IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 April 2024, accepted 17 May 2024, date of publication 27 May 2024, date of current version 3 June 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3405586*

RESEARCH ARTICLE

# BCD-MM: Multimodal Sentiment Analysis Model With Dual-Bias-Aware Feature Learning and Attention Mechanisms

**LEI MA**[1], **JINGTAO LI**[1], **DANGGUO SHAO**[1,2], **(Member, IEEE), JIANGKAI YAN**[1], **JIAWEI WANG**[1], **AND YUKUN YAN**[1]

[1]Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China
[2]Key Laboratory of Artificial Intelligence in Yunnan Province, Kunming University of Science and Technology, Kunming 650500, China

Corresponding author: Dangguo Shao (cancerjohn@126.com)

This work was supported by the National Natural Science Foundation of China under Grant 62266025.

**ABSTRACT** Multimodal Sentiment Analysis (MSA) is gaining attention, but faces two main challenges: efficient extraction of cross-modal features without redundancy and removing spurious correlations between sentiment labels and multimodal features. In this paper, we propose a novel multimodal learning debiasing model, named Bilateral Cross-modal Debias Multimodal sentiment analysis Model (BCD-MM), to address these issues. Specifically, BCD-MM ultimately enhances the generalisation of the model to out-of-distribution (OOD) situations by improving the ability of cross-modal low-redundancy feature extraction and reducing the reliance on non-causal correlations. First, BCD-MM utilizes an attention score-based method to preserve critical information and eliminate redundancy within modalities. It also employs a gated crossmodal attention mechanism to filter inconsistencies through modal interaction, thereby enhancing the extraction of cross-modal specific features. Second, BCD-MM incorporates a debiasing approach with double bias extraction, using a Tanh-based Mean Absolute Error (TMAE) loss function and inverse probability weighting to mitigate spurious correlations. Finally, extensive testing on three public datasets (MOSI, MOSEI, and SIMS) and two OOD datasets (OOD MOSI and OOD MOSEI) demonstrates our model's effectiveness in both MSA and debiasing tasks.

**INDEX TERMS** Attention mechanisms, debiased learning, deep learning, multimodal sentiment analysis, out-of-distribution generalization.

## I. INTRODUCTION

The rise of platforms like YouTube and Instagram has led to more expressive multimodal communication, including voice, text, emojis, and body language. Traditional sentiment analysis [1], [2], [3] usually relies on only one modality in text, speech, or facial expression, leading to limitations such as incomplete information. Consequently, this inability to accurately assess emotional tone, such as sarcasm and humor, significantly limits the accuracy of the analysis. Multimodal Sentiment Analysis (MSA) combines these different modalities to provide a more comprehensive understanding of sentiment [4], which is important for market trend analysis, personalized marketing and brand management [5], [6], [7], [8]. In this study, we focus on addressing two major challenges in the field of multimodal sentiment analysis (MSA). First, the problem of inter-modality inconsistency, for instance, how to extract and integrate key features efficiently from different modalities (e.g. text, sound, and video), while excluding those factors that may lead to inconsistent or confusing information. Second, we address the problem of non-causal links between sentiment labels and multimodal features. This spurious correlation may not only lead to a degradation of the model's generalization ability when dealing with real-world data but also lead to a significant performance degradation when confronted with out-of-distribution (OOD) data. At the heart of these two challenges is ensuring that the model not only

The associate editor coordinating the review of this manuscript and approving it for publication was Maria Chiara Caschera.

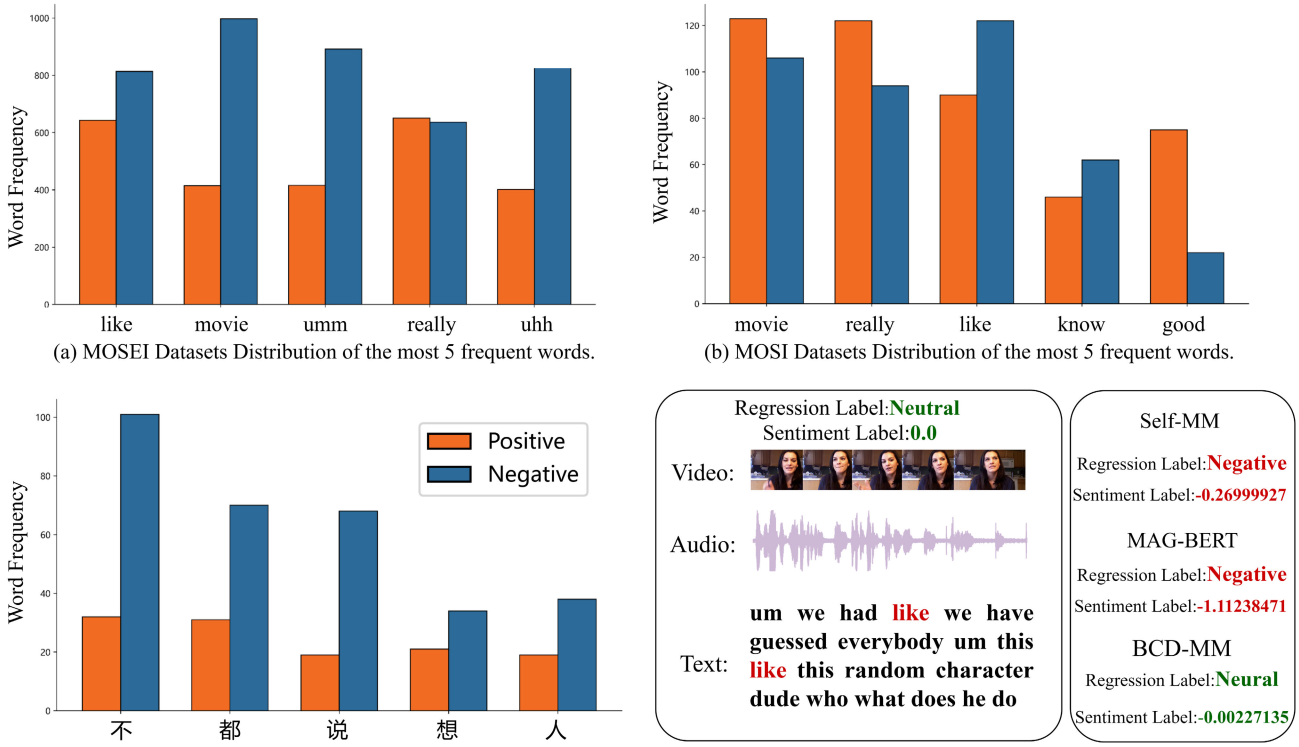
extracts valuable information from multiple modalities but also effectively distinguishes and excludes misleading or irrelevant signals, thus improving the accuracy and robustness of multimodal sentiment analysis.

The first challenge is among the current research areas in which less attention has been given to multimodal feature representation learning. Previous studies [6], [7], [8], [14], [15] tend to extract features within each modality independently during the feature extraction process and only consider inter-modality information sharing in the subsequent feature fusion phase. However, this separate extraction may lead to multiple inconsistencies, and the presence of semantic conflicts may trigger inconsistencies in feature expression. For example, in verbal sarcasm, there may be differences between verbal and auditory emotional expressions. Moreover, external factors such as noise or absence in the data can also lead to inter-modality inconsistencies. Therefore, it is a major challenge to extract reliable features that are complementary across modalities and remove nonessential ambiguities to prevent such inconsistencies.

The second challenge lies in the problem of prone to incorrect correlation matching between multimodal features and sentiment labels. This can lead to a decrease in the predictive accuracy of the model, especially when confronted with out-of-distribution (OOD) data [16]. False correlation occurs when a model incorrectly learns an association between a specific feature (e.g. certain words in a text, background color of a video) and a sentiment label, even if this association does not always hold in reality [17]. For example, as shown in the analysis of Fig.1(a), Fig.1(b), and Fig.1(c), some words in the text modality seem to be closely related to sentiment labels, but this may be unreliable. Also the video and audio modals suffer from the same problem [18]. For example, we see a smiling woman with a brown background, which is a surface feature that can be easily captured by the model, while words such as ''like'' appear more frequently in the positive sample of the training set than in the negative sample of the training set. Thus, 'like' and a brown background create biased, misleading links with emotion labels, which standard MSA models fail to address effectively. Self-MM [12] and MAG-BERT [13] models classify neutral test samples with a brown background and ''like'' as negative samples. In fact, a brown background and ''like'' are not reliable cues for identifying negative samples. Due to short-cut bias [19], deep neural networks can easily adapt to this correlation to make predictions. Consequently, this diminishes the ability to generalize in out-of-distribution (OOD) situations. Traditional studies have used an identical pair of robust and bias extractors for three single modalities [18], thus ignoring the importance of inter-modality complementarity for bias extraction. Therefore, improving the out-of-distribution (OOD) generalization of MSA models by minimizing multimodal pseudo-correlations presents a significant challenge.

To address the two main challenges in MSA, we propose an integrated innovative analysis model, the Bilateral Cross-Modal Debias Multimodal Sentiment Analysis model (BCD-MM), featuring distinct modules for specific challenges. First, we introduce the Top Attention Extractor (TAE) module and the Bimodal Cross Attention Gate Interaction (BCAG) module to address inter-modality inconsistency. The TAE module efficiently filters key information and eliminates invalid ambiguities based on attention scores. The BCAG module, which we have improved based on the work of Sun et al. [20]. It is based on the Transformer architecture and adopts a unique gating mechanism for effective inter-modality interaction, and its parallel structure is designed to simultaneously extract inter-modality complementary features and filter out invalid information, thus enhancing inter-modality coherence. Second, to cope with the problem of miscorrelation matching, we design the Dual Unbiased Extract Robust Debiasing (DUERD) module, which employs an inverse probability weighting (IPW) method to assign fewer training weights to biased samples. Then, the strategy employs a robust feature extractor and integrates it with a dual bias extraction approach, which combines a traditional bias feature extractor with our newly proposed cross-modal bias extractor. This method aims to effectively remove biases from multimodal sentiment analysis models. Finally, inspired by Sun et al. [18], we adopted a tanh-based mean absolute error (TMAE) loss function, aiding in distinguishing significant from minor errors and enhancing model performance, particularly in addressing spurious correlations. Furthermore, to counteract the impact of learned false correlations, especially in out-of-distribution (OOD) scenarios, we extensively tested our model using the OOD MOSI and OOD MOSEI datasets from Sun et al. [16], confirming its strong debiasing effectiveness. In addition, we verify that our model is still competitive on the IID dataset. To summarize, this research has made four main contributions.

