MDF: A Modality-Aware Disentanglement and Fusion Framework for Multimodal Sentiment Analysis

Zhongquan Jian[1,*], Wenhan Lv[2,*], Yanhao Chen[2], Guanran Luo[3], Wentao Qiu[3], Shaopan Wang[3], Bingbing Hu[2,†], Qingqiang Wu[2,3,†]

[1]School of Computer and Data Science, Minjiang University, Fuzhou, China
[2]School of Film, Xiamen University, Xiamen, China
[3]School of Informatics, Xiamen University, Xiamen, China

Abstract

The homogeneity and heterogeneity across modalities are critical factors that influence multimodal fusion. In Multimodal Sentiment Analysis (MSA), the inherent textual information within the audio modality induces cross-modality homogeneity with the text modality. Conversely, the mutual independence between text and vision modalities results in their cross-modal heterogeneity. Although existing disentangle-based methods achieve notable performance gains by separating modality features into distinct subspaces, they overlook the characteristics of cross-modality heterogeneity and homogeneity among different modalities. To this end, we propose a novel **M**odality-aware **D**isentangle and **F**usion (**MDF**) framework to investigate the role of core modality features. Specifically, we first use text as the anchor to disentangle the audio modality and extract its unique modality-specific features, thereby establishing cross-modal heterogeneity among text, audio, and vision. We then introduce a Cross-Modality Heterogeneity Enhancement (CHE) module to refine these features, further reinforcing their heterogeneous characteristics. Finally, a Modality Adaptive Weighting (MAW) module is employed to dynamically assign weights to the text, sound, and vision modalities based on their potential contributions to sentiment prediction, achieving a more effective multimodal representation for MSA. Experimental evaluations on different benchmarks demonstrate MDF's superiority, with extensive ablation studies confirming its effectiveness.

Code — https://github.com/jian-projects/msa-mdf

Introduction

With the prevalence of social networks, individuals are increasingly utilizing various media, such as text, audio, and vision, to convey their sentiments. As a result, Multimodal Sentiment Analysis (MSA) has become a prominent research topic in artificial intelligence (Poria et al. 2023), with significant applications across diverse fields, including human-computer interaction (Wadley et al. 2022; Zhou et al. 2023), intelligent healthcare (Dhuheir et al. 2021), and beyond. In multimodal representation learning, distinct media exhibit varying representational formats and often contain complementary information. When integrated effectively, these diverse modality signals provide a richer and more comprehensive understanding of human sentiment. This presents the challenge of efficiently extracting and integrating multimodal information to enhance the performance of MSA systems.

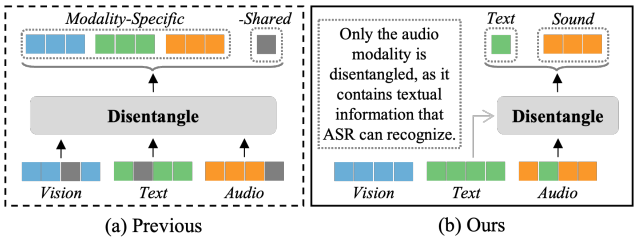


Figure 1: Comparison with previous disentanglement-based methods. From a human intuition perspective, vision consists of a sequence of images that differ significantly in format from text and audio, resulting in cross-modal heterogeneity. In contrast, audio and text exhibit cross-modal homogeneity, as audio inherently contains textual information.

Most recent MSA methods focus on two aspects (Poria et al. 2023): 1) the exploration of effective feature extraction methods to distill pertinent information from multimodal data (Hazarika, Zimmermann, and Poria 2020; Yang et al. 2022a), thereby enhancing the performance of the fusion process in terms of accuracy, reliability, and computational efficiency, and 2) the development of robust fusion strategies that integrate data knowledge from multimodal sources to maximize the complementary strengths and minimize the inherent weaknesses of each modality, including Tensor-based (Liu et al. 2018; Verma et al. 2020), MLP-based (Sun et al. 2022; Zhuang et al. 2024; Liu, Luo, and Fu 2025) and Attention-based (Tsai et al. 2019; Tang et al. 2019; Su et al. 2021; Zong et al. 2023; Feng et al. 2024; Wang, Ratnavelu, and Shibghatullah 2025) fusion methods. Recently, researchers have increasingly focused on representation learning-oriented approaches, with disentangle-based methods (Hazarika, Zimmermann, and Poria 2020; Yang et al. 2022a; Li, Wang, and Cui 2023; Liang et al.

*These authors contributed equally.
†Corresponding authors.

2023a; Zhu et al. 2025) gaining significant attentions for their abilities to separate modality-invariant and modality-specific features. This separation helps filter out irrelevant information, thereby facilitating feature fusion and enhancing the robustness of generated multimodal representations.

Despite their conceptual clarity and superior performance, the disentangled modality-invariant and modality-specific features are abstract and lack interpretability, as they are understood by machines rather than humans. As shown in Fig. 1, it remains unclear what the modality-invariant and modality-specific features represent, or how they contribute to sentiment prediction. From a human perspective, audio and vision are distinct because they are perceived through the sensory organs of hearing and sight, respectively, resulting in their cross-modality heterogeneity (Fan et al. 2024). Text, on the other hand, is an advanced form of language expression that relies on human linguistic ability and cognitive systems for interpretation and understanding. As is commonly known, the text is typically extracted from audio by the Automatic Speech Recognition (ASR) (Yu and Deng 2016) technique, meaning that the audio modality inherently contains textual information, leading to the cross-modality homogeneity with the text modality (Liu, Luo, and Fu 2025). Therefore, humans can intuitively understand and interpret the meanings of modality-specific features, while modality-invariant features remain abstract and imperceptible, limiting the potential of disentangle-based models.

