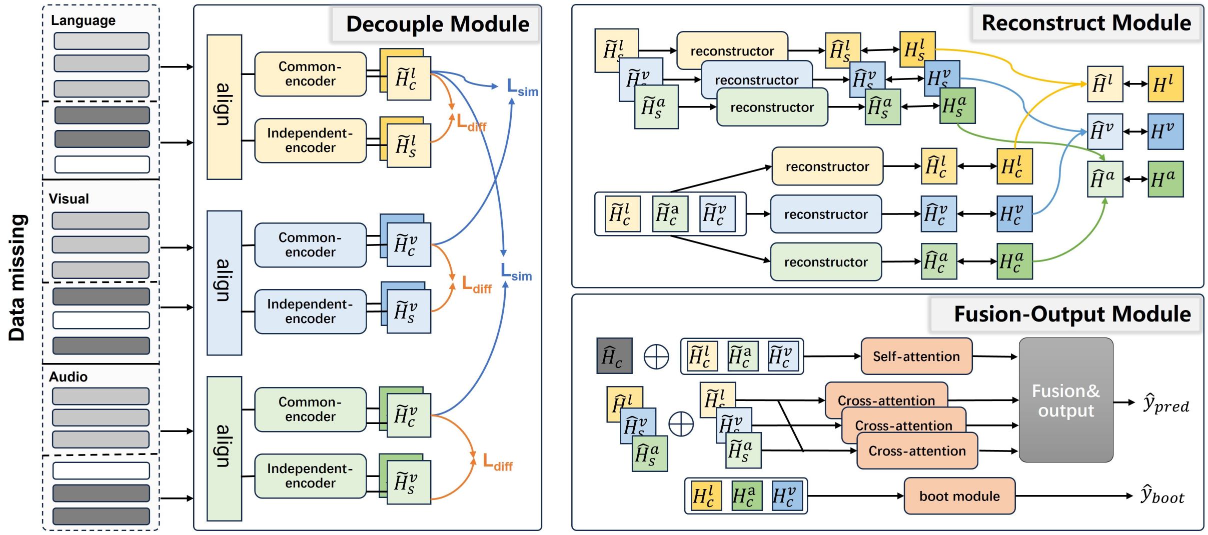

Figure 2: The overall architecture of our proposed model. Light gray blocks on the left side indicate complete inputs, dark gray blocks indicate missing inputs, and blanks indicate missing parts. The model consists of three main components: (a) decouple module, (b) reconstruct module, and (c) Fusion-Output module, where the marker s denotes modal independent features, c denotes modal common features and two-way arrows represent comparative losses.

number of tokens (T_l), for modality m and d_m denotes the respective feature dimensions.

Given these sequences $\mathbf{U}_{m\in\{l,v,a\}}$, the primary task is to predict the affective orientation of utterance U from either a predefined set of C categories $y \in \mathbb{R}^C$ or as a continuous intensity variable $y \in \mathbb{R}$.

3.2 Overview

The general structure of the model is shown in Figure 2. It first obtains incomplete multimodal data through the datamissing operation. Model DAR first uses an alignment layer to adjust the input features of all modalities to the same dimension to ensure data consistency. Then, for each modal input, we use independent modal-common feature encoder and modal-independent feature encoder to obtain modal-common representation and modal-independent representation of the features. Next, the modal reconstruction module corrects the decomposed two feature reconstructions to restore the feature representation corresponding to the full input. Finally, the feature fusion module utilizes the self-attention mechanism and the cross-attention mechanism to process the two kinds of features, fuse them, and output the classification results through the output layer.

3.3 Input Construction and Multimodal Input

Following the previous method ([Zhang *et al.*, 2024]), for each modality, we randomly erase changing proportions of information (from 0% to 90%). These pre-processed inputs are represented as sequences, denoted by $\mathbf{U}_m \in \mathbb{R}^{T_m \times d_m}$, $m \in \{l, v, a\}$ representing language, visual and acoustic features respectively where T_m denotes the length of the sequence for modality m (such as number of tokens for $m = l$), and d_m denotes the respective feature dimensions . With obtained $\mathbf{U}_m$, we apply random data missing to $\mathbf{U}_m$, thus forming the noise-corrupted multimodal input $\widetilde{\mathbf{U}}_m$.

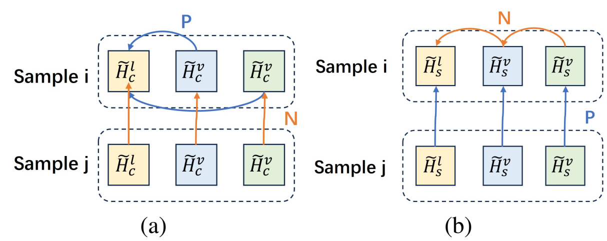

Figure 3: Method of dividing positive and negative examples. (a) represents the modal-common features pairing; (b) represents the modal-independent features pairing.