- We introduce an innovative module to minimize superfluous details within modalities, maintaining crucial features and boosting the model's accuracy in managing complex multimodal data.
- We employ a Transformer-based parallel structure enhancing inter-modality interactions and model's capacity to learn integrated sentiment. This mechanism effectively filters out inconsistent subsequences by adaptively regulating inter-modality interactions.
- We crafted a new debiasing model to a precisely identify unusual features and bolster model robustness. We designed two sets of bias extractors and one robust extractor for each modality, exploited the intra-modality diversity and inter-modality complementary information, and combined the TMAE loss function and IPW-enhanced training method to effectively boost the model's generalization capabilities.
- We perform comprehensive tests on three IID datasets (MOSI [9], MOSEI [10], and SIMS [11]) and two OOD datasets [16](OOD MOSI, OOD MOSEI). The results



(a) MOSEI Datasets Distribution of the most 5 frequent words.

(b) MOSI Datasets Distribution of the most 5 frequent words.

(c) SIMS Datasets Distribution of the most 5 frequent words.

(d) Examples of test sets in MOSI

**FIGURE 1. MOSEI [9], MOSI [10], and SIMS [11] training samples, the distributions of the top 5 most common words in MOSEI, MOSI, and SIMS, as well as the predictions of Self-MM [12], the MAG-BERT [13] model, and ours for the test samples.**

highlight the exceptional generalizability and debiasing capabilities of BCD-MM.

The rest of this paper is organized as follows. Section II reviews the related work, and Section III presents a detailed description of the proposed BCD-MM method. Section IV provides the experimental details. Section V provides the results and analysis. Finally, section VI provides the conclusion.

## II. RELATED WORK

This study aims to categorize sentiment propensities across text, audio, and video modalities. It centers on performing inter-modality interactions and learning to counter biased correlations between multimodal features and sentiment labels. In this section, we delve into multimodal sentiment analysis, Transformer-based multimodal interaction, and debiased learning. In addition, we focus on the innovation of our work.

### A. MULTIMODAL SENTIMENT ANALYSIS

MSA extends traditional text-based sentiment analysis to speech and visual features, aiming to determine the overall sentiment in a discourse. Research in MSA has mainly focused on representation learning and multimodal fusion. In representation learning, approaches include: 1) Using multivariate Gaussian distributions with KL divergence for temporal distribution similarity [20]. 2) Developing shared subspace learning models that map modalities to both modality-invariant and modality-specific representations [21]. 3) Employing self-supervised models for single-peak labels and multitask training [12]. Regarding multimodal fusion, researchers have applied two strategies based on the fusion stage: 1) Early fusion. Zadeh et al. [8] designed a memory fusion network for cross-view interactions. Tsai et al. [15] proposed a cross-modal Transformer that enhances the target modality through cross-modal attention. 2) Late fusion. Zadeh et al. [6] applied a tensor fusion network by computing the outer product between single-peak representations to obtain a tensor representation. Liu et al. [7] proposed a low-rank multimodal fusion method designed to lower the computational complexity of tensor-based approaches.

Despite the great success of existing studies, they might ignore inter-modality inconsistencies and spurious correlations between sentiment labels. Our study introduces an innovative module to promote model consistency, mitigate intra-modality redundancy, and retain critical information. Specifically, focus is placed on debiased learning to address the common problem of spurious correlations in sentiment analysis. In summary, we propose a robust debiasing model BCD-MM based on double bias extraction with cross-modal attention mechanisms.

### B. TRANSFORMER-BASED MULTIMODAL INTERACTION

The widespread adoption of Transformer networks [22] in natural language processing and computer vision has led researchers to utilize specific self-attention mechanisms for exploring correlations between different modalities [23], [24]. This Transformer-based self-attention mechanism, particularly when employing queries ($Q$) and keys ($K$), is adept at processing multimodal data and capturing extensive global information. In multimodal scenarios, most Transformer-based methods [24] employ dual modalities ($Q$ and $K$) to create a shared attention graph for exchanging information, a technique termed intermodal attention. For instance, Han et al. [25] implemented a symmetric inter-modality attention structure in their work to equilibrate information across various modalities. However, traditional Transformer networks have certain limitations in dealing with data containing multiple modalities, especially in calculating the inter-modality correlations between different modalities. For example, the cross-attention mechanism using $Q$ and $K$ modalities can usually only compute the correlation between two modalities, which is insufficient when dealing with video data containing multiple modalities. As a result, many studies consider text to be the primary modality [23], [24], [25] and employ other modalities to augment linguistic information.

To address these challenges, our research presents a novel three-branch parallel model to enhance inter-modal data transfer. Each branch enhances an inter-modality attention module for deeper multimodal relationship analysis. This approach overcome the limitations of traditional models in multimodal processing and enables more in-depth information exchange and analysis.

### C. DEBIASED LEARNING

Debiasing learning methods are categorized into three main approaches [26], each targeting different stages of the model development process. 1) Preprocessing debiasing [27], [28] focuses on correcting bias and imbalance before model training through data distribution adjustments or transformations. 2) Debiasing during training [29], [30], [31] address biases during the model training phase by incorporating fairness metrics into the optimization process to balance accuracy and fairness. 3) Postprocessing debiasing [32], [33] involves applying adjustments after training to strengthen the fairness of model predictions for protected variables and their subgroups.

Despite the great success of existing research, however, multimodal data still have complex biases that are difficult to identify. For this reason, the second class is resorted to for performing debiasing during training. By customizing novel dual bias extractors and robust extractors for each modality, combined with the proposed TMAE loss function and IPW-enhanced training method, our model obviously develops the generalizability to OOD data, further validating the debiasing ability of the model.

## III. METHODOLOGY

In this study, we treat multimodal sentiment analysis (MSA) as a regression task, described in Fig.2, comprising three main components: 1) Multimodal Feature Representation Module: This module utilizes a robust feature extractor $E_T^R(m)$ along with double bias extractors (traditional $E_T^B(m)$ and a new cross-modal $E_C^B(m)$) for each of the three modalities, aiming to enhance both the robust and biased features.Therefore, this section is divided into two subsections: the Traditional Biased/Robust Feature Encoder and the Cross-modal bias extractor. 2) Fusion Module: The module fuses cross-modal biased features with robust features through simple splicing. 3) Debiased Optimization Module: Utilizes TMAE loss for bias extractor training and employs an Inverse Probability Weighting (IPW) augmented Mean Absolute Error (MAE) loss for the robust extractor. This approach calculates the absolute differences between multimodal bias feature predictions and sentiment labels, using these values to estimate bias weights for each sample. IPW is applied to adjust sample weights according to their bias, minimizing the impact of biased samples and fostering debiased learning. Each module is elaborated further in the sections that follow.

### A. TRADITIONAL BIASED/ROBUST FEATURE ENCODER

#### 1) TEXT ROBUST BIAS EXTRACTOR

Our approach adopts a traditional multimodal feature extraction method. For the text module, a pretrained 12-layer BERT model is utilized to derive sentence representations. Consistent with previous studies [12], the first word vector from the last layer is selected to represent the entire sentence. Subsequently, a linear layer maps these features into a lower-dimensional semantic space. Then, the original text 't' is processed by the robust text extractor and biased extractor to obtain the robust potential text vector $F_t^R$ and the biased potential text vector $F_t^B$. The comprehensive structure of these text extractors is outlined as follows:

$$F_t^\gamma = E_T^\gamma(t) = ReLU(W_t^\gamma \left(BERT_t^\gamma(t)\right) + \mathbf{b}_t^\gamma), \tag{1}$$

where RuLU is the ReLU activation function [34], $\gamma \in \{R, B\}$, $F_t^\gamma \in \mathbb{R}^{d_{st}}$, $\mathbf{W}_t^\gamma \in \mathbb{R}^{d_{st} \times d_t}$, $\mathbf{b}_t^\gamma \in \mathbb{R}^{d_{st}}$, $d_{st}$, and $d_t$ denote the text potential vector and the output of BERT, respectively.