Furthermore, modality fusion is another crucial step in MSA, as it combines information from multiple sources to maximize their collective strengths. Traditional fusion methods (Sun et al. 2022; Zhuang et al. 2024; Liu, Luo, and Fu 2025) often treat all modalities equally, learning their interactions implicitly. However, recent studies (Wang et al. 2023a; Feng et al. 2024) reveal that modalities contribute unequally, with text typically conveying the most sentiment information and dominating the final prediction, while vision and audio modalities often serve as complementary sources. To address this, attention-based methods have become prevalent (Singh, Abhishek, and Azad 2024), assigning adaptive weight to each modality to emphasize the most informative features, thereby enabling more effective multimodal representations. A representative example is AcFormer (Zong et al. 2023), which employs the cross-attention mechanism to compute attention weights based on inter-modal correlations. However, strong cross-modal heterogeneity, particularly between audio and visual modalities, makes it challenging to capture these correlations accurately. Another reasonable approach is to assign modality weights based on the uncertainty of each modality (Feng et al. 2024; Wang, Ratnavelu, and Shibghatullah 2025). However, without label guidance during inference, accurately estimating each modality's uncertainty becomes difficult, resulting in a mismatch between training and inference.

To address these challenges, we attempt to propose a **M**odality-aware **D**isentangle and **F**usion (**MDF**) framework for MSA. Building on the observation that audio and text modalities exhibit cross-modality homogeneity, while vision shows cross-modality heterogeneities with text and audio, MDF first disentangles the audio modality to extract its unique features (referred to as "sound" for distinction), creating cross-modality heterogeneities among text, sound, and vision. This disentangling process is intuitive and interpretable for humans. Subsequently, a Cross-modality Heterogeneity Enhancement (CHE) module is introduced to strengthen these heterogeneities by refining modality-specific features, enabling them to more accurately reflect the inherent independence of modalities in real-world scenarios. Cross-modal heterogeneities are essential for effective multimodal fusion, as they highlight the complementary strengths of modalities. For multimodal fusion, we develop a Modality Adaptive Weighting (MAW) module that treats text, sound, and vision as primary colors, blending them in adaptive proportions to generate rich and dynamic representations. Simply put, these proportions are determined by a weight generator. During training, we estimate modality confidences and use them as the supervision signals to train the generator, allowing it to learn modality contributions and generate appropriate fusion weights, thus ensuring consistency between training and inference.

In summary, our contributions are fourfold:

- We develop MDF, an effective modality disentangle and fusion framework for MSA. By disentangling the audio modality to extract its unique sound component, MDF creates cross-modality heterogeneities among text, sound, and vision, thereby facilitating multimodal fusion.
- We introduce the CHE module to reinforce cross-modal heterogeneities among text, sound, and vision, enhancing modality-specific features and emphasizing their complementary strengths.
- We design the MAW module to leverage the contribution of each modality as supervision, guiding the weight generator to learn modality importance and produce effective fusion weights.
- Experiments carried out on two datasets demonstrate the superiority of MDF. Extensive ablation studies and qualitative analyses confirm the effectiveness of MDF's disentangle process, as well as the weighted fusion process.

Related Work

MSA seeks to utilize information from multiple modalities, such as text, audio, and vision, to predict sentiment intensity or polarity. In the literature, mainstream MSA methods are broadly divided into two categories: representation learning-oriented approaches (Yu et al. 2021; Mai et al. 2023; Zeng et al. 2024), which emphasize extracting meaningful unimodal and multimodal features to enhance cross-modal interactions, and fusion-oriented approaches (Liu et al. 2018; Tsai et al. 2019; Zhang et al. 2023; Wang, Ratnavelu, and Shibghatullah 2025), which focus on designing effective strategies to combine these multimodal features.

Multimodal Representation Learning

With the rapid advancements in deep learning, particularly in pre-trained models and representation learning techniques, effective multimodal representations of data can be derived, thereby improving the performance of MSA. For

unimodal representations, Self-MM (Yu et al. 2021) leveraged pre-trained models to extract unimodal features and employed a self-supervised learning approach to obtain informative unimodal embeddings. For multimodal representations, several studies (Mai et al. 2023; Lin et al. 2022; Fan et al. 2024) have focused on utilizing contrastive learning to enhance multimodal interactions by drawing the anchor and positive samples closer, while pushing away the anchor and negative samples. Recently, numerous works (Hazarika, Zimmermann, and Poria 2020; Yang et al. 2022b; Zeng et al. 2024) proposed to disentangle different modalities into shared and unique components to improve the robustness of multimodal representations.

Disentangling learning aims to separate distinct features across multiple modalities into independent subspaces (Poria et al. 2023; Singh, Abhishek, and Azad 2024), thereby clarifying the beneficial factors and facilitating effective multimodal data fusion. In MSA, (Hazarika, Zimmermann, and Poria 2020; Yang et al. 2022a; Zhu et al. 2025; Liu, Luo, and Fu 2025) proposed learning modality-invariant and modality-specific features, while (Yang et al. 2022b) proposed learning modality-specific and modality-agnostic representations. Similarly, (Li, Wang, and Cui 2023) decoupled each modality into modality-irrelevant and modality-exclusive components. FactorCL (Liang et al. 2023b) further decomposed the representations into task-relevant shared and unique components, and PID (Liang et al. 2023a) extended this decomposition to encompass unique, redundant, and synergistic multimodal information.

Previous disentangle-based methods attempt to separate modality-invariant features, which are shared across modalities, and modality-specific features, which are unique to each modality. Despite their conceptual clarity, these features are understood by machines rather than by humans, limiting the model's practical applicability. In contrast, MDF disentangles the textual component from the audio modality, preserving the unique sound component specific to audio, and forming cross-modality heterogeneities among text, sound, and vision. Therefore, MDF is concrete and interpretable, as it clearly emphasizes the unique characteristics of each modality and explicitly evaluates their contributions.