3.4 Decouple Module

It is essential to standardize the feature representations across modalities for ease of further processing. To achieve this, we apply 1D convolutions followed by a simple nonlinear layers to process the input features. Given features corresponding to complete input data and random missing data be represented as $\mathbf{U}_m \in \mathbb{R}^{T_m \times d_m}$ and $\widetilde{\mathbf{U}}_m \in \mathbb{R}^{T_m \times d_m}$, $m \in \{l, v, a\}$. After the alignment operation, the output feature $\mathbf{U}_m^1 \in \mathbb{R}^{t \times d}$ and $\widetilde{\mathbf{U}}_m^1 \in \mathbb{R}^{t \times d}$ have unified length of utterance, t and feature dimension d across all modalities, making it suitable for subsequent model processing.

Given the incomplete sequence $\widetilde{\mathbf{U}}_m^1 \in \mathbb{R}^{t \times d}$ for modality m, we employ common feature extractors and independent feature extractors to extract the modal-common features $\widetilde{\mathbf{H}}_m^{\text{com}}$ and modal-independent features $\widetilde{\mathbf{H}}_m^{\text{spec}}$ using the encoding functions.

$$\widetilde{\mathbf{H}}_m^{\text{com}} = E_c(\widetilde{\mathbf{U}}_m^1; \theta_m^c), \quad \widetilde{\mathbf{H}}_m^{\text{spec}} = E_s(\widetilde{\mathbf{U}}_m^1; \theta_m^s) \quad (1)$$

Similarly, for the complete input corresponding to feature $\mathbf{U}_m^1$ we also use the same encoder to obtain the corresponding modal-common input and modal-independent inputs $\mathbf{H}_m^{\text{com}}$ and $\mathbf{H}_m^{\text{spec}}$. We reserve two types of features for the generation of restoration features under supervision.

Based on the characteristics of the modal-common and modal-independent features, we aim to ensure that the common features from the same sample across different modalities exhibit high consistency, while the independent features within the same modality show high consistency as well. Simultaneously, we seek to reduce the information redundancy between the two types of features. To achieve this, we define a decoupling loss function $\mathcal{L}_{\text{decouple}}$ as:

$$\mathcal{L}_{\text{decouple}} = \lambda(\mathcal{L}_{\text{sim}} + \mathcal{L}_{\text{diff}}) + \mathcal{L}_{\text{re}} \tag{2}$$

Where λ is a hyperparameter, $\mathcal{L}_{\text{re}}$ is the restoration loss that reduces the decomposed feature to the original feature and I for mutual information. The mutual information between the two distributions is represented as follows:

$$I(\mathbf{z}_1;\mathbf{z}_2) = \int\int p(\mathbf{z}_1,\mathbf{z}_2)\log\frac{p(\mathbf{z}_1,\mathbf{z}_2)}{p(\mathbf{z}_1)p(\mathbf{z}_2)}\,d\mathbf{z}_1 d\mathbf{z}_2 \tag{3}$$

where:$p(\mathbf{z}_1,\mathbf{z}_2)$ is the joint probability distribution of $\mathbf{z}_1$ and $\mathbf{z}_2$,$p(\mathbf{z}_1)$ and $p(\mathbf{z}_2)$ are the marginal distributions of $\mathbf{z}_1$ and $\mathbf{z}_2$, respectively.

Specifically, for sets of data in batches B we have:

$$\begin{aligned}\mathcal{L}_{\text{sim}} = &-I(\widetilde{\mathbf{H}}_a^{\text{com}};\widetilde{\mathbf{H}}_v^{\text{com}};\widetilde{\mathbf{H}}_l^{\text{com}}) \\ &-\sum_m^M I(\widetilde{\mathbf{H}}_{m,i}^{\text{spec}};\widetilde{\mathbf{H}}_{m,j}^{\text{spec}})\end{aligned} \tag{4}$$

where i, j represent two different batches of data.

$$\mathcal{L}_{\text{diff}} = \sum_m^M I(\widetilde{\mathbf{H}}_m^{\text{spec}};\widetilde{\mathbf{H}}_m^{\text{com}}) \tag{5}$$

where $\widetilde{\mathbf{H}}_m^{\text{com}}$ and $\widetilde{\mathbf{H}}_m^{\text{spec}}$ represent the modal-common features and modal-independent features, respectively, $m \in M$ and $M = \{l, v, a\}$. The objective is to maximize the mutual information between the common features of different modalities for the same sample and the independent features of different batches within the same modality, while minimizing the mutual information between the common and independent features of the same sample.