#### 2) AUDIO AND VIDEO MODAL ROBUST BIAS EXTRACTOR

In the video and audio modalities, we adopt the method of Yu et al. [12], with 'a' and 'v' indicating the extracted audio and video features, respective. Unidirectional long short-term memory (sLSTM) is manipulated to capture temporal features, following previous studies [6], [12], and the final state vector of sLSTM is chosen as the entire modality representation as follows:

$$F_a^\gamma = E_T^\gamma(a) = ReLU(W_a^\gamma \left(sLSTM_a^\gamma(a)\right) + b_a^\gamma), \tag{2}$$
$$F_v^\gamma = E_T^\gamma(v) = ReLU(W_v^\gamma \left(sLSTM_v^\gamma(v)\right) + b_v^\gamma), \tag{3}$$

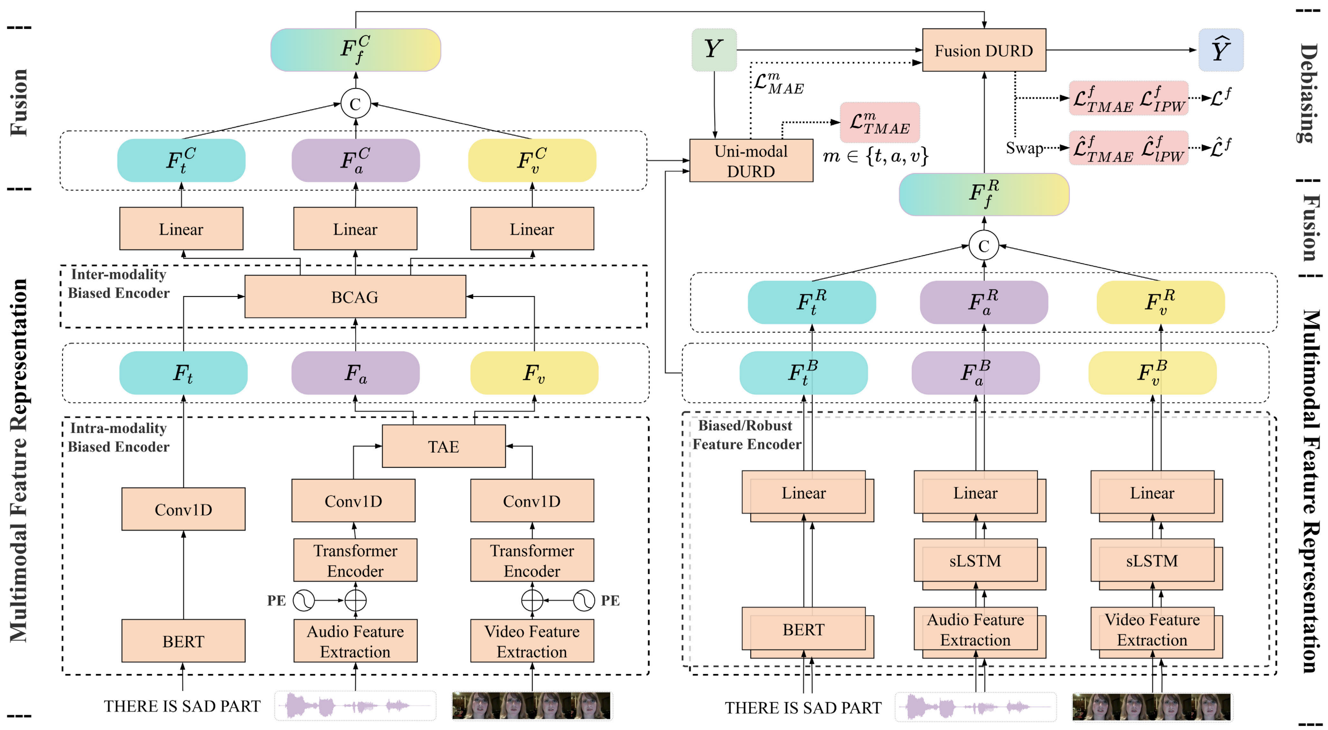


**FIGURE 2.** **The overall structure of our proposed model. It includes multimodal feature representation, fusion and debiased optimization. $F_t^C$, $F_a^C$, $F_v^C$ are the cross-modal biased feature output by the Inter-modality Biased Encoder. $F_t^B$, $F_a^B$, and $F_v^B$ are traditional bias features that have been output by Biased Feature Encoder, $F_t^R$, $F_a^R$, and $F_v^R$ are robust features that have been output by Robust Feature Encoder, $F_f^C$, $F_f^R$, are multimodal features fused by a simple splicing of cross-modal bias features and robust features, respectively. ⊕ represents Element wise addition, PE stands for Positional Encoding. Finally, $Y$ is the human multimodal annotation, $\widehat{Y}$ is the predicted emotional output.**

where $\gamma \in \{R, B\}$, $F_a^\gamma \in \mathbb{R}^{d_{sa}}$, $F_v^\gamma \in \mathbb{R}^{d_{sv}}$, $\mathbf{W}_a^\gamma \in \mathbb{R}^{d_{sa}\times d_a}$, $\mathbf{W}_v^\gamma \in \mathbb{R}^{d_{sv}\times d_v}$, $\mathbf{b}_a^\gamma \in \mathbb{R}^{d_{sa}}$, $\mathbf{b}_v^\gamma \in \mathbb{R}^{d_{sv}}$, $d_{sa}$, $d_a$ and $d_{sv}$, $d_v$ denote the potential vector and output dimensions of the audio and video passing through the sLSTM, respectively.

## B. CROSS-MODAL BIAS EXTRACTOR

### 1) INTRA-MODALITY BIASED ENCODER

In the text module, a traditional robust bias extraction method is used to select the first word vector from the last layer of the BERT output to represent the entire sentence. To ensure that each element in the input sequence adequately captures its neighboring elements, convolution is very effective in analyzing the relationships between sequentially adjacent feature components and integrating global information. Therefore, the sequence is passed through a one-dimensional temporal convolutional layer as follows:

$$F_t = ReLU\left(BN\left(\text{Conv1D}((BERT(t), k_t)\right)\right), \tag{4}$$

where $F_t \in \mathbb{R}^{d_{st}\times d_t}$, BN is batch normalized to make predictions more stable [35], Conv1D is the temporal convolution and $k_t$ is the size of the convolution kernel for the text modality.

In the audio and video modalities, previous studies have suggested that conventional LSTM may lose key features in extended sequences [36], thus compromising its effectiveness in specific applications. Especially in extracting bias, part of the lost information may be bias information, leading to a degradation of the debiasing performance. Therefore, positional embedding [22] is first introduced to augment the sequences, and intra-modality Transformers are introduced to model the intra-modality interactions of audio and video sequences. Subsequently, we employ the Transformer encoder, which uses the features that have undergone positional embedding as the query, key, and value for the extracted features $F_m^T \in \mathbb{R}^{d_{sm}\times d_m}$. Finally, the inter-modality features are extracted from the various encoders using the temporal convolution operation algorithm as follows:

$$F_m^* = E_C^B(m) = ReLU\left(BN\left(\text{Conv1D}(F_m^T, k_m)\right)\right), \tag{5}$$

where, $m \in \{a, v\}$, $F_m^* \in \mathbb{R}^{d_{sm}\times d_m}$, and $k_m$ is the size of the convolution kernel of audio and video modality.

### 2) TOP ATTENTION EXTRACTOR (TAE) MODULE

To retain the key information while removing the redundant information within the modality, and at the same time provide dimensionally consistent features for the subsequent cross-modal attention mechanism to ensure seamless integration and interaction between the modalities, the features from both modalities are mapped into a new feature space, so we design the TAE (Top Attention Extractor) module, as shown in Algorithm 1.The algorithm is described in detail as shown below.

**Algorithm 1** Top Attention Extractor

```
Input: m-modal redundant features $\boldsymbol{F_m^*}$, Top dimensions $\boldsymbol{K}$
Output: Key features of Top $\boldsymbol{K}$ dimensions $\boldsymbol{F_m}$
for $F_m$ in $\{a, v\}$ do
    attn_scores$_m$ ← Calculate $\boldsymbol{F_m^*}$ using Eqn.(6)
    attn_weights$_m$ ← Allocation based on $\boldsymbol{attn\_scores_m}$ using Eqn.(7)
    for $i = 1$ to $\boldsymbol{K}$ do
        $\boldsymbol{i_K}$ ← top $\boldsymbol{K}$ indices from $\boldsymbol{attn\_weight_m}$
        $\boldsymbol{v_K}$ ← $\boldsymbol{i_K}$ value corresponding to $\boldsymbol{attn\_weight_m}$
    end for
    $\boldsymbol{v_K}$ ← Quicksort($\boldsymbol{v_K}$)
    $\boldsymbol{i_K}$ ← Update based on the position of $\boldsymbol{v_K}$
    $\boldsymbol{F_m}$ ← Elements in $\boldsymbol{F_m^*}$ indexed by $\boldsymbol{i_K}$
end for
```

First, the $F_m^*$ linear projection of m modes, containing redundant information, is mapped into a tensor as follows:

$$attn\_scores_m = squeeze(W_m F_m^* + b_m), \tag{6}$$

where, $F_m^* \in \mathbb{R}^{d_{sm} \times d_m}$, $W_m \in \mathbb{R}^{1 \times d_{sm}}$, $b_m \in \mathbb{R}^{d_m}$, $attn\_scores_m \in \mathbb{R}^{d_m}$, squeeze() denote the squeezing operation, i.e. removing dimensions of size 1 from the tensor, which is as follows:

$$attn\_weigths_m = Softmax(attn\_scores_m), \tag{7}$$

where, $attn\_weigths_m \in \mathbb{R}^{d_m}$, and Softmax is the Softmax activation function. Then, based on the calculated attention weights $attn\_weights_m$, we select the top $\boldsymbol{K}$ important dimensions for indexing and value correspondence. After that, we apply the quicksort algorithm to sort and update the indices of these top $\boldsymbol{K}$ important dimensions, thereby extracting the key features of the m modality $F_m$.

Finally, after the TAE module, the a, v features contain the key information $F_m \in \mathbb{R}^{d_k \times d_m}$. Here, $d_k$ is a preserved K-dimensional feature. To simplify the model's complexity and streamline the operations of subsequent modules, $d_k$ was chosen with the same dimensions as the textual latent vector as $d_{st}$.

### 3) BIMODAL CROSS ATTENTION GATE INTERACTION (BCAG) MODULE

To effectively extract useful inter-modality complementary features and eliminate invalid ambiguities, thus preventing inter-modality inconsistencies, the BCAG module was developed, as depicted in Fig.3 built upon the Transformer Encoder framework [22], BCAG features two parallel inter-modalty attention streams, and our gating mechanism is inspired by Sun et al. [20] and decomposing bilinear pools [37] (FBP), which is designed to generate temporal gating signals.