Multimodal Fusion

Multimodal fusion is central to MSA, as it integrates information from diverse modalities, thereby enhancing the accuracy and robustness of sentiment prediction. In the literature, various fusion methods have been proposed, including: Tensor-based methods (Zadeh et al. 2017; Liu et al. 2018; Verma et al. 2020), which represent multimodal data as multi-dimensional tensors and employ tensor decomposition techniques to learn unified representations; Translation-based methods (Pham et al. 2019; Mai, Hu, and Xing 2020; Wang, Wan, and Wan 2020), which translate data between modalities (*e.g.*, from text to vision or audio), enabling the seamless fusion of cross-modal features; Graph-based methods (Yang et al. 2021; Qian et al. 2023), which model multimodal data as a graph where nodes represent features from each modality and edges capture inter-modal relationships, thereby facilitating effective fusion and sentiment inference;

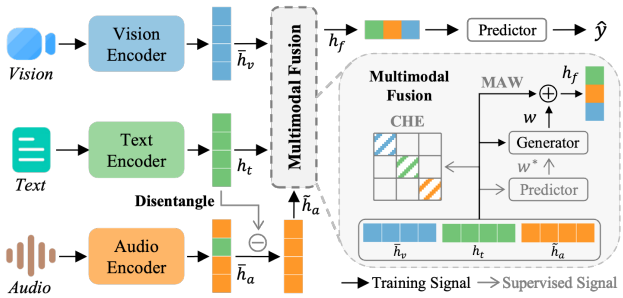


Figure 2: The Overview of MDF. Three modality signals are inputted, with the audio features disentangled to achieve its essential sound part. Subsequently, the text, sound, and vision are treated as primary colors and combined in varying proportions to generate rich and dynamic multimodal representations for MSA.

MLP-based methods (Sun et al. 2022; Zhuang et al. 2024; Liu, Luo, and Fu 2025), which employ Multi-Layer Perceptrons (MLPs) to learn the interactions between different modalities; and Attention-based methods (Gu et al. 2018; Akhtar et al. 2019; Tsai et al. 2019; Hazarika, Zimmermann, and Poria 2020; Han et al. 2021; Yang et al. 2022a; Zong et al. 2023; Feng et al. 2024; Wang, Ratnavelu, and Shibghatullah 2025), which assign dynamic weights to different parts of multimodal data, enabling the model to focus on the most informative features from each modality.

Vector addition is the simplest and most efficient feature fusion method (Bengio 2009) in the same dimension space. The challenge lies in determining the appropriate weights for modalities, as numerous studies (Feng et al. 2024; Wang, Ratnavelu, and Shibghatullah 2025) have shown that different modalities contribute unequally to the final prediction. In KuDA (Feng et al. 2024), a dynamic attention fusion module was introduced to directly assign different weights to different modalities based on the uncertainty of unimodal predictions. Conversely, we train a generator to derive modality weights, supervised by their actual contributions to sentiment prediction, enabling the model to perceive and assign the correct weight to each modality. Features from different modalities are treated as primary colors and, after being weighted by their corresponding contributions, are fused into rich and dynamic multimodal representations.

Methodology

Model Overview

Generally, MSA can be framed as either a regression task aimed at predicting sentiment intensity or a classification task focused on predicting sentiment polarity. Following (Yu et al. 2021), we approach MSA as a regression task and additionally report the corresponding sentiment polarity based on the predicted sentiment intensity. Fig. 2 provides an overview of MDF, where text (X_t), audio (X_a), and vision (X_v) serve as inputs to predict the specific sentiment intensity $y \in \mathbb{R}$, forming a sample represented as a quadruple (X_t, X_a, X_v, y).

Modality Encoding and Mapping

With the great success of deep learning (Wang et al. 2023b), we leverage pre-trained models (Vaswani et al. 2017) and toolkits to extract initial features from raw modality signals. Specifically, for the text modality, BERT (Devlin et al. 2019) is employed to encode token embeddings, with the embedding of the [CLS] token, *i.e.*, the first vector from the last layer, selected as the overall text representation:

$$h_t = \textit{BERT}(X_t)[0] \tag{1}$$

where $h_t \in \mathbb{R}^{d_t}$ denotes the encoded text representation, and d_t is the vector dimension.

Following Self-MM (Yu et al. 2021), existing pre-trained toolkits can be utilized to extract the initial features from the audio and vision modalities, denoted as $H_a \in \mathbb{R}^{l_a \times d_a}$ and $H_v \in \mathbb{R}^{l_v \times d_v}$, respectively. Here, l_a and l_v represent the sequence lengths of the audio and vision features, while d_a and d_v denote their vector dimensions. Then, two bidirectional LSTMs are applied separately to capture the temporal characteristics of these sequential features:

$$h_a = \textit{BiLSTM}_a(H_a); \quad h_v = \textit{BiLSTM}_v(H_v) \tag{2}$$

where $\textit{BiLSTM}_*$ represents a bidirectional LSTM model that captures the temporal dependencies of the input sequences. $h_a \in \mathbb{R}^{d_a}$ and $h_v \in \mathbb{R}^{d_v}$ denote the returned audio and vision representations, respectively. Then, these modality features are mapped into a unified representation space:

$$\overline{h}_a = \textit{Mapper}_a(h_a); \quad \overline{h}_v = \textit{Mapper}_v(h_v) \tag{3}$$

where $\textit{Mapper}_a : \mathbb{R}^{d_a} \rightarrow \mathbb{R}^{d_t}$ and $\textit{Mapper}_v : \mathbb{R}^{d_v} \rightarrow \mathbb{R}^{d_t}$ are two MLPs that map different input modality representations into the text representation space.

Modality Disentangle and Enhancement

From the machine perspective, representations from different modalities typically vary in nature because they are encoded by distinct pre-trained models, resulting in differences in knowledge type and content across modalities. From the human perspective, audio and vision are distinct modalities perceived through the sensory organs of hearing and sight, resulting in their inherent cross-modality heterogeneity. In contrast, audio and text modalities exhibit cross-modality homogeneity, as the audio modality inherently contains textual information. Vision can also carry textual information, as indicated by text-to-image and image-to-text techniques (Radford et al. 2021). However, their connection is fragile, as there is no direct one-to-one correspondence between text and vision, unlike the more straightforward alignment between audio and text.

In our work, we attempt to disentangle the audio modality to isolate its unique characteristics and investigate the influence of different modalities on MSA, aiming to enhance the model's interpretability from a human perspective.

Audio Modality Disentangle As is well known, text is typically extracted from audio using the ASR technique, meaning that the audio modality inherently contains textual information, resulting in cross-modality homogeneity between the audio and text modalities. MDF attempts to disentangle the unique characteristics of audio. A straightforward approach is to subtract text representation from audio representation, *i.e.,* $\overline{h}_a - h_t$. However, this method worked based on a strong assumption that both text and audio representations come from the same pre-trained model, with the audio representation containing more information than the text. Generally, the unified multimodal model can address the encoding differences between various modalities, but the abundance of text training data compared to audio data often leads to text representation outperforming audio representation. Hence, the aforementioned method, while reasonable, appears infeasible in our scenario.