Since it is difficult to compute the mutual information directly, we use the mutual information approximate upper and lower bounds to optimize the loss function as above. For the similarity loss, we use the noise comparison lower bounds of the mutual information for optimization; for the dissimilarity loss, we use the CLUB upper bounds of the mutual information for optimization, and we achieve the minimization of decoupling loss by optimizing the upper and lower bounds of the mutual information.

InfoNCE-based Mutual Information Maximization: InfoNCE([Oord *et al.*, 2018]) is a commonly used lower bound for mutual information loss, contrastive methods enhance this by utilizing sample pairs from positive set $\mathcal{P}$ and negative set $\mathcal{N}$. The goal is to pull positive pairs closer in the representation space while pushing negative pairs apart. The commonly used InfoNCE loss is defined as:

$$\begin{aligned}\mathcal{L}_{\text{sim}} = -\frac{1}{|\mathcal{P}|}\sum_{(\mathbf{z}_1,\mathbf{z}_2)\in\mathcal{P}} &\log[\exp(\text{sim}(\mathbf{z}_1,\mathbf{z}2)/\tau)/ \\ \sum_{(\mathbf{z}_1,\mathbf{z}_i)\in\mathcal{N}} &\exp(\text{sim}(\mathbf{z}_1,\mathbf{z}_i)/\tau)]\end{aligned} \tag{6}$$

where: $\text{sim}(\cdot,\cdot)$ is a similarity function, in this paper, we use the cosine similarity, and τ is a temperature parameter. $|\mathcal{P}|$ denotes the cardinality of the positive pair set. We maximize the mutual information between positive examples by constructing positive and negative examples, chosen as shown in Figure 3. According to 3a, 3b in Figure 3, we compute the $\mathcal{L}_{\text{sim}}^{\text{com}}$ and $\mathcal{L}_{\text{sim}}^{\text{spec}}$ corresponding to the common and independent features respectively, and add the two together to obtain the final $\mathcal{L}_{\text{sim}}$.

$$\mathcal{L}_{\text{sim}} = \mathcal{L}_{\text{sim}}^{\text{com}} + \mathcal{L}_{\text{sim}}^{\text{spec}} \tag{7}$$

We average the original time series features in the time dimension as the sample features, obtain the corresponding feature $\mathbf{z}$, calculate the InfoNCE as the loss of the lower bound of the mutual information.

CLUB-based MI Minimization: CLUB can effectively optimize the MI upper bound, demonstrating superior advantages in information disentanglement [Cheng *et al.*, 2020]. Given two variables $\mathbf{x}$ and $\mathbf{y}$, the objective function of CLUB is defined as:

$$\begin{aligned}I_{\text{vCLUB}}(\mathbf{x};\mathbf{y}) := &\mathbb{E}_{p(\mathbf{x},\mathbf{y})}[\log q_\theta(\mathbf{y}|\mathbf{x})] \\ &-\mathbb{E}_{p(\mathbf{x})}\mathbb{E}_{p(\mathbf{y})}[\log q_\theta(\mathbf{y}|\mathbf{x})],\end{aligned} \tag{8}$$

where q_θ is the variational approximation of ground-truth posterior of $\mathbf{y}$ given $\mathbf{x}$ and can be parameterized by a network θ. We use CLUB to optimize the MI upper bound between the common features $\widetilde{\mathbf{H}}_m^{\text{com}}$ and modal-specific features $\widetilde{\mathbf{H}}_m^{\text{spec}}$. To better measure the mutual information between the two temporal features, we use a combination of a bidirectional lstm([Huang *et al.*, 2015]) and a nonlinear fully connected layer as a variational approximation network q_θ, we modify I_{vCLUB} into following:

$$\begin{aligned}\mathcal{L}_{\text{diff}} = &\frac{1}{N}\sum_{i=1}^{N}[\log q_\theta(\widetilde{\mathbf{H}}_m^{\text{com}}|\widetilde{\mathbf{H}}_m^{\text{spec}}) \\ &-\frac{1}{N}\sum_{j=1}^{N}\log q_\theta(\widetilde{\mathbf{H}}_m^{\text{com}}|\widetilde{\mathbf{H}}_m^{\text{spec}})],\end{aligned} \tag{9}$$

The approximation network and the main networks are optimized alternatively during training process.