Taking the $F_{m1_{m2}}$ side of the BCAG module as an example, the query $Q_{m1}$ of the $m_1$ modality feature $F_{m1}$, the key $K_{m2}$ and the value $V_{m2}$ of the $m_2$ modality feature $F_{m2}$ are mapped to different tensors using separate linear projections, in addition to $Q_{m1}$, $K_{m2}$ to generate time-gated signals as follows:

$$\begin{aligned} Q_{m_1} &= F_{m_1} W_Q, \\ K_{m_2} &= F_{m_2} W_K, \\ V_{m_2} &= F_{m_2} W_V, \\ Q'_{m_1} &= Q_{m_1} W'_Q, \\ K'_{m_2} &= K_{m_2} W'_K, \end{aligned} \tag{8}$$

where, $F_{m_1} \in \mathbb{R}^{d_k \times d_{m_1}}$, $F_{m_2} \in \mathbb{R}^{d_k \times d_{m_2}}$, $Q_{m_1} \in \mathbb{R}^{d_k \times d_{km_1}}$,$K_{m_2} \in \mathbb{R}^{d_k \times d_{km_2}}$, $V_{m_2} \in \mathbb{R}^{d_k \times d_{vm_2}}$, $Q'_{m_1} \in \mathbb{R}^{d_k \times d_g}$, $K'_{m_2} \in \mathbb{R}^{d_k \times d_g}$,$W_Q \in \mathbb{R}^{d_{m_1} \times d_{km_1}}$, $W_K \in \mathbb{R}^{d_{m_2} \times d_{km_2}}$, $W_V \in \mathbb{R}^{d_{m_2} \times d_{vm_2}}$,$W_{Q'} \in \mathbb{R}^{d_{km_1} \times d_g}$, $W'_K \in \mathbb{R}^{d_{km_2} \times d_g}$, $d_{km_1}$, $d_{km_2}$, and $d_{vm_2}$ are the hidden dimensions of the modality $m1, m2 \in \{t, a, v\}$, and $d_g$ is the hidden dimension of the FBP calculation. Then, the stream aiming to augment modality $m_1$ by modality $m_2$ inputs $Q_{m1}$, $K_{m2}$, $V_{m2}$ to Cross-Modal Attention (CMA) and performs the cross-attention operation to produce the interaction feature $A_{m_2 \to m_1}$ as follows:

$$\begin{aligned} A_{m_2 \to m_1} &= CMA(Q_{m_1}, K_{m_2}, V_{m_2}) \\ &= softmax(\frac{Q_{m_1} K_{m_2}^\top}{\sqrt{d_{km_1}}}) V_{m_2} \\ &= softmax(\frac{F_{m_1} W_Q W_K^\top F_{m_2}^\top}{\sqrt{d_{km_1}}}) F_{m_2 W_V} \end{aligned} \tag{9}$$

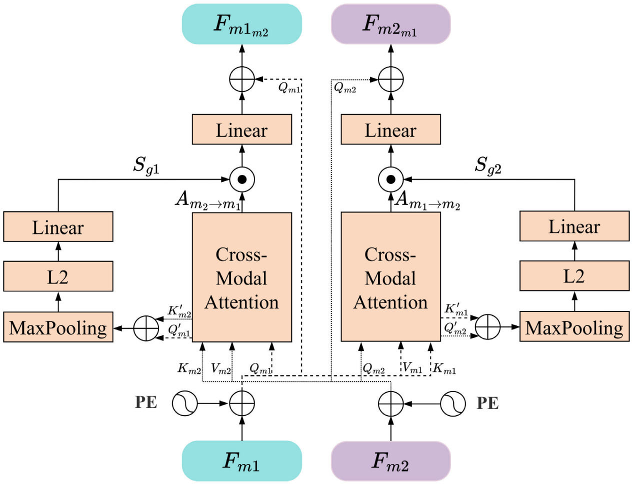


**FIGURE 3. Overview structure of the bimodal cross attention gate interaction (BCAG).**

Moreover, the gating mechanism utilizes $Q'_{m_1}$ and $K'_{m_2}$ as inputs to generate gating signals to control the interaction. The formula can be expressed as:

$$\begin{aligned} F^{mp}_{m_1,m_2} &= MaxPool(Q'_{m_1} \oplus K'_{m_2}, p) \\ F^{norm}_{m_1,m_2} &= \frac{F^{mp}_{m_1,m_2}}{\parallel F^{mp}_{m_1,m_2} \parallel_2}, \end{aligned}$$

$$S_{g1} = F^{norm}_{m_1,m_2} W_{norm} + b_{norm}, \tag{10}$$

where, $F^{mp}_{m_1,m_2}, F^{norm}_{m1,m2} \in \mathbb{R}^{d_k \times \frac{2d_g}{p}}$, $S_{g1} \in \mathbb{R}^{d_k}$, $W_{norm} \in \mathbb{R}^{\frac{2d_g}{p} \times 1}$, $b_{norm} \in \mathbb{R}^{d_k}$, $\oplus$ represents element-by-element summation, and $MaxPool(, p)$ denotes maximal pooling with a pooling window of size p. In summary, the gating mechanism we construct is able to adaptively determine the modality strength of the association between $m_1$ and $m_2$ while filtering out irrelevant timestamps. Filtering is performed at the time level through gating signals, enabling finer, more precise inter-modality interactions as follows:

$$F_{m1_{m2}} = (Sign(S_{g1}) A_{m_2 \to m_1}) W_A \oplus Q_{m1}, \tag{11}$$

where, $Sign$ is the sign function that produces 0 or 1. Therefore, the modality $m_2$ augmented modality $m_1$ is generated, using the same process in the other stream to generate $F_{m2_{m1}}$.

After three pairs of BCAG modules, there are a total of six outputs $(F_{t_a}, F_{a_t})$, $(F_{a_v}, F_{v_a})$, and $(F_{v_t}, F_{t_v})$. To compute the enhanced multimodal representation, drawing inspiration from Sun et al. [20], the mean value of each modality's outputs is calculated to derive its representative description as follows:

$$\begin{aligned}
F^C_t &= E^B_C(t) = \frac{1}{2}(F_{t_a} + F_{t_v}), \\
F^C_a &= E^B_C(a) = \frac{1}{2}(F_{a_t} + F_{a_v}), \\
F^C_v &= E^B_C(v) = \frac{1}{2}(F_{v_a} + F_{v_t}),
\end{aligned} \tag{12}$$

where, by means of the $E^B_C(m)$, features that are complementary between modalities are extracted while filtering out invalid ambiguities to avoid inconsistent modality features $F^C_m, m \in \{t, a, v\}$, they perceive all the information involving the modality.

To make the cross-modal feature $F^C_m$ consistent with the conventional robust, biased feature (Eqs.(1)(2)(3)), averaging is applied to compress the first-dimensional dimension as follows:

$$F^C_m = \frac{1}{N} \sum_{i=1}^{N} F^C_m[i, :], \tag{13}$$

where, $F^C_m \in \mathbb{R}^{d_k}$, $d_k$ denotes the dimension of the potential vector.

### C. FUSION MODULE

To reduce the complexity of the model, a complex fusion network is employed, generating multimodal embeddings from the $E^R_T(m)$ and $E^B_C(m)$ outputs, which may introduce redundancy and potentially ignore discriminative unimodal information [38]. Therefore, we chose the same fusion approach as in previous studies [12], [15], [38] using simple concatenation to fuse the multimodal features in the features extracted by the robust extractor $E^R_T(m)$ as follows:

$$F^R_f = E^R_T(f) = [F^R_t; F^R_a; F^R_v], \tag{14}$$

where, $F^R_f \in \mathbb{R}^{d_f}$, $d_f = d_{st} + d_{sa} + d_{sv}$. In the features of the output of $E^B_C(m)$, The $F^C_m$ features obtained after compressing the first dimension using the mean are concatenated to obtain the fused cross-modal deviation feature as follow:

$$F^C_f = E^B_C(f) = W^C_f [F^C_t; F^C_a; F^C_v], \tag{15}$$

where, $F^c_f \in \mathbb{R}^{d_f}$, $W^C_f \in \mathbb{R}^{d_f \times 3d_k}$.

### D. DEBIASED OPTIMIZATION MODULE

To perform the debiasing operation accurately, we need to extract high-quality biased features. The TMAE loss, which is a loss function for debiasing optimization that amplifies the model's learning of biased features and advances sensitivity to outliers, is proposed. In the multimodal task, the Uni-modal Dual Unbiased Extraction Robust Removal Module (Uni-modal DUERD) and Fusion Dual Unbiased Extraction Robust Removal Module (Fusion DUERD) are used to further enhance the debiasing effect, and the TMAE loss is applied to the single-modality and fusion features, respectively, to achieve effective learning of bias and debiasing optimization. High-quality bias features are employed to assess the bias in each modality and calculate the bias weights for the samples.

#### 1) TMAE LOSS

It is well known that in the early stages of training, models tend to be more inclined to learn features that are biased and simple rather than more complex and robust features [39]. By using Generalized Cross Entropy (GCE) loss, researchers can more effectively construct models that focus on these biased features to train biased models. However, GCE is considered more applicable to classification tasks, while MSA is viewed as a regression task. Inspired by [16], [18], and [40], the mean absolute error (MAE) is less sensitive to outliers. This is because it does not square the error as MSE does, so the error growth is not as sharp for large errors. This makes the MAE particularly useful in datasets where outliers are present. However, using only the MAE to give the same weight to all sizes of errors means that large and small errors are not distinguished in terms of their impact on model performance. Since the tanh function has a larger slope as the input approaches zero, adding tanh to the loss function can increase the sensitivity of the model to smaller errors. Hence, the TMAE loss is applied to encourage biased extractors to acquire high-quality biased features and amplify the bias. The TMAE loss is defined as follows:

$$\mathcal{L}^m_{TMAE} = \tanh\left(\left(\frac{1}{|Y_i - \hat{y}_i|}\right)^q\right) |Y_i - \hat{y}_i|, \tag{16}$$

where || indicates the absolute value, $Y_i$ denotes the real label, and $\hat{y}_i$ denotes the sentiment prediction in $m \in \{t, a, v\}$ for three single modalities.

#### 2) UNI-MODAL DUERD

The traditional bias extractor $E^B_T(m)$ and cross-modal bias extractor $E^B_C(m)$ are trained to amplify the "bias" through

TMAE loss so that the biased model excels at making predictions using biased features. To achieve effective debiasing, a Uni-modal DUERD is introduced. This module employs a dual bias extractor to better capture potential biases present in different modalities (text, audio, video), thereby facilitating within-modality debiasing learning.