In this section, we propose an indirect approach to achieve our objective. Following FDMER (Yang et al. 2022a), we utilize an MLP to extract the textual content from the audio modality representation $\overline{h}_a$:

$$\tilde{h}_t = \textit{Extractor}(\overline{h}_a) \tag{4}$$

where $\textit{Extractor} : \mathbb{R}^{d_t} \rightarrow \mathbb{R}^{d_t}$ is an MLP employed to separate the textual component $\tilde{h}_t$ from $\overline{h}_a$. Consequently, the unique sound component is then obtained by subtracting the textual component from the audio representation:

$$\tilde{h}_a = \overline{h}_a - \tilde{h}_t \tag{5}$$

Here, we impose constraints on the attributes of the textual component $\tilde{h}_t$ to ensure that the sound component $\tilde{h}_a$ effectively represents the unique characteristics of audio. As previously noted, the extracted textual component differs from the actual text representation since they are generated by different encoders. Although we map the audio representation into the text representation space, a simple MLP may struggle to bridge the information gap between the textual component $\tilde{h}_t$ separated from the audio and the actual text representation h_t. However, they should at least share similar characteristics, *i.e.*, exhibit equal contributions to sentiment prediction. Hence, an alignment loss is introduced to align their predicted sentiment intensities:

$$\mathcal{L}_a = ||\textit{Predictor}(\tilde{h}_t) - \textit{Predictor}(h_t)||_2 \tag{6}$$

where $\textit{Predictor} : \mathbb{R}^{d_t} \rightarrow \mathbb{R}^1$ is an MLP that serves as the sentiment predictor, and $||*||_2$ denotes the L_2 norm function.

Cross-modality Heterogeneity Enhancement After disentangling the audio modality to isolate its unique sound component, the text, sound, and vision form cross-modality heterogeneities, each exhibiting distinct characteristics. To preserve this, as shown in Fig. 2, we introduce an orthogonal constraint to reinforce their cross-modal heterogeneity:

$$\mathcal{L}_o = \sum_{h \in H} \sum_{h' \in H} \mathbb{I}_{h \neq h'} \cdot \textit{LN}(h) \otimes \textit{LN}(h') \tag{7}$$

where $H = [h_t, \tilde{h}_a, \overline{h}_v]$ is the set of modality representations, $\otimes$ denotes the dot product operation, and *LN* represents the Layer Normalization function. $\mathbb{I}_*$ denotes an identity function that returns 1 if h and h' are different, and 0 otherwise. This loss encourages the model to learn modality-specific features that are orthogonal to one another, thereby

reflecting their inherent nature of mutual independence in the real world, and further enhancing their complementary strengths in multimodal fusion.

Modality Adaptive Weighting Fusion

Given the nature of regression tasks, we treat the text, sound, and vision components as primary colors, combining them in varying proportions to generate rich and dynamic multimodal representations that correspond to the sentiment intensity, which is represented by a real number.

Modality Weighting Fusion Vector addition is the simplest and most efficient method for feature fusion (Bengio 2009). However, treating each modality equally may not be optimal, as different modalities contribute differently to sentiment prediction (Feng et al. 2024). Hence, to put it simply, we develop a weight generator to directly assign appropriate weights to modalities:

$$w = \textit{Softmax}(\textit{Generator}(h_c)) \tag{8}$$

where $h_c = [h_t; \tilde{h}_a; \overline{h}_v] \in \mathbb{R}^{3d_t}$ represents the concatenation of the text, sound, and vision representations, and $\textit{Generator} : \mathbb{R}^{3d_t} \rightarrow \mathbb{R}^3$ is an MLP that generates the corresponding weights for different modalities to achieve their adaptive weighting fusion. *Softmax* is the softmax function that maps the weights to probabilities summing to 1. Therefore, $w = [w_t, w_a, w_v]$ indicates the weights of the text, sound, and vision modalities, respectively. The multimodal representation is then computed as:

$$h_f = w \otimes H = w_t h_t + w_a \tilde{h}_a + w_v \overline{h}_v \tag{9}$$

where $h_f \in \mathbb{R}^{d_t}$ represents the fused multimodal representation that considers the varying contributions of modalities.

Weight Generator Training Although the weight generator can be trained by backpropagating the error from the predicted sentiment intensity, the generated modality weights lack interpretability and may not accurately reflect the true contributions of different modalities. To address this, we first leverage the label signal y to evaluate the uncertainty of each modality, *i.e.*, $|y - \hat{y}_m|$, where $\hat{y}_m$ is the predicted sentiment intensity derived solely from modality m. Then, these uncertainty scores are transformed into weights to reflect the true contributions of different modalities to sentiment prediction:

$$w^* = \textit{Softmax}(\{(|y - \textit{Predictor}(h)| + \epsilon)^{-1}\}_{h \in H}) \tag{10}$$

Intuitively, the lower the uncertainty, the greater the modality's importance. Hence, we use the reciprocal of the uncertainty to assess the modality's importance. ϵ is a small constant to prevent division by zero. Therefore, w^* describes the true contributions of modalities, and is utilized as the ground truth to supervise the training of the weight generator:

$$\mathcal{L}_w = ||w^* - w||_2 \tag{11}$$

Notably, w^* varies dynamically during training, guiding the weight generator to learn the patterns of modality combinations for sentiment prediction. This enables the generator to produce proper weights (w) that reflect the potential contributions of different modalities, thereby achieving the robust and effective fusion of multimodal representations.