Restoration loss: To distinguish the differences between $\widetilde{\mathbf{H}}_m^{\text{com}}$ and $\widetilde{\mathbf{H}}_m^{\text{spec}}$ and mitigate the feature ambiguity, we synthesize the vanilla coupled features $\widetilde{\mathbf{U}}_m^1$ in a self-regression manner. Mathematically speaking, for each modality m, we concatenate the features from the other two modalities with $\widetilde{\mathbf{H}}_m^{\text{spec}}$ and exploit a private decoder $\mathcal{D}_m$ to produce the coupled feature. Specifically: For modality l:

$$\mathcal{L}_{\text{re}}^l = \|\widetilde{\mathbf{U}}_l^1 - \mathcal{D}_l(\text{Concat}(\widetilde{\mathbf{H}}_v^{\text{com}},\widetilde{\mathbf{H}}_a^{\text{com}},\widetilde{\mathbf{H}}_l^{\text{spec}}))\|_F^2 \tag{10}$$

For the other two modalities, we also use the same way to get the losses $\mathcal{L}_{\text{re}}^v$ and $\mathcal{L}_{\text{re}}^a$. Adding up these losses, we get the overall restoration loss $\mathcal{L}_{\text{re}}$:

$$\mathcal{L}_{\text{re}} = \mathcal{L}_{\text{re}}^l + \mathcal{L}_{\text{re}}^v + \mathcal{L}_{\text{re}}^a \tag{11}$$

3.5 Reconstruct Module

We hypothesize that the independent features of a complete modality can be predicted through the corresponding independent features of the missing modality feature, while the common features of a complete modality can be predicted by the common features of all the input missing modalities feature.

To implement this, we propose two distinct feature reconstruction modules for each modality: the Independent Feature correction module and the Common Feature reconstruction module. The Independent Feature reconstruction module takes as input the decoupled independent features and outputs the corrected independent features $\hat{\mathbf{H}}_m^{\text{spec}}$. In contrast, the Common Feature Reconstruction module uses the combined common features from all modalities as input and generates the reconstructed features $\hat{\mathbf{H}}_m^{\text{com}}$ as output. Finally, after obtaining the two features, we use a specially set up private decoder $\mathcal{D}_m$ to reconstruct the coupled complete input $\mathbf{U_m^1}$.

$$\hat{\mathbf{H}}_m^{\text{com}} = E_{com}^m(\text{Concat}(\widetilde{\mathbf{H}}_l^{\text{com}}, \widetilde{\mathbf{H}}_v^{\text{com}}, \widetilde{\mathbf{H}}_a^{\text{com}}), \theta_{com}^m), \tag{12}$$

$$\hat{\mathbf{H}}_m^{\text{spec}} = E_{spec}^m(\widetilde{\mathbf{H}}_m^{\text{spec}}, \theta_{spec}^m), \tag{13}$$

$$\hat{\mathbf{U}}_m^1 = \mathcal{D}_m(\text{Concat}(\widetilde{\mathbf{H}}_m^{\text{com}}, \widetilde{\mathbf{H}}_m^{\text{spec}}) \tag{14}$$

where θ_{com} denotes the parameters of the common feature reconstruction module E_{com} and θ_{spec} denotes the parameters of the independent feature reconstruction module E_{spec}.

Finally, we combine reconstructed features with original input features to obtain features for downstream tasks.

$$\mathbf{g} = \sigma(\mathbf{W}_g[\hat{\mathbf{H}}, \widetilde{\mathbf{H}}] + \mathbf{b}_g) \tag{15}$$

$$\mathbf{H}_{\text{fused}} = \mathbf{g} \odot \hat{\mathbf{H}} + (1 - \mathbf{g}) \odot \widetilde{\mathbf{H}} \tag{16}$$

To ensure that the reconstructed features are consistent with the common and independent features obtained from the complete input through the encoder, hereafter referred to as the complete common and complete independent features, we construct the alignment loss minimizing the loss between the corrected features and the complete features as following:

$$\mathcal{L}_{\text{recon}} = \|\hat{\mathbf{H}} - \mathbf{H}\|_F^2 + \|\hat{\mathbf{U}}^1 - \mathbf{U}^1\|_F^2 \tag{17}$$

3.6 Fusion-Output Module

For the modal-common features, which exhibit relatively similar distributions, we apply a multi-layer self-attention model for further refinement. In contrast, for the modal-independent features, where there are significant distributional differences between features, we employ a cross-attention mechanism.

Modal-common Features Fusion Module. Given the modified modal-common feature $\mathbf{H}_{\text{fused}}^{\text{com}}$, we perform feature fusion in the temporal dimension using a multilayer self-attention module for each modal counterpart, while using the features of the last frame of the output of the last layer as the overall feature output $\mathbf{h}_{\text{fused}}$.

$$\mathbf{h}^{\text{com}} = \text{SelfAttention}(\mathbf{H}_{\text{fused}}^{\text{com}})[-1] \tag{18}$$