The $E_T^B(m)$ output ($F_t^B, F_a^B, F_v^B$) from Eqs.(1)(2)(3) and the $E_C^B(m)$ output ($F_t^C, F_a^C, F_v^C$) from Eqn.(13) are combined into a modality pair: ($F_t^C$, $F_t^B$), ($F_a^C$, $F_a^B$), and ($F_v^C$, $F_v^B$), which is then inputted into the Uni-modal DUERD. The designed Uni-modal DUERD is specified as follows.

In our model, a multimodal loss function $\mathcal{L}_{MAE_m}^{\beta}$ is employed to simultaneously account for the consistency of textual, audio, and visual modality bias information. The performance for each modality result is evaluated by the corresponding MAE loss as follows:

$$\mathcal{L}_{MAE_m}^{\beta} = |Y_i - F_m^{\beta}|, \tag{17}$$

where, $\beta \in \{C, B\}$, $Y_i$ denotes the real label, and $F_m^{\beta} \in \mathbb{R}^{d_{sm}}$. These loss values reflect the accuracy of the model's label predictions for each individual modality.

The average errors from the robust and deviation extractors are summed, which can be expressed as follows:

$$\mathcal{L}_{MAE}^{m} = \mathcal{L}_{MAE_m}^{C} + \mathcal{L}_{MAE_m}^{B}, \tag{18}$$

Subsequently, to amplify the bias, the modality pairs are also input into the $\mathcal{L}_{TMAE_m}^{\beta}$ loss function separately as follows:

$$\mathcal{L}_{TMAE_m}^{\beta} = \tan\mathrm{h}\left(\left(\frac{1}{|\ Y_i - F_m^{\beta}\ |}\right)^q\right) |\ Y_i - F_m^{\beta}\ |, \tag{19}$$

Finally, the final loss function for this module can be expressed as follows:

$$\mathcal{L}_{TMAE}^{m} = \lambda_1^m \mathcal{L}_{TMAE_m}^{C} + \lambda_2^m \mathcal{L}_{TMAE_m}^{B}, \tag{20}$$

where,$\lambda_1^m$, $\lambda_2^m$ were adjusted to weight the TMAE to balance the difference between the conventional bias extractor and the cross-modal attention bias extractor.

### 3) FUSION DUERD

Given the significant advantages of multimodal features in sentiment analysis tasks, an innovative fusion debiasing module for dual unbiased extraction, named "Fusion DUERD", is designed and implemented to process and optimize fused multimodal features. The core objective of Fusion DUERD is to enhance the generalization ability of the model to the fused features while effectively removing potential biases from these fused features. First, similar to the Uni-modal DUERD module, the fused robust and biased features $F_f^R$ and $F_f^C$ are evaluated and input to the MAE loss, respectively. The details are as follows:

$$\mathcal{L}_{MAE}^{\alpha} = |Y_i - F_f^{\alpha}|, \tag{21}$$

where $\alpha \in \{R, C\}$, $F_f^{\alpha} \in \mathbb{R}^{d_f}$

Subsequently, to amplify the bias of the fused features, the fused biased features $F_f^C$ are also input into the $\mathcal{L}_{TMAE_f}$ loss function as follows:

$$\mathcal{L}_{TMAE}^{f} = \tanh\left(\left(\frac{1}{|\ Y_i - F_f^C\ |}\right)^q\right) |\ Y_i - F_f^C\ |, \tag{22}$$

Then, in order to efficiently measure the bias of each sample, we apply the MAE strategy to each fusion pattern by taking the minimum of the outputs of $\{\mathcal{L}_{MAE}^t, \mathcal{L}_{MAE}^a, \mathcal{L}_{MAE}^v\}$ in Eqn.18 and $\mathcal{L}_{MAE}^C$ in Eqn.21 and taking the reciprocal of them to compute the bias weights. The details are as follows:

$$\psi_{min}(Y_i, F_f^R) = \frac{1}{min(\mathcal{L}_{MAE}^t, \mathcal{L}_{MAE}^a, \mathcal{L}_{MAE}^v, \mathcal{L}_{MAE}^C)}, \tag{23}$$

where $\psi_{min}()$ signifies the function for estimating a sample's bias weight, the larger the bias weight is, the more significant the sample's bias. This approach is applied to determine the modality with the highest bias, reflecting the overall bias of the sample.

To learn more robust features in the face of biased data containing spurious correlations, IPW is employed to enhance the mean square error loss. The core idea of IPW is to assign smaller weights to samples with larger biases during training so that the feature extractor focuses more on robust features of unbiased samples. This approach improves the model's ability to generalize over diverse data. Therefore, our loss function employs the MAE loss combined with inverse probability weighting. The details are as follows:

$$\mathcal{L}_{IPW}^{f} = \mathcal{L}_{\mathrm{MAE}}^{\mathrm{R}} \frac{1}{\psi_{min}\left(Y_i, F_f^R\right)}, \tag{24}$$

If sample $m \in \{t, a, v\}$ shows a greater likelihood of association with its biased features, the loss needs to be reduced to discourage reliance on such biased samples.

Consequently, the final loss function for this module is formulated as follows:

$$\mathcal{L}^f = \mathcal{L}_{IPW}^f + \lambda_f \mathcal{L}_{TMAE}^f, \tag{25}$$

### 4) DIVERSIFICATION OF SAMPLES THROUGH SWAPPING

Previous studies have argued that sample diversity is critical for robust and biased feature unraveling [18]. We follow the idea [18] that to facilitate robust and biased feature unraveling of fused features, features are exchanged by targeting the robust extractor $E_T^R(m)$ and the cross-modal bias extractor $E_C^B(m)$, which were initially solved. First, the feature outputs from $E_T^R(m)$ and $E_C^B(m)$ are fused to obtain $F_f^R$ (from Eqn.(14)) and $F_f^C$ (from Eqn.(15)), respectively. And an exchange batch $S_e$ is set up. Second, when the set $S_e$ is reached, the deviation vectors $\hat{F}_f^c$ are randomly selected to replace each of the original deviation vectors $F_f^C$ to create different combinations of samples. Subsequently, the robust features $F_f^R$ are connected with the corresponding bias potential vectors $F_f^C$, and the robust features $F_f^R$ are also connected with a randomly selected bias potential vector $F_f^C$.

Finally, both connected vectors are inputted to a pair of biased and robust linear layers. The details are as follows:

$$F_f^o = ReLU(W_k^O[F_f^R; F_f^C] + b_k^O),$$
$$\widehat{F}_f^O = ReLU(\widehat{W}_k^O[F_f^R; \widehat{F}_f^C] + \widehat{b}_k^O), \quad (26)$$

where, $W_k^O, \widehat{W}_k^O \in \mathbb{R}^{d_f \times 2d_f}, b_k^O, \widehat{b}_k^O \in \mathbb{R}^{d_f}, F_f^O, \widehat{F}_f^O \in \mathbb{R}^{d_f}$ represent the latent vectors combining both robust latent vectors and biased latent vectors, with and without swapping. By this swapping, additional latent vectors $\hat{F}_f^O$ are generated, which share the same robust potential vectors but differ in bias potential vectors compared to $F_f^O$. This method yields a broader range of samples featuring diverse combinations of robust and bias features.

Consistent with Section III-D3, the bias and robust connection vectors $F_f^O$ $\hat{F}_f^O$ obtained for the exchange are evaluated by being inputted to the MAE loss, to obtain $\mathcal{L}_{\mathrm{MAE}}^O$ and $\hat{\mathcal{L}}_{MAE}^O$, respectively. The biased connection feature $\hat{F}_f^O$ is input to the TAME loss function to obtain $\hat{\mathcal{L}}_{TMAE}^f$. Meanwhile, the MAE loss for the swapped samples was also computed, combined with inverse probability weighting, to obtain $\hat{\mathcal{L}}_{lPW}^f$. Then, the final loss function for that module can be expressed as follows:

$$\hat{\mathcal{L}}f = \hat{\mathcal{L}}_{lPW}^f + \hat{\lambda}_f \hat{\mathcal{L}}_{TMAE}^f, \quad (27)$$

### E. OPTIMIZATION OBJECTIVES

Finally, the MAE loss was augmented with the TMAE and IPW methods as the underlying optimization objective. The details are as follows:

$$\mathcal{L} = \mathcal{L}_{TMAE}^t + \mathcal{L}_{TMAE}^a + \mathcal{L}_{TMAE}^v + \mathcal{L}^f + \lambda_s \hat{\mathcal{L}}^f, \quad (28)$$

where $\lambda_s$ is a coefficient before the loss of exchange. If the batch does not reach the exchange batch $S_e$, we set $\lambda_s = 0$. If $S_e$ is reached, $\lambda_s = 1$.

## IV. EXPERIMENTS

In this section, we describe the experiments that validate the effectiveness of our approach, covering implementation, datasets, and baseline methods, to validate our approach.

### A. DATASETS

Drawing on previous research [16], we evaluate the performance of our model using three public multimodal sentiment analysis datasets that serve as independent and identically distributed (IID) datasets. They are MOSI [9], MOSEI [10], and SIMS [11]. A brief overview of these datasets is provided in Table 1. In addition, we used the out-of-distribution (OOD) test sets developed by Sun et al. [16] based on the MOSI and MOSEI datasets to evaluate the debiasing ability of BCD-MM.