Training Objective

With the multimodal representation h_f generated by Eq. (9), the sentiment intensity is predicted as follows:

$$\hat{y} = \textit{Predictor}(h_f) \tag{12}$$

Following prior studies (Zadeh et al. 2017; Hazarika, Zimmermann, and Poria 2020; Sun et al. 2022), the mean squared error loss is employed for the regression task:

$$\mathcal{L}_{\textit{task}} = ||y - \hat{y}||_2 \tag{13}$$

Finally, the training objective is to minimize the overall loss:

$$\mathcal{L} = \mathcal{L}_{\textit{task}} + \mathcal{L}_a + \mathcal{L}_o + \mathcal{L}_w \tag{14}$$

Experimental Setups

Datasets

MDF is evaluated on two widely used MSA datasets, namely CMU-MOSI and CMU-MOSEI. The descriptions and statistics of these datasets are provided in the Appendix.

Evaluation Metrics

Following the previous works (Yu et al. 2021; Yang et al. 2023), a set of widely-recognized metrics is adopted, including Mean Absolute Error (MAE), Pearson Correlation Coefficient (Corr), 7-class accuracy (Acc-7), binary accuracy (Acc-2), and F1-Score. Particularly, the Acc-2 and F1 scores are reported in two forms using the segmentation marker '-/-': the first score represents negative/non-negative performance, while the second score corresponds to negative/positive performance. The distinction between non-negative and positive scores lies in the fact that the former includes scores ≥ 0, while the latter encompasses scores > 0.

Compared Methods

MDF is compared with recent advanced MSA methods: LMF (Liu et al. 2018), MulT (Tsai et al. 2019), MISA (Hazarika, Zimmermann, and Poria 2020), Self-MM (Yu et al. 2021), CubeMLP (Sun et al. 2022), ConFEDE (Yang et al. 2023), AcFormer (Zong et al. 2023), EMT (Sun et al. 2024), HyDiscGAN (Wu et al. 2024), KuDA (Feng et al. 2024), DNT (Zeng et al. 2024), ULMD (Zhu et al. 2025), DLF (Wang et al. 2025), DEVA (Wu et al. 2025) and MFON (Zhang, Wei, and Zou 2025).

Implementation Details

For a fair comparison, we employ the bert-base-uncased model as the text encoder, with the text representation space having a dimension of 768. Following (Yu et al. 2021; Feng et al. 2024), we directly utilize the features of the audio and vision modalities as provided in the original datasets. During model training, the AdamW optimizer is used with an initial learning rate $\{5e^{-5}, 1e^{-3}\}$ for BERT and other parameters, respectively. The batch sizes and epochs are set to $\{32, 64\}$ and $\{50, 30\}$ for CMU-MOSI and CMU-MOSEI, respectively. All experiments are conducted using a single NVIDIA RTX 3090 GPU with 24GB memory.

Methods	CMU-MOSI					CMU-MOSEI				
	$MAE_{(\downarrow)}$	Corr	Acc-7	Acc-2	F1-Score	$MAE_{(\downarrow)}$	Corr	Acc-7	Acc-2	F1-Score
LMF♣	0.950	0.651	33.82	77.90/79.18	77.80/79.15	0.576	0.717	51.59	80.54/83.48	80.94/83.66
MulT♣	0.879	0.702	36.91	79.71/80.98	79.63/80.95	0.559	0.733	52.84	81.15/84.63	81.56/84.52
†MISA♣	0.776	0.778	41.37	81.84/83.54	81.82/83.58	0.557	0.751	52.05	80.67/84.67	81.12/84.66
Self-MM♣	0.708	0.796	46.67	83.44/85.46	83.36/85.43	0.531	0.764	53.87	83.76/85.15	83.82/84.90
CubeMLP♣	0.755	0.772	43.44	80.76/82.32	81.77/84.23	0.537	0.761	53.35	82.36/85.23	82.61/85.04
ConFEDE	0.742	0.784	42.27	84.17/85.52	84.13/85.52	**0.522**	0.780	**54.86**	81.65/85.82	82.17/85.83
AcFormer	0.715	0.794	44.2	82.3/85.4	82.1/85.2	0.531	0.786	54.7	84.3/86.5	**84.2**/85.8
EMT	0.705	0.798	47.4	83.3/85.0	83.2/85.0	0.527	0.774	54.5	83.4/86.0	83.7/86.0
HyDiscGAN	0.749	0.782	43.2	84.1/86.7	83.7/86.3	0.533	0.761	54.4	81.9/86.3	82.1/86.2
KuDA♣	0.705	0.795	47.08	84.40/86.43	84.48/86.46	0.529	0.776	52.89	83.26/86.46	82.97/**86.59**
†DTN	0.714	0.807	48.1	-/86.2	-/86.2	0.579	**0.788**	52.5	-/86.3	-/86.3
†ULMD	0.700	0.799	47.81	-/85.82	-/85.71	0.531	0.770	53.81	-/85.95	-/85.91
†DLF	0.731	0.781	47.08	-/85.06	-/85.04	0.536	0.764	53.90	-/85.42	-/85.27
DEVA	0.730	0.787	46.32	84.40/86.29	84.48/86.30	0.541	0.769	52.26	83.26/86.13	82.93/86.21
MFON	0.725	0.797	44.90	84.84/**86.89**	84.75/**86.86**	0.528	0.780	53.72	82.70/86.32	83.13/86.29
†**MDF(ours)**	**0.692**	**0.810**	**48.25**	**84.99**/85.82	**85.03**/85.90	0.525	**0.788**	54.80	**85.01/86.61**	83.92/86.41

Table 1: Comparison with the advanced baselines. ♣indicates results excerpted from KuDA, while others are from their original papers. †indicates methods based on the concept of disentangling.

Experimental Results and Analysis

Compared With Advanced Methods

Table 1 presents the comparative performance of MDF with recent advanced methods on two benchmark datasets, marking the best and second-best results in **bold** and underlined, respectively. Note that, the compared methods are all BERT-based, and thus we utilize the bert-base pre-trained model as the text encoder for a fair comparison. Overall, MDF surpasses recent advanced methods, achieving the best or second-best performance across most evaluation metrics on all datasets, establishing new benchmarks on the CMU-MOSI dataset in MAE, Corr, and Acc-7 metrics.