Modal-independent Features Fusion Module. For modal-independent features, we use a cross-attention mechanism to fuse different modal information. The core of the multimodal transformer is the crossmodal attention unit (CA), which receives features from a pair of modalities and fuses cross-modal information. Take the language modality $\mathbf{H}_{\text{fused-L}}^{\text{spec}}$ as the source and the visual modality $\mathbf{H}_{\text{fused-V}}^{\text{spec}}$ as the target, the cross-modal attention can be defined as: $\mathbf{Q}_V = \mathbf{H}_{\text{fused-V}}^{\text{spec}}\mathbf{P}_q$, $\mathbf{K}_L = \mathbf{H}_{\text{fused-L}}^{\text{spec}}\mathbf{P}_k$, and $\mathbf{V}_L = \mathbf{H}_{\text{fused-L}}^{\text{spec}}\mathbf{P}_v$, where $\mathbf{P}_q$, $\mathbf{P}_k$, $\mathbf{P}_v$ are the learnable parameters, formulated as:

$$\mathbf{h}_{L\to V}^{\text{spec}} = \text{softmax}\left(\frac{\mathbf{Q}_V\mathbf{K}_L^\top}{\sqrt{d}}\right)\mathbf{V}_L[-1], \tag{19}$$

where $\mathbf{h}_{L\to V}^{\text{spec}}$ is the enhanced features from Language to Visual, d means the dimension of $\mathbf{Q}_V$ and $\mathbf{K}_L$. For the three modalities in MER, feature of each modality $\mathbf{h}_m^{\text{spec}}$ will be reinforced by the two others and the resulting features will be concatenated. Take visual modality as an example the formula is expressed as follows:

$$\mathbf{h}_V^{\text{spec}} = \text{Concat}(\mathbf{h}_{L\to V}^{\text{spec}}, \mathbf{h}_{A\to V}^{\text{spec}}) \tag{20}$$

Prediction/Inference. Finally, we splice the obtained fused features and input the nonlinear fully connected layer to generate predictions $\hat{y}$, we also use the bootstrap module to predict the results $\hat{y}_{\text{boot}}$ using common features generated from the complete information, ensuring that the encoder learns features that facilitate classification.

$$\hat{y} = \text{MLP}(\text{Concat}(\mathbf{h}^{\text{com}}, \mathbf{h}^{\text{spec}})) \tag{21}$$

$$\hat{y}_{\text{boot}} = \text{MLP}(\mathbf{H}) \tag{22}$$

The task loss $\mathcal{L}_{\text{task}}$ and overall model loss $\mathcal{L}_{\text{total}}$ are formulated as follows:

$$\mathcal{L}_{\text{task}} = \text{Loss}(y, \hat{y}) + \text{Loss}(y, \hat{y}_{\text{boot}}) \tag{23}$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \alpha\mathcal{L}_{\text{decouple}} + \beta\mathcal{L}_{\text{recon}} \tag{24}$$

where α and β are hyperparameters.

4 Experiments and Analysis

In this section, we provide a comprehensive and fair comparison between the proposed DAR and previous representative MSA methods on MOSI ([Zadeh *et al.*, 2016]) and MOSEI ([Bagher Zadeh *et al.*, 2018]) datasets.

4.1 Datasets

MOSI The dataset includes 2,199 multimodal samples, integrating visual, audio, and language modalities. It is divided into a training set of 1,284 samples, a validation set of 229 samples, and a test set of 686 samples. Each sample is given a sentiment score, varying from -3, indicating strongly negative sentiment, to 3, signifying strongly positive sentiment.

MOSEI The dataset consists of 22,856 video clips sourced from YouTube. The sample is divided into 16,326 clips for training, 1,871 for validation, and 4,659 for testing. Each clip is labeled with a score, ranging from -3, denoting the strongly negative, to 3, denoting the strongly positive.

Method	MOSI						MOSEI					
	Acc-7	Acc-5	Acc-2	F1	MAE↓	Corr	Acc-7	Acc-5	Acc-2	F1	MAE↓	Corr
MISA	28.90	31.67	69.15 / 70.74	68.50 / 70.23	1.092	0.508	38.92	39.28	76.21 / 72.12	70.76 / 65.50	0.800	0.490
Self-MM	30.78	34.03	68.75 / 70.89	65.47 / 67.90	1.070	0.518	46.40	46.78	71.18 / 72.75	70.45 / 70.99	0.695	0.498
MMIM	31.51	34.92	69.22 / 71.08	67.34 / 69.42	1.077	0.511	44.04	44.42	75.99 / 71.47	70.63 / 64.97	0.739	0.459
CENET	29.78	33.23	66.41 / 69.47	62.65 / 65.38	1.088	0.496	**47.18**	47.93	75.96 / 74.10	73.28 / 70.51	0.685	0.525
TETFN	29.89	33.20	68.66 / 70.89	65.11 / 67.64	1.087	0.512	46.31	47.03	71.63 / 71.84	68.91 / 68.14	0.714	0.508
TFR-Net	29.54	34.67	68.15 / 66.35	61.73 / 60.06	1.200	0.459	46.83	34.67	73.62 / 77.23	68.80 / 71.99	0.697	0.489
ALMT	30.35	32.92	68.27 / 70.55	64.47 / 67.07	1.083	0.506	42.01	42.58	76.75 / 72.96	72.00 / 67.16	0.754	0.511
LNIN	32.80	36.12	71.11 / 72.22	71.33 / 72.34	**1.066**	0.505	45.42	46.17	75.27 / 76.98	74.97 / 77.39	0.692	0.530
Ours	**34.47**	**38.65**	**71.60 / 73.18**	**71.51 / 73.15**	1.069	**0.520**	47.01	**48.02**	**77.48 / 78.14**	**77.44 / 77.51**	**0.665**	**0.583**

Table 1: Performance comparison on MOSI and MOSEI datasets.