#### 1) IID DATASETS

To ensure a fair comparison of our model's performance with prior research, we utilized the same IID dataset as employed in previous studies [6], [8], [12], [15], specifically relying on the publicly available dataset as follows:

- **MOSI**. CMU-MOSI was created by Zadeh et al. [9] and contains 93 vlog-tagged YouTube videos. It includes 89 speakers (41 female, 48 male) and is notable as the first dataset for multimodal sentiment analysis, providing annotations for subjectivity and sentiment intensity on a scale of -3 (strongly negative) to 3 (strongly positive).
- **MOSEI**. CMU-MOSEI was created by Zadeh et al. [10], a prominent resource in sentiment analysis and emotion recognition for online videos that encompasses over 65 hours of video from more than 1000 speakers across 250 topics. It features 23,453 sentences from 3,228 videos, each with phoneme-level transcriptions synchronized with the audio. This dataset predominantly includes product and service reviews (16.2%), debates (2.9%), and advice (2.9%), and each video is assigned a sentiment score ranging from -3 to 3.
- **SIMS**. CH-SIMS was created by Yu et al. [11] and offers a comprehensive resource for both unimodal and multimodal sentiment analysis in Chinese. It includes 2,281 finely detailed video clips extracted from 60 original videos. Each clip in this dataset was meticulously annotated by human reviewers with sentiment scores ranging from -1 for strongly negative emotions to 1 for strongly positive emotions.

**TABLE 1. Dataset statistics for the MOSI, MOSEI, and SIMS.**

| Datasets | Years | Modal | Train | Valid | Test | All |
|---|---|---|---|---|---|---|
| MOSI | 2016 | T+A+V | 1284 | 229 | 686 | 2199 |
| MOSEI | 2018 | T+A+V | 16326 | 1871 | 4659 | 22856 |
| SIMS | 2022 | T+A+V | 1368 | 456 | 457 | 2281 |

#### 2) OOD DATASETS

To verify whether the model removes spurious correlations during training, and if it is able to remove bias, then it may perform well in the presence of OOD data. Therefore, we used the provided OOD datasets of Sun et al. [16], who partitioned the MOSI, MOSEI dataset into four parts, which include IID training, IID validation, IID testing, and OOD test set. For dataset segmentation, they applied a simulated annealing algorithm for each dataset (MOSI, MOSEI). The modified simulated annealing algorithm operates until the distributional difference of words across various sentiment categories between the IID and OOD sets approximates the pre-set distributional difference $\phi_\Delta$. This process involves 800 iterations. The temperature parameter of the algorithm is initially 0.5, with a decay rate of 0.99. Consequently, the OOD test set, sourced from MOSI, comprises 12 videos. To align the OOD and IID test sets in terms of size, 12 videos are randomly chosen from the IID set as its test set, with the remainder split into two, namely, 59 videos (85%) for IID training and 10 videos (15%) for IID validation. Correspondingly, there are 1830 videos in the MOSEI dataset for IID training, 324 for IID validation, 330 for IID testing, and 330 for the OOD test set. After the above processing of the dataset, the distribution of the IID dataset is significantly different from that of the OOD dataset. Therefore, by using this dataset for experiments, the debiasing ability of the model can be further demonstrated.

### B. BASELINE

To comprehensively assess the performance of BCD-MM, we conducted a fair comparative analysis against various baseline and cutting-edge models in multimodal sentiment analysis.

- **TFN**. Tensor fusion network [6], utilizes multidimensional tensors, derived from outer products, to effectively capture interactions across unimodal, bimodal, and trimodal modalities.
- **LMF**. Low-rank multimodal fusion [24], is an adaptation of the TFN approach. It incorporates low-rank multimodal tensor fusion methods to enhance efficiency.
- **MFN**. Memory Fusion Network [8] is a network that continuously models both view-specific and cross-view interactions. It efficiently summarizes these interactions over time through the employ of multi-view gated memory system.
- **MulT**. Multimodal Transformer [15] is designed to enhance the multimodal Transformer model. It achieves this by employing directed pairwise cross-modal attention mechanisms, enabling the transformation of information from one modality to another.
- **MISA**. Modal Invariant and Specific Representation [21] is adept at learning both modality invariance and modality-specific features. It utilizes a blend of losses including distributional similarity, orthogonality, reconstruction, and task prediction to achieve this nuanced understanding.
- **MAG-BERT**. Bert Multimodal Adaptation Gate [13], which represents an enhancement of the RAVEN model on aligned data. It achieves this by implementing multimodal adaptation gates at various layers within the BERT backbone structure.
- **Self-MM**. Self-MM model introduces a label generation strategy based on a self-supervised approach [12], specifically focusing on generating single-peak labels. Furthermore, it incorporates a novel weight self-adjustment strategy to balance various task loss constraints.
- **BBFN**. This model effectively controls inter-modality correlation through a bimodal fusion network and a gating mechanism [25].
- **CubeMLP**. CubeMLP is an MLP-based modal [41] for sentiment analysis that effectively blends multimodal features.
- **DEAN**. Deep Emotional Arousal Network introduces multimodal gating blocks to simulate activation mechanisms in human emotional arousal models [42].
- **PS-Mixer**. This method uses a polarity vector (PV) and an intensity vector (SV) to gauge emotion polarity and intensity, respectively [43]. These vectors are blended to obtain a fusion vector that determines the emotional state.
- **EMT-DLFR**. This model elevates the efficiency and robustness of multimodal sentiment analysis in incomplete modal environments through two-layer feature recovery and effective multimodal interaction [13].
- **CLUE**. This model is a modality independent of specific models, facilitates multimodal sentiment analysis [16], which discerns the direct impacts of text modalities using an auxiliary textual model, and computes the indirect effects via a multimodal approach.
- **GEAR**. A model that mitigates bias in multimodal sentiment analysis, enhances the model's generalization capability by integrating an inverse probability weighting model with a conventional feature segregation strategy [18].

### C. EVALUATION TASKS AND METRIC

Building on the methodology of previous studies [12], [13], [21], we present our experimental findings in two distinct categories: classification and regression. For classification, we report the weighted F1 score (F1-Score) and binary classification accuracy (Acc2). Specifically, for the MOSI and MOSEI datasets, we calculate Acc-2 and F1-Score are computed as negative/non-negative ("/" left) [6] and negative/positive ("/" right) [13]. For regression, we report Mean Absolute Error (MAE) and Pearson correlation (Corr), where higher values (except MAE) indicate better performance.

### D. IMPLEMENTATION DETAILS

Our network is implemented using the PyTorch framework and an RTX 3090 GPU. The unaligned MOSI and MOSEI datasets were applied. In our approach, experiments were executed under two distinct settings: IID and OOD. In the IID scenario, the test set aligns with the training set in terms of distribution. Conversely, in the OOD context, the aim is for the test set's sample distribution for each word across sentiment categories to markedly differ from the training set's distribution. Therefore, to validate the debiasing ability of our model, only the IID training set, the IID validation set, and the OOD test set proposed by Sun et al. [16] are used. Similar to previous IID studies, Adam is used as the optimizer, and the grid search method is employed to select the best hyperparameters, as shown in Table 2. Specifically, the dimension K in the TAE module is retained, where the values in K"/" represent the K-dimensional features of retained text/audio/video, respectively. In the TAME loss (from Eqn.(16)), the parameter number q is chosen as its index. To amplify the bias from different bias extractors, the bias amplified by TAME for different bias extractors is multiplied by different weights $\lambda_2^m$, "/", which represent the TMAE coefficients before the text/audio/video. To simplify the parameters, the parameters in the loss function are set to $\lambda_1^m = \lambda_f = \hat{\lambda}_f = 15$. To promote the robustness of fusion features and bias feature unraveling, different exchange batches for different datasets $S_e$ are used. Finally, an early stopping strategy is adopted, which stops training if the loss does not increase/decrease for 8 consecutive epochs.

**TABLE 2. BCD-MM hyperparameters for multimodal sentiment analysis.**

| Hyper parameter | MOSI | MOSEI | SIMS |
|---|---|---|---|
| Batch size | 16 | 32 | 32 |
| Text lr | 5e-5 | 5e-5 | 5e-5 |
| Audio lr | 5e-3 | 5e-4 | 7e-3 |
| Video lr | 5e-3 | 2e-4 | 7e-3 |
| Other lr | 1e-3 | 1e-3 | 1e-3 |
| K | 50/50/50 | 50/50/50 | 39/39/39 |
| q | 0.7 | 0.7 | 0.7 |
| $\lambda_2^m$ | 18/12/14 | 18/11/15 | 18/12/14 |
| $S_e$ | 4 | 4 | 5 |

## V. RESULTS AND ANALYSIS

In this section, we showcase the experimental outcomes of our suggested approach, juxtaposed with comparisons to other leading-edge methods. Following this, we detail our ablation study, which undergoes further analysis to verify the efficacy of our proposed model.

### A. QUANTITATIVE RESULTS

Our dataset, Similar to previous studies, uses an unaligned corpus. For a fair comparison, we evaluate our model against this ‘‘unaligned’’ format. Key findings from Table 3, Table 4, and Table 5, we obtain the following observations. 1) Elevated accuracy: On the MOSI dataset, there is a 2.43% increase in negative/positive accuracy and a 0.78% increase on the MOSEI. This reveals our model's (BCD-MM) advantage over previous methods. 2) Generalization: A 2.23% improvement in Acc-5 accuracy on the SIMS dataset result in relatively strong generalizability, even for Chinese data. 3) Debiasing ability: BCD-MM shows enhanced performance on OOD tests: 1.30% better on MOSI and 1.88% better on MOSEI in negative/positive accuracy. In particular, the performance improvement is very obvious on the MOSEI dataset, where more bias information exists, as shown in Fig.1, which further highlights the debiasing ability of our model. Overall, these results demonstrate the generality, improved generalization ability, and preferable debiasing performance of our model in various data scenarios.

### B. ABLATION STUDIES

#### 1) TAE MODULE

We first conducted an ablation study on the TAE module, as shown in Table 6. There is a need to provide dimensionally consistent features for subsequent cross-modal attention mechanisms to ensure seamless integration and interaction between modalities. Therefore, we cannot ignore this module, and to verify the effectiveness of the TAE module, we compare the current mainstream dimensionality reduction methods. These mainly include the following 1) Average Pooling. 2) Max Pooling. 3) Downsampling. 4) Linear layer. 5) Convolution.

Table 6 shows the results of the ablation study. Replacing the TAE module with other dimensionality reduction methods, the results on the MOSI and OOD MOSI datasets are degraded, and the linear layer and convolutional methods have relatively better performance but are still lower than that of the TAE module. The experimental results demonstrate that our TAE module can remove redundant information while retaining critical information and obviously improve the performance on the dataset.