From Table 1, it is evident that representation learning plays a crucial role in constructing effective multimodal representations for MSA. Among them, ConFEDE employs the contrastive learning mechanism to extract robust modality features and demonstrates strong performance on the CMU-MOSEI dataset, achieving the best results in the MAE and Acc-7 metrics. MFON also utilizes contrastive learning to enhance the model's performance, achieving competitive results on the CMU-MOSI datasets. ULMD and DTN disentangle modality features into common and private spaces to improve multimodal fusion, thereby delivering competitive performance on both datasets. MDF surpasses these representation learning-oriented methods, including MISA, DTN, ULMD, and DLF, highlighting its superiority in disentangling audio modality-specific features and effectively integrating cross-modality heterogeneous features for MSA.

Additionally, compared to MLP-based fusion methods, such as EMT, ConFEDE, CubeMLP, and Self-MM, the weighted sum fusion strategy indicates greater effectiveness, as evidenced by the superior performance of AcFormer (utilizing a cross-attention mechanism) on the CMU-MOSEI dataset and KuDA (employing a weight assignment mechanism) on both datasets. Compared with KuDA, MDF trains a weight generator to assign the correct weights to different modalities, thereby enhancing the model's ability to recognize the dominant modality and achieving better results.

$\mathcal{L}_{task}$	$\mathcal{L}_a$	$\mathcal{L}_o$	$\mathcal{L}_w$	CMU-MOSI		CMU-MOSEI	
				$MAE_\downarrow$	$Corr_\uparrow$	$MAE_\downarrow$	$Corr_\uparrow$
✓				0.720	0.781	0.536	0.760
✓	✓			0.718	0.780	0.536	0.769
✓	✓	✓		0.710	0.798	0.530	0.774
✓			✓	0.713	0.793	0.532	0.775
✓	✓		✓	0.708	0.803	0.530	0.780
✓	✓	✓	✓	0.692	0.810	0.525	0.788

Table 2: Model performance with different losses.

Ablation Studies

Influences of Different Modules Table 2 presents the ablation results of introducing different losses in MDF, including $\mathcal{L}_a$ for disentangling the audio modality, $\mathcal{L}_o$ for enhancing modality-specific features, and $\mathcal{L}_w$ for training the weight generator. It can be observed that the model performance is significantly affected by the removal of any loss term, with the most pronounced performance degradation observed when all loss terms are removed. Particularly, the MAW module ($\mathcal{L}_w$) is pivotal to the model's performance, significantly enhancing its effectiveness across various settings. Additionally, the CHE module ($\mathcal{L}_o$) activates the ability of audio disentangle ($\mathcal{L}_a$), amplifying the performance gains achieved through audio disentangle.

Contributions of Different Modalities The experimental results summarized in Table 3 analyze the impact of different

Modalities	CMU-MOSI		CMU-MOSEI	
	MAE$_\downarrow$	Corr$_\uparrow$	MAE$_\downarrow$	Corr$_\uparrow$
A (only)	1.455	0.162	0.998	0.184
A' (only)	1.456	0.145	1.005	0.163
T (only)	0.786	0.705	0.566	0.740
$T + A$	0.739	0.777	0.547	0.752
$T + A'$	0.722	0.795	0.535	0.768
V (only)	1.480	0.114	1.003	0.112
$V + T + A$	0.720	0.781	0.536	0.760
$V + T + A'$	0.692	0.810	0.525	0.788

Table 3: Model performance with different modalities.

modalities on the model's performance by progressively incorporating them, where T, A, V represent the original text, audio, and video modality features, respectively. A' represents the audio-specific features after disentangling, *i.e.*, the sound component. Our results under unimodal settings are consistent with previous studies (Zong et al. 2023), which demonstrate the dominance of the text modality in MSA.

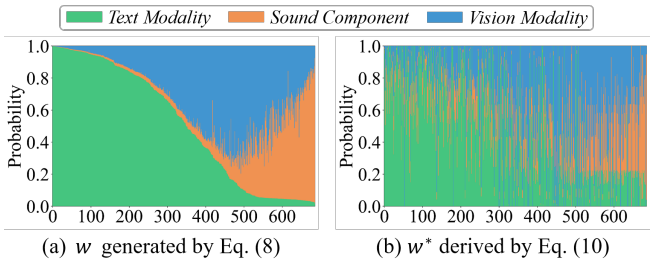
As seen, using only the sound component (A') slightly underperforms the original audio (A), indicating that the original audio modality contains additional textual information distinct from the original text modality. This confirms that we cannot directly align the text component disentangled from the audio modality with the text modality, which is why we align their characteristics in Eq. (6). However, fusing A' with other modalities significantly outperforms counterparts fused with A, highlighting the effectiveness of the disentangled sound component in complementing the other modalities, particularly the combination of $V + T + A'$, which achieves the best performance among all combinations.

Visualization Analysis

To further demonstrate MDF's effectiveness, we visualize the effects of different modules, with experiments conducted on the CMU-MOSI dataset. More visualization results can be found in the Appendix.

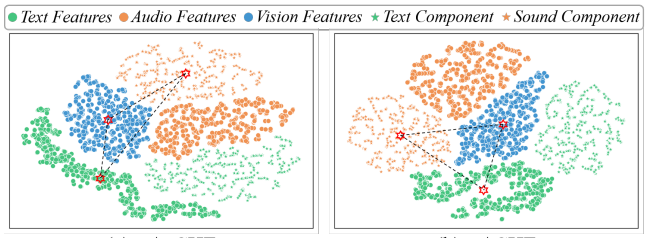
Modality Weights Distributions Fig. 3 visualizes and compares the generated modality weights w by Eq. (8) alongside the transformed weights w^* by Eq. (10). For clarity, the samples are first ranked based on the values w in the text modality, as shown in Fig. 3(a). Correspondingly, the samples in Fig. 3(b) are ranked using the same indices.

As observed, Fig. 3(a) indicates that text dominates the MSA task, with most samples exhibiting higher text modality weights. Samples with lower text modality weights tend to have higher weights on the sound component, indicating that sound effectively supplements the text. In contrast, the vision modality plays a more supportive role when the text modality is dominant, as reflected in its higher weights compared to sound in the lower-ranked samples. Additionally, the weights in Fig. 3(b) represent the modality contributions based on their unimodal features. A comparison between Fig. 3(a) and Fig. 3(b) shows that the generated weights w generally align with the derived weights w^*, demonstrating the effectiveness of the MAW module in recognizing the true contributions of modalities to sentiment prediction.