4.2 Evaluation Settings and Criteria

For each sample in the dataset, we incorporate data from three modalities: language, audio, and visual data. Consistent with previous works ([Zhang *et al.*, 2023]), each modality is processed using widely-used tools: language data is encoded using BERT([Devlin, 2018]), audio features are extracted through Librosa ([McFee *et al.*, 2015]), and visual features are obtained using OpenFace ([Baltrusaitis *et al.*, 2018]). Specifically, for visual and audio modalities, we fill the erased information with zeros. For language modality, we fill the erased information with [UNK] which indicates the unknown word in BERT ([Devlin, 2018]).

Following the previous works ([Zhang *et al.*, 2024]), we report our results in classification and regression with the average of 3 runs of different seeds and 10 missing rates from 0.0 to 0.9 at 0.1 intervals. For classification, we report the multiclass accuracy and weighted F1 score. We calculate the accuracy of 2-class prediction, 5-class prediction (Acc-5) and 7-class prediction (Acc-7) for MOSI and MOSEI. Besides, Acc-2 and F1-score of MOSI and MOSEI have two forms: negative/non-negative (non-exclude zero) ([Zadeh *et al.*, 2017]) and negative/positive (exclude zero) ([Tsai *et al.*, 2019]1). For regression, we report Mean Absolute Error (MAE) and Pearson correlation (Corr). Except for MAE, higher values indicate better performance for all metrics.

In training process, for hyperparameters, we choose that $\lambda = 0.7$, $\alpha = 0.1$, $\beta = 0.1$. On the mosi dataset, we choose the missing rate $k = 0.3$, and on the mosei dataset, we choose $k = 0.4$.

Compared with the baseline LNLN([Zhang *et al.*, 2024]) which uses the best model under different metrics for testing, we use the same model with the smallest overall loss as the optimal model for testing, and at the same time, in order to ensure the stability of the results, we randomly test three times and take the average value as the final result following the baseline settings.

Inaddition, the result of MISA, Self-MM, MMIM, CENET, TETFN, ALMT is reproduced by the authors from open source code in the MMSA([Mao *et al.*, 2022]),which is a unified framework for MSA, using default hyperparameters, LNLN([Zhang *et al.*, 2024]) model is implemented using the author's open source code and for TFR-Net, We use the results reported in the LNLN article, and since that article uses the best modeling results under the corresponding metrics, we consider this comparison to be fair.

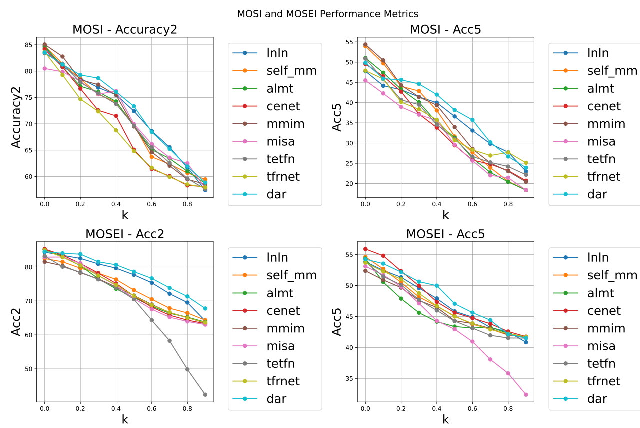

Figure 4: Variation of acc2 and acc5 of the model with training data of different missing rates

4.3 Robustness Comparison

Table 1 shows the robustness evaluation results on the MOSI and MOSEI datasets. As shown in Table 1, DAR achieves state-of-the-art performance on most metrics, demonstrating the robustness of DAR in the term of different noise effects. For seven categorical metrics on the mosi dataset MAE versus the mosei dataset, our model is able to achieve sub-optimal results. Considering the unpredictability of the impact of stochastic factors on the quality of missing data, and some of the extremes of the data have a huge impact on the overall results, in this case, given the inherent instability of missing data, we can assume that DAR achieves the optimal overall performance on both datasets compared to the other models compared.