#### 2) BCAG MODULE

To further explore the contribution of BCD-MM in extracting effective inter-modality complementary features and filtering out invalid ambiguities to validate the effectiveness of the proposed cross-modal bias extractor model, our ablation study of our proposed method is shown in Table 6. First, we replace the feature extraction part with the traditional method. w/o-Transformer, i.e. as consistent with Section III-A1, Eqn.(1) is used for the feature extraction of text, Eqn.(2) and Eqn.(3) are used for feature extraction of audio and video, respectively. Second, we verify that BCAG solves the inconsistency problem of feature representation, w/o-BCAG. The BCAG module is directly removed, and the remainder is left unchanged.

- **W/o-Transformer's limitation**. Removing the Transformer (w/o-Transformer) significantly decreases the performance on the MOSI and OOD MOSI datasets. This indicates that conventional LSTM may lose key features in extended sequences, affecting the ability to capture long-range dependencies and crucial modality information, especially in cross-modal bias extraction.
- **Importance of intra-modality information**. The same results for w/o-Transformer across both datasets underscore the necessity of capturing key intra-modality information, particularly for cross-modal bias extraction.
- **Intra-modality encoding's effectiveness**. Excluding the inter-modality encoding module (w/o-BCAG) confirms its importance in multimodal feature extraction. Its absence, particularly impacting the OOD MOSI dataset, emphasizes the need for complementarity in bias extraction.

#### 3) DURD MODULE

To delve deeper into the debiasing capabilities of BCD-MM, Table 6 presented an ablation study of our method, analyzing it through various modifications: 1) w/o-Cmbe. The Cross-modal bias extractor (Cmbe) is removed, and to validate the enhanced bias extraction capability of our dual bias extractor, we use the traditional bias extractor instead. 2) w/o-Tbe. The Traditional bias extractor (Tbe) is removed, and to validate the enhanced bias extraction of our dual bias extractor, we use the traditional bias extractor instead. 3) w/o-DbeC. The Cross-modal extractor with the Dual bias extractor (DbeC) is removed, retaining only a standard robust

**TABLE 3.** Performance of models using BERT for text encoding on the CMU-MOSI and CMU-MOSEI datasets. $^1$ is from the unified framework of Yu et al. And $^2$ is from Yu et al. [12]. And other results are from the original paper. The highest results are in bold and the next highest results are underlined.

| Model | Years | MOSI | | | | MOSEI | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Acc-2(↑) | F1-Score(↑) | MAE(↓) | Corr(↑) | Acc-2(↑) | F1-Score(↑) | MAE(↓) | Corr(↑) |
| TFN$^1$ | 2017 | 77.99/79.08 | 77.95/79.11 | 0.947 | 0.673 | 78.50/81.89 | 78.96/81.74 | 0.573 | 0.714 |
| LMF$^1$ | 2018 | 77.9/79.18 | 77.8/79.15 | 0.95 | 0.651 | 80.54/83.48 | 80.94/83.36 | 0.576 | 0.717 |
| MulT$^1$ | 2019 | 79.71/80.98 | 79.63/80.95 | 0.88 | 0.702 | 81.15/84.63 | 81.56/84.52 | 0.559 | 0.733 |
| MISA$^1$ | 2020 | 81.84/83.54 | 81.82/83.58 | 0.777 | 0.778 | 80.67/84.67 | 81.12/84.66 | 0.558 | 0.752 |
| Self-MM$^1$ | 2021 | 83.44/85.46 | 83.36/85.43 | 0.708 | 0.796 | 83.76/85.15 | 83.82/84.90 | 0.53 | 0.765 |
| MAG-BERT$^2$ | 2021 | 82.54/84.3 | 82.59/84.3 | 0.731 | 0.789 | 83.79/85.23 | 83.74/85.08 | 0.539 | 0.753 |
| BBFN [25] | 2021 | -/84.30 | -/84.30 | 0.776 | 0.755 | -/86.2 | -/86.1 | 0.529 | 0.767 |
| CubeMLP [41] | 2022 | -/85.6 | -/85.5 | 0.77 | 0.767 | -/85.1 | -/84.5 | 0.529 | 0.76 |
| DEAN [42] | 2022 | -/82.7 | -/82.6 | 0.843 | - | -/83.3 | -/83.2 | 0.571 | - |
| C-MIB [38] | 2022 | -/85.2 | -/85.2 | 0.728 | 0.793 | -/86.2 | -/86.2 | 0.584 | 0.789 |
| PS-Mixer [43] | 2023 | 80.3/82.1 | 80.3/82.1 | 0.794 | 0.748 | 83.1/86.1 | 83.1/86.1 | 0.537 | 0.765 |
| EMT-DLFR [45] | 2023 | 83.3/85.0 | 83.2/85.0 | 0.705 | 0.798 | 83.4/86.0 | 83.7/86.0 | 0.527 | 0.774 |
| GEAR [18] | 2023 | 83.29/84.96 | 83.22/84.95 | - | - | 84.06/85.88 | 84.30/85.79 | - | - |
| BCD-MM(Ours) | | **85.13/86.89** | **85.01/86.82** | **0.688** | **0.805** | **84.73/86.78** | **84.52/86.73** | **0.511** | **0.797** |
| ΔSOTA | | ↑1.69/2.43 | ↑1.65/2.39 | ↓0.017 | ↑0.007 | ↑0.67/0.78 | ↑0.22/0.53 | ↓0.016 | ↑0.008 |

**TABLE 4.** Performance of the different models on the SIMS dataset, in which all models use Chinese-BERT [45] as a text encoder, where $^1$ is from the unified framework for multimodal sentiment analysis of Yu et al.

| Model | Acc-2(↑) | Acc-3(↑) | Acc-5(↑) | F1-Score(↑) | MAE(↓) | Corr(↑) |
|---|---|---|---|---|---|---|
| TFN$^1$ | 78.38 | 65.12 | 39.3 | 78.62 | 0.432 | 0.591 |
| LMF$^1$ | 77.77 | 64.68 | 40.53 | 77.88 | 0.441 | 0.576 |
| MulT$^1$ | 78.56 | 64.77 | 37.94 | 79.66 | 0.453 | 0.564 |
| Self-MM$^1$ | 80.04 | 65.47 | 41.53 | 80.44 | 0.425 | 0.595 |
| BCD-MM(Ours) | **81.84** | **67.18** | **43.76** | **81.44** | **0.412** | **0.608** |
| ΔSOTA | ↑1.8 | ↑1.71 | ↑2.23 | ↑1.00 | ↓0.013 | ↑0.013 |

**TABLE 5.** The performance of different models on the MOSI and MOSEI datasets for the OOD test performance of the baseline in the table are taken from the work of Sun et al. [16].

| Model | OOD MOSI | | | OOD MOSEI | | |
|---|---|---|---|---|---|---|
| | Acc-2(↑) | F1-Score(↑) | Acc-7(↑) | Acc-2(↑) | F1-Score(↑) | Acc-7(↑) |
| TFN | 73.02/74.62 | 72.93/74.56 | 32.95 | 71.23/69.76 | 70.46/69.02 | 41.05 |
| LMF | 73.54/75.27 | 73.40/75.18 | 29.1 | 68.16/69.58 | 68.31/69.58 | 31.11 |
| MulT | 75.00/76.72 | 74.75/76.52 | 29.8 | 72.56/73.73 | 72.44/73.58 | 40.58 |
| MISA | 75.90/77.39 | 75.82/77.35 | 38.05 | 74.48/76.45 | 74.39/76.33 | 43.15 |
| CLUE+MISA | 78.25/79.17 | 78.28/79.19 | 42.25 | 77.17/78.77 | 77.08/78.74 | 46.86 |
| Self-MM | 76.70/78.12 | 76.68/78.13 | 40.25 | 74.68/74.33 | 74.50/74.22 | 45.81 |
| CLUE+Self-MM | 78.75/79.94 | 78.75/79.93 | 41.75 | 77.76/79.48 | 77.72/79.47 | 48.09 |
| MAG-BERT | 75.57/77.28 | 75.52/77.26 | 39.85 | 74.59/76.41 | 74.48/76.27 | 45.88 |
| CLUE+ MAG-BERT | 77.25/78.65 | 77.46/78.83 | 40.75 | 78.34/80.51 | 78.23/80.46 | 48.66 |
| BCD-MM(Ours) | **80.05/81.32** | **80.04/81.30** | **43.05** | **80.22/83.43** | **79.86/82.19** | **49.32** |
| ΔSOTA | ↑1.30/1.46 | ↑1.29/1.37 | ↑0.8 | ↑1.88/1.71 | ↑1.63/1.73 | ↑0.66 |

extractor and a conventional bias extractor. 4) w/o-DbeT. Similarly, the traditional extractor with the Dual bias extractor (DbeT) is removed, and a conventional robust extractor and a Cross-Modal Bias Extractor are employed. 5) w/o-Swap. The swap batch $S_e$ is set to a high number, effectively nullifying the swap operation. 6) w/o-IPW. $\mathcal{L}_{IPW}$ and $\hat{\mathcal{L}}_{IPW}$ are replaced with $\mathcal{L}_{MAE}$ and $\hat{\mathcal{L}}_{MAE}$, removing the weights from the MAE losses as specified in Eqn.(24). 7) w/o-TMAE. To evaluate impact of TMAE loss, we replace the TMAE loss in Eqn.(16) with the MAE loss. The ablation study of the DURD module section in Table 6 reveals several aspects of BCD-MM debiasing performance.