(a) w generated by Eq. (8) (b) w^* derived by Eq. (10)

Figure 3: Visualization of modality weights.



(a) w/o CHE (b) w/ CHE

Figure 4: Visualization of cross-modality heterogeneity.

Modality Feature Distributions Using the t-SNE toolkit, we visualize the distributions of different modality features in Fig. 4, with features extracted from models trained with and without the CHE module. Since different modalities represent distinct information types, Fig. 4(a) shows that modality features are clearly separated, supporting the assumption of cross-modality heterogeneity. After disentangling, the text component and sound component are separated obviously, and the text component aligns closely with the text modality, demonstrating the effectiveness of our proposed indirect alignment method (*i.e.*, $\mathcal{L}_a$). Furthermore, the introduced CHE module successfully enhances cross-modality heterogeneity among the text, sound, and vision, as demonstrated in Fig. 4(b), where the corresponding features exhibit a three-legged confrontation. This enhancement is crucial for strengthening the complementarity of modalities and improving the multimodal fusion process, as illustrated below.

Conclusion

We propose MDF to extract and refine modality-specific features and fuse them via adaptive weights for MSA. MDF first disentangles the audio modality to extract its unique sound component, then employs a CHE module to enhance cross-modal heterogeneity among text, sound, and vision. Building on this, an MAW module further guides a weight generator to assign modality-specific fusion weights. Together, these components address both cross-modality heterogeneity and homogeneity challenges in MSA.

Acknowledgments

This work is supported by the Pre-research Project for Introduced Talents of Minjiang University (No.MJY25025) and the Public Technology Service Platform Project of Xiamen City (No.3502Z20231043).

References

Akhtar, M. S.; Chauhan, D.; Ghosal, D.; Poria, S.; Ekbal, A.; and Bhattacharyya, P. 2019. Multi-task Learning for Multi-modal Emotion Recognition and Sentiment Analysis. In *NAACL*, 370–379.

Bengio, Y. 2009. *Learning Deep Architectures for AI*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, 4171–4186.

Dhuheir, M.; Albaseer, A.; Baccour, E.; Erbad, A.; Abdallah, M.; and Hamdi, M. 2021. Emotion recognition for healthcare surveillance systems using neural networks: A survey. In *IWCMC*, 681–687.

Fan, C.; Zhu, K.; Tao, J.; Yi, G.; Xue, J.; and Lv, Z. 2024. Multi-level Contrastive Learning: Hierarchical Alleviation of Heterogeneity in Multimodal Sentiment Analysis. *IEEE Trans. Affect. Comput.*, 1–17.

Feng, X.; Lin, Y.; He, L.; Li, Y.; Chang, L.; and Zhou, Y. 2024. Knowledge-Guided Dynamic Modality Attention Fusion Framework for Multimodal Sentiment Analysis. In *EMNLP*, 14755–14766.

Gu, Y.; Yang, K.; Fu, S.; Chen, S.; Li, X.; and Marsic, I. 2018. Multimodal Affective Analysis Using Hierarchical Attention Strategy with Word-Level Alignment. In *ACL*, 2225–2235.

Han, W.; Chen, H.; Gelbukh, A.; Zadeh, A.; Morency, L.-p.; and Poria, S. 2021. Bi-Bimodal Modality Fusion for Correlation-Controlled Multimodal Sentiment Analysis. In *ICMI*, 6–15.

Hazarika, D.; Zimmermann, R.; and Poria, S. 2020. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. In *MM*, 1122–1131.

Li, Y.; Wang, Y.; and Cui, Z. 2023. Decoupled multimodal distilling for emotion recognition. In *CVPR*, 6631–6640.

Liang, P. P.; Cheng, Y.; Fan, X.; Ling, C. K.; Nie, S.; Chen, R. J.; Deng, Z.; Allen, N. B.; Auerbach, R.; Mahmood, F.; Salakhutdinov, R.; and Morency, L. 2023a. Quantifying & Modeling Multimodal Interactions: An Information Decomposition Framework. In *NeurIPS*.

Liang, P. P.; Deng, Z.; Ma, M. Q.; Zou, J. Y.; Morency, L.; and Salakhutdinov, R. 2023b. Factorized Contrastive Learning: Going Beyond Multi-view Redundancy. In *NeurIPS*.

Lin, Z.; Liang, B.; Long, Y.; Dang, Y.; Yang, M.; Zhang, M.; and Xu, R. 2022. Modeling Intra- and Inter-Modal Relations: Hierarchical Graph Contrastive Learning for Multimodal Sentiment Analysis. In *COLING*, 7124–7135.

Liu, S.; Luo, Z.; and Fu, W. 2025. Fcdnet: Fuzzy Cognition-Based Dynamic Fusion Network for Multimodal Sentiment Analysis. *IEEE Trans. Fuzzy Syst.*, 33(1): 3–14.

Liu, Z.; Shen, Y.; Lakshminarasimhan, V. B.; Liang, P. P.; Bagher Zadeh, A.; and Morency, L.-P. 2018. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In *ACL*, 2247–2256.

Mai, S.; Hu, H.; and Xing, S. 2020. Modality to Modality Translation: An Adversarial Representation Learning and Graph Fusion Network for Multimodal Fusion. In *AAAI*, 164–172.

Mai, S.; Zeng, Y.; Zheng, S.; and Hu, H. 2023. Hybrid Contrastive Learning of Tri-Modal Representation for Multimodal Sentiment Analysis. *IEEE Trans. Affect. Comput.*, 14(3): 2276–2289.

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 *AAAI*, volume 33, 6892–6899.

Poria, S.; Hazarika, D.; Majumder, N.; and Mihalcea, R. 2023. Beneath the Tip of the Iceberg: Current Challenges and New Directions in Sentiment Analysis Research. *IEEE Trans. Affect. Comput.*, 14(1): 108–132.

Qian, S.; Xue, D.; Fang, Q.; and Xu, C. 2023. Integrating Multi-Label Contrastive Learning With Dual Adversarial Graph Neural Networks for Cross-Modal Retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(4): 4794–4811.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, volume 139, 8748–8763.