Figure 4 shows the performance of all models under two of the most commonly used binomial and multiclassification metrics, non0acc2 and acc5, at different missing rates. The results show that although DAR loses part of its performance compared to other models when facing complete inputs, its

Method	Acc-7	Acc-5	Acc-2	F1	MAE↓	Corr
w/o L_{sim}	34.14	38.42	71.50 / 72.71	71.30 / 72.62	1.084	0.505
w/o L_{diff}	34.28	38.27	71.54 / 72.85	71.48 / 72.62	1.089	0.507
w/o L_{sim}&L_{diff}	34.15	38.35	71.32 / 72.46	71.10 / 72.35	1.113	0.504
w/o L_{recon}	33.57	38.31	71.02 / 72.20	70.45 / 71.13	1.123	0.493
w/o L_{boot}	33.03	36.93	70.50 / 72.26	69.90 / 71.80	1.123	0.475
Ours	**34.47**	**38.65**	**71.60 / 73.18**	**71.51 / 73.15**	**1.069**	**0.520**

Table 2: Effects of different component. Where L_{boot} denotes the task loss corresponding to the boot module.

Method	Acc-7	Acc-5	Acc-2	F1	MAE↓	Corr
k=0.0	31.84	35.45	68.99 / 71.03	66.39 / 63.10	1.069	0.514
k=0.2	33.18	37.88	70.57 / 71.06	70.57 / 70.56	1.160	0.502
k=0.4	32.95	36.39	71.22 / 72.73	70.98 / 72.62	1.078	0.515
k=0.6	30.12	32.58	70.63 / 71.99	70.27 / 71.75	1.118	0.475
k=0.8	24.56	24.64	69.16 / 70.96	67.50 / 69.51	1.173	0.460

Table 3: Performance of the model at different missing rates k in training process.

performance under other missing rates is significantly improved compared to other models without missing data, and also compared to TFR-Net and LNLN trained with missing data, which proves the effectiveness of our method.

4.4 Ablation Experiment

To evaluate the effectiveness of our proposed approach, we conduct a series of ablation experiments. These experiments systematically remove or modify key components of our model to assess their individual contributions to performance. By comparing the results of these ablations with the full model, we are able to quantify the impact of each design choice. This analysis provides a deeper understanding of the strengths and limitations of our method.

The effect of the ablation experiment is shown in Table 2. The results of the ablation experiments demonstrate the effectiveness of our proposed multimodal fusion framework based on the decomposition-reconstruction idea. Compared to the complete model, eliminating either similarity or dissimilarity loss causes information redundancy in the feature correction reconstruction process, which reduces the performance of the model to varying degrees.

Besides, we also verified the effect of eliminating the alignment loss and bootstrap loss in the incomplete feature reconstruction process on the model effectiveness, and the elimination of the alignment loss increases the uncertainty in the incomplete feature reconstruction process and affects the model performance. While eliminating the bootstrap loss causes the model to focus too much on the effect of the incomplete feature reconstruction, in order to minimize the difference losses between the incomplete input and the complete input after encoding. This leads to the degradation of the encoder's ability to extract features, the reduction of the variability of the extracted features, and ultimately impairing the model's ability. For these reasons, we believe that mapping the bootstrap loss forces the model encoders to learn to benefit the downstream tasks of the features, mitigating encoder degradation.

4.5 Missing Rates Sensitivity Experiment

During the training of the model, we found that the manually selected missing rate of the multimodal data has a critical impact on the training process, and the following demonstrates the specific impact of the missing rate on the model output results. We tested the performance of the model under different missing training sets constructed with different missing rates k. The results are shown in Table 3.

Analyzing the experimental results, it can be seen that the performance of the model appears to increase and then decrease overall as the missing rate increases. After analyzing the results, we believe that too low missing rate will lead to the missing data is not distinct enough from the original complete input data, and the model degenerates into an ordinary multimodal fusion model. In this case,the DAR model is unable to learn the ability of feature reconstruction, while too high missing rate will lead to the features being corrupted seriously, especially for the modal common features, which may lead to the fact that all the modal features corresponding to all modal features are after alignment under too high missing rate. The model is therefore unable to learn the ability to reconstruct complete features from incomplete features.

The experiments show that choosing the appropriate missing rate is very important for the final performance of the model, and the model should be robustly trained by choosing the appropriate missing rate according to the task features.

5 Conlusion

In this paper, we propose a novel method for multimodal sentiment analysis called Decoupled-Adaptive Reconstruction (DAR). The framework uses a reconstruction method based on feature decoupling, and adopts different reconstruction methods for the modal-common features and modal-independent features of the missing data according to their own properties, and achieves a more obvious improvement in the robustness test of the mosi and mosei datasets compared with the existing methods. In addition, we validate the effectiveness of our proposed feature decomposition-reconstruction framework through ablation experiments, showing that our method can alleviate problems such as information redundancy in the feature reconstruction process.