Singh, U.; Abhishek, K.; and Azad, H. K. 2024. A Survey of Cutting-edge Multimodal Sentiment Analysis. *ACM Comput. Surv.*, 56(9).

Su, J.; Tang, J.; Jiang, H.; Lu, Z.; Ge, Y.; Song, L.; Xiong, D.; Sun, L.; and Luo, J. 2021. Enhanced aspect-based sentiment analysis models with progressive self-supervised attention learning. *Artif. Intell.*, 296: 103477.

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 *MM*, 3722–3729.

Sun, L.; Lian, Z.; Liu, B.; and Tao, J. 2024. Efficient Multimodal Transformer With Dual-Level Feature Restoration for Robust Multimodal Sentiment Analysis. *IEEE Trans. Affect. Comput.*, 15(1): 309–325.

Tang, J.; Lu, Z.; Su, J.; Ge, Y.; Song, L.; Sun, L.; and Luo, J. 2019. Progressive Self-Supervised Attention Learning for Aspect-Level Sentiment Analysis. In *ACL*, 557–566.

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 *ACL*, 6558–6569.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*, 6000–6010.

Verma, S.; Wang, J.; Ge, Z.; Shen, R.; Jin, F.; Wang, Y.; Chen, F.; and Liu, W. 2020. Deep-HOSeq: Deep Higher

Order Sequence Fusion for Multimodal Sentiment Analysis. In *ICDM*, 561–570.

Wadley, G.; Kostakos, V.; Koval, P.; Smith, W.; Webber, S.; Cox, A.; Gross, J. J.; Höök, K.; Mandryk, R.; and Slovák, P. 2022. The future of emotion in human-computer interaction. In *CHI EA*, 1–6.

Wang, D.; Guo, X.; Tian, Y.; Liu, J.; He, L.; and Luo, X. 2023a. TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis. *Pattern Recognit.*, 136: 109259.

Wang, P.; Zhou, Q.; Wu, Y.; Chen, T.; and Hu, J. 2025. DLF: Disentangled-Language-Focused Multimodal Sentiment Analysis. In *AAAI*.

Wang, S.; Ratnavelu, K.; and Shibghatullah, A. S. B. 2025. UEFN: Efficient uncertainty estimation fusion network for reliable multimodal sentiment analysis. *Appl. Intell.*, 55(2): 171.

Wang, X.; Chen, G.; Qian, G.; Gao, P.; Wei, X.; Wang, Y.; Tian, Y.; and Gao, W. 2023b. Large-scale Multi-modal Pre-trained Models: A Comprehensive Survey. *Mach. Intell. Res.*, 20(4): 447–482.

Wang, Z.; Wan, Z.; and Wan, X. 2020. TransModality: An End2End Fusion Method with Transformer for Multimodal Sentiment Analysis. In *WWW*, 2514–2520.

Wu, S.; Wang, X.; Wang, L.; He, D.; and Dang, J. 2025. Enriching Multimodal Sentiment Analysis through Textual Emotional Descriptions of Visual-Audio Content. In *AAAI*.

Wu, Z.; Zhang, Q.; Miao, D.; Yi, K.; Fan, W.; and Hu, L. 2024. HyDiscGAN: A Hybrid Distributed cGAN for Audio-Visual Privacy Preservation in Multimodal Sentiment Analysis. In *IJCAI*, 6550–6558.

Yang, D.; Huang, S.; Kuang, H.; Du, Y.; and Zhang, L. 2022a. Disentangled Representation Learning for Multimodal Emotion Recognition. In *MM*, 1642–1651.

Yang, D.; Kuang, H.; Huang, S.; and Zhang, L. 2022b. Learning Modality-Specific and -Agnostic Representations for Asynchronous Multimodal Language Sequences. In *MM*, 1708–1717.

Yang, J.; Wang, Y.; Yi, R.; Zhu, Y.; Rehman, A.; Zadeh, A.; Poria, S.; and Morency, L.-P. 2021. MTAG: Modal-Temporal Attention Graph for Unaligned Human Multimodal Language Sequences. In *NAACL*, 1009–1021.

Yang, J.; Yu, Y.; Niu, D.; Guo, W.; and Xu, Y. 2023. ConFEDE: Contrastive Feature Decomposition for Multimodal Sentiment Analysis. In *ACL*, 7617–7630.

Yu, D.; and Deng, L. 2016. *Automatic speech recognition*, volume 1. Berlin: Springer.

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 *AAAI*, volume 35, 10790–10797.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *EMNLP*, 1103–1114.

Zeng, Y.; Yan, W.; Mai, S.; and Hu, H. 2024. Disentanglement Translation Network for multimodal sentiment analysis. *Inf. Fusion*, 102: 102031.

Zhang, H.; Wang, Y.; Yin, G.; Liu, K.; Liu, Y.; and Yu, T. 2023. Learning Language-guided Adaptive Hyper-modality Representation for Multimodal Sentiment Analysis. In *EMNLP*, 756–767.

Zhang, X.; Wei, W.; and Zou, S. 2025. Modal Feature Optimization Network with Prompt for Multimodal Sentiment Analysis. In *COLING*, 4611–4621.

Zhou, C.; Liang, Y.; Meng, F.; Zhou, J.; Xu, J.; Wang, H.; Zhang, M.; and Su, J. 2023. A Multi-Task Multi-Stage Transitional Training Framework for Neural Chat Translation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(7): 7970–7985.

Zhu, L.; Zhao, H.; Zhu, Z.; Zhang, C.; and Kong, X. 2025. Multimodal sentiment analysis with unimodal label generation and modality decomposition. *Inf. Fusion*, 116: 102787.

Zhuang, Y.; Zhang, Y.; Hu, Z.; Zhang, X.; Deng, J.; and Ren, F. 2024. GLoMo: Global-Local Modal Fusion for Multimodal Sentiment Analysis. In *MM*, 1800–1809.

Zong, D.; Ding, C.; Li, B.; Li, J.; Zheng, K.; and Zhou, Q. 2023. AcFormer: An Aligned and Compact Transformer for Multimodal Sentiment Analysis. In *MM*, 833–842.