Finally, we explore the performance of the trained models with different levels of data missing rates, and the results show that choosing the appropriate data missing rate has an extremely important impact on the robust performance of the models. In this experiment, we only discuss the case of the same missing rate for multiple modalities, however, in practice, due to the different quality and noise immunity of different modalities, choosing different missing rates for different modalities or using methods that can adapt the missing rate is a more promising direction for future improvement.

References

[Bagher Zadeh *et al.*, 2018] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[Baltrusaitis *et al.*, 2018] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 59–66, 2018.

[Cheng *et al.*, 2020] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, pages 1779–1788. PMLR, 2020.

[Devlin, 2018] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Guo *et al.*, 2022] Jiwei Guo, Jiajia Tang, Weichen Dai, Yu Ding, and Wanzeng Kong. Dynamically adjust word representations using unaligned multimodal information. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 3394–3402, New York, NY, USA, 2022. Association for Computing Machinery.

[Han *et al.*, 2021] Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9192, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[Hazarika *et al.*, 2020] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1122–1131, New York, NY, USA, 2020. Association for Computing Machinery.

[Huang *et al.*, 2015] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

[Li *et al.*, 2023] Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6631–6640, June 2023.

[Li *et al.*, 2025] Mingcheng Li, Dingkang Yang, Yuxuan Lei, Shunli Wang, Shuaibing Wang, Liuzhen Su, Kun Yang, Yuzheng Wang, Mingyang Sun, and Lihua Zhang. A unified self-distillation framework for multimodal sentiment analysis with uncertain missing modalities. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2025.

[Liang *et al.*, 2020] Jingjun Liang, Ruichen Li, and Qin Jin. Semi-supervised multi-modal emotion recognition with cross-modal distribution matching. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 2852–2861, New York, NY, USA, 2020. Association for Computing Machinery.

[Liang *et al.*, 2021] Tao Liang, Guosheng Lin, Lei Feng, Yan Zhang, and Fengmao Lv. Attention is not enough: Mitigating the distribution discrepancy in asynchronous multimodal sequence fusion. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8128–8136, 2021.

[Lv *et al.*, 2021a] Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2554–2562, 2021.

[Lv *et al.*, 2021b] Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2554–2562, 2021.

[Mai *et al.*, 2020] Sijie Mai, Songlong Xing, and Haifeng Hu. Locally confined modality fusion network with a global perspective for multimodal human affective computing. *IEEE Transactions on Multimedia*, 22(1):122–137, 2020.

[Mao *et al.*, 2022] Huisheng Mao, Ziqi Yuan, Hua Xu, Wenmeng Yu, Yihe Liu, and Kai Gao. M-SENA: An integrated platform for multimodal sentiment analysis. In Valerio Basile, Zornitsa Kozareva, and Sanja Stajner, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 204–213, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[McFee *et al.*, 2015] Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *SciPy*, 2015.

[Mittal *et al.*, 2020] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 1359–1367, 2020.

[Oord *et al.*, 2018] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[Rahman *et al.*, 2020] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2359–2369, Online, July 2020. Association for Computational Linguistics.

[Tsai *et al.*, 2019] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019.

[Xia *et al.*, 2024] Yan Xia, Hai Huang, Jieming Zhu, and Zhou Zhao. Achieving cross modal generalization with multimodal unified representation. *Advances in Neural Information Processing Systems*, 36, 2024.

[Yang *et al.*, 2022] Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 1642–1651, New York, NY, USA, 2022. Association for Computing Machinery.

[Yang *et al.*, 2023] Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. ConFEDE: Contrastive feature decomposition for multimodal sentiment analysis. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7617–7630, Toronto, Canada, July 2023. Association for Computational Linguistics.

[Yu *et al.*, 2021] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 10790–10797, 2021.

[Yuan *et al.*, 2021] Ziqi Yuan, Wei Li, Hua Xu, and Wenmeng Yu. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 4400–4407, New York, NY, USA, 2021. Association for Computing Machinery.

[Yuan *et al.*, 2024] Ziqi Yuan, Yihe Liu, Hua Xu, and Kai Gao. Noise imitation based adversarial training for robust multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 26:529–539, 2024.

[Zadeh *et al.*, 2016] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016.

[Zadeh *et al.*, 2017] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[Zhang *et al.*, 2023] Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 756–767, Singapore, December 2023. Association for Computational Linguistics.

[Zhang *et al.*, 2024] Haoyu Zhang, Wenbin Wang, and Tianshu Yu. Towards robust multimodal sentiment analysis with incomplete data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS 2024)*, 2024.

[Zhao *et al.*, 2024] Fei Zhao, Chengcui Zhang, and Baocheng Geng. Deep multimodal data fusion. *ACM Comput. Surv.*, 56(9), April 2024.