- **Dual bias extractor importance**. Models with two conventional (w/o-Cmbe) or two cross-modal (w/o-Tbe) bias extractors outperformed those with a single extractor (w/o-DbeC, w/o-DbeT). This underscores the critical role of dual bias extractors in enhancing debiasing, particularly in the OOD MOSI dataset. For instance, the Acc-2 of w/o-Tbe is 1.12%/1.03% higher than that of w/o-Cmbe, and w/o-DbeC outperforms

**TABLE 6. BCD-MM ablation studies on the MOSI and MOSI OOD test sets.**

| | MOSI | | | OOD MOSI | | |
|---|---|---|---|---|---|---|
| | Acc-2(↑) | F1-Score(↑) | Acc-7(↑) | Acc-2(↑) | F1-Score(↑) | Acc-7(↑) |
| BCD-MM | **85.13/86.89** | **85.01/86.82** | **48.69** | **80.05/81.32** | **80.04/81.30** | **43.05** |
| **TAE** | | | | | | |
| Average Pooling | 83.88/86.17 | 83.80/86.16 | 43.17 | 78.91/80.10 | 78.82/80.04 | 41.89 |
| Max Pooling | 81.49/83.23 | 81.38/83.19 | 42.27 | 77.21/78.21 | 77.18/78.20 | 38.19 |
| Downsampling | 81.20/83.23 | 81.14/83.25 | 45.19 | 77.12/78.10 | 77.07/78.18 | 40.89 |
| Linear | 82.94/85.21 | 82.88/85.22 | 47.15 | 77.89/79.37 | 77.86/79.35 | 40.37 |
| Convolution | 82.94/83.69 | 82.99/83.78 | 45.63 | 77.64/78.83 | 77.64/78.81 | 40.57 |
| **BCAG** | | | | | | |
| w/o-Transformer | 83.65/84.76 | 83.65/84.69 | 45.58 | 78.57/79.21 | 78.52/79.20 | 41.23 |
| w/o-BCAG | 84.02/85.38 | 84.00/85.28 | 46.19 | 79.37/80.63 | 79.34/80.61 | 41.71 |
| **DUERD** | | | | | | |
| w/o- Cmbe | 83.69/85.91 | 83.66/85.87 | 45.62 | 77.65/78.79 | 77.61/78.76 | 39.94 |
| w/o- Tbe | 83.09/85.31 | 83.01/85.30 | 45.27 | 78.53/79.76 | 78.50/79.75 | 41.01 |
| w/o-DbeC | 82.41/84.72 | 82.36/84.66 | 45.15 | 76.96/78.03 | 76.96/78.02 | 39.12 |
| w/o-DbeT | 82.92/85.16 | 82.85/85.09 | 46.15 | 77.21/78.45 | 77.19/78.41 | 39.01 |
| w/o-Swap | 84.84/86.43 | 84.76/86.39 | 47.06 | 79.42/80.97 | 79.40/80.96 | 42.35 |
| w/o-IPW | 83.91/85.71 | 83.89/85.63 | 47.32 | 78.62/80.33 | 78.41/80.30 | 40.96 |
| w/o-TMAE | 84.01/85.83 | 83.98/85.81 | 47.34 | 78.98/80.06 | 78.98/80.03 | 40.71 |

w/o-DbeT by 0.25%/0.42%. This further validates that dual bias extractors not only advance bias feature extraction but also demonstrate the superior debiasing performance of the cross-modal bias extractor.

- **Effectiveness of different bias extractors**. Comparing w/o-Tbe with BCD-MM shows that varying the bias extractor method enhances bias information extraction and overall model performance.
- **Significance of sample diversity (w/o-Swap)**. In the MOSI and OOD MOSI datasets, the w/o-Swap model, while performing better than the other baselines, compares poorly with the BCD-MM. This implies the importance of diverse swapping samples for obtaining the best results, suggesting that swapping has potential advantages in the field of debiasing.
- **Role of weighted losses (w/o-IPW)**. The notable performance drop in the w/o-IPW model highlights the necessity of assigning smaller weights to highly biased samples for effective debiasing.
- **TMAE loss contribution (w/o-TMAE)**. The subpar performance of the w/o-TMAE variant relative to the original BCD-MM model underscores the value of the TMAE loss. This specific loss is designed to intensify bias during training, in contrast to the standard MAE loss, which might result in only partial learning of robust features. This partial learning can weaken the overall effectiveness of debiasing, as it may lead to inaccurate bias estimation across different modalities.

## VI. CONCLUSION

In this work, we propose a novel multimodal learning debiasing model, abbreviated as BCD-MM. The model not only removes intra-modality redundancies while preserving critical information but also adaptively captures inter-modality and intra-modality information, filters inter-modality incongruent subsequences, and further enhances the debiasing performance of the model through a dual bias extractor. It strengthens generalization by integrating a standard robust extractor, a typical bias extractor, and an innovative cross-modal bias extractor. This setup effectively distinguishes between robust and biased elements in text, audio, and visual data, and it calculates bias weights. The model employs the IPW enhancement loss during training. Thorough testing on the MOSI, MOSEI, and SIMS datasets not only reveals spurious correlations but also highlights the excellent debiasing performance of BCD-MM, especially in out-of-domain test scenarios. Furthermore, our proposed multimodal learning debiasing model, BCD-MM, holds promise for applications in various machine learning domains, including healthcare. We are currently evaluating its effectiveness in mood and depression estimation. The goal of future research will be to expand the application of the BCD-MM by focusing on more tasks and exploring the integration of physiological signals (e.g., heart rate variability and skin conductance) and other modalities, such as facial expressions, to further enhance the robustness and depth of sentiment analysis.

## REFERENCES

[1] Z. Lu, L. Cao, Y. Zhang, C.-C. Chiu, and J. Fan, "Speech sentiment analysis via pre-trained features from end-to-end ASR models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7149–7153.

[2] C. Song, X.-K. Wang, P.-F. Cheng, J.-Q. Wang, and L. Li, "SACPC: A framework based on probabilistic linguistic terms for short text sentiment analysis," *Knowl.-Based Syst.*, vol. 194, Apr. 2020, Art. no. 105572.

[3] L. Zhang, X. Hong, O. Arandjelovic, and G. Zhao, "Short and long range relation based spatio-temporal transformer for micro-expression recognition," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 1973–1985, Oct. 2022.

[4] J. An and W. M. N. Wan Zainon, "Integrating color cues to improve multimodal sentiment analysis in social media," *Eng. Appl. Artif. Intell.*, vol. 126, Nov. 2023, Art. no. 106874.

[5] H. Huang, A. Asemi Zavareh, and M. Begum Mustafa, "Sentiment analysis in e-commerce platforms: A review of current techniques and future directions," *IEEE Access*, vol. 11, pp. 90367–90382, 2023.

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

[7] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Bagher Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2247–2256.

[8] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 1–8.

[9] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," 2016, *arXiv:1606.06259*.

[10] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2236–2246.

[11] W. Yu, H. Xu, F. Meng, Y. Zhu, Y. Ma, J. Wu, J. Zou, and K. Yang, "CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 3718–3727.

[12] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, May 2021, vol. 35, no. 12, pp. 10790–10797.

[13] W. Rahman, M. K. Hasan, S. Lee, A. Bagher Zadeh, C. Mao, L.-P. Morency, and E. Hoque, "Integrating multimodal information in large pretrained transformers," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 2359–2369.

[14] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 3, no. 1, pp. 7216–7223.

[15] Y.-H.-H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 6558–6569.

[16] T. Sun, W. Wang, L. Jing, Y. Cui, X. Song, and L. Nie, "Counterfactual reasoning for out-of-distribution multimodal sentiment analysis," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 15–23.

[17] J. Pearl, *Causality*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[18] T. Sun, J. Ni, W. Wang, L. Jing, Y. Wei, and L. Nie, "General debiasing for multimodal sentiment analysis," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 5861–5869.

[19] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nature Mach. Intell.*, vol. 2, no. 11, pp. 665–673, Nov. 2020.

[20] H. Sun, J. Liu, Y.-W. Chen, and L. Lin, "Modality-invariant temporal representation learning for multimodal sentiment classification," *Inf. Fusion*, vol. 91, pp. 504–514, Mar. 2023.

[21] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-invariant and -Specific representations for multimodal sentiment analysis," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1122–1131.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[23] J.-B. Delbrouck, N. Tits, M. Brousmiche, and S. Dupont, "A transformer-based joint-encoding for emotion recognition and sentiment analysis," in *Proc. ACL*, 2020, pp. 1–7.

[24] S. Sahay, E. Okur, S. H. Kumar, and L. Nachman, "Low rank fusion based transformers for multimodal sequences," in *Proc. 2nd Grand-Challenge Workshop Multimodal Lang.*, 2020, pp. 29–34.

[25] W. Han, H. Chen, A. Gelbukh, A. Zadeh, L.-P. Morency, and S. Poria, "Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2021, pp. 6–15.

[26] A. Chhabra, K. Masalkovaite, and P. Mohapatra, "An overview of fairness in clustering," *IEEE Access*, vol. 9, pp. 130698–130720, 2021.

[27] I. Chen, F. D. Johansson, and D. Sontag, "Why is my classifier discriminatory?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–12.

[28] J. Jung, S. Corbett-Davies, J. D. Gaebler, R. Shroff, and S. Goel, "Mitigating included- and omitted-variable bias in estimates of disparate impact," 2018, *arXiv:1809.05651*.

[29] B. Glymour and J. Herington, "Measuring the biases that matter: The ethical and casual foundations for measures of fairness in algorithms," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2019, pp. 269–278.

[30] B. Salimi, L. Rodriguez, B. Howe, and D. Suciu, "Interventional fairness: Causal database repair for algorithmic fairness," in *Proc. Int. Conf. Manage. Data*, 2019, pp. 793–810.

[31] V. Iosifidis, B. Fetahu, and E. Ntoutsi, "FAE: A fairness-aware ensemble framework," in *Proc. IEEE Int. Conf. Big Data*, Dec. 2019, pp. 1375–1380.

[32] H. Jiang and O. Nachum, "Identifying and correcting label bias in machine learning," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 702–712.

[33] B. Fish, J. Kun, and Á. D. Lelkes, "A confidence-based approach for balancing fairness and accuracy," in *Proc. SIAM Int. Conf. Data Mining*, Jun. 2016, pp. 144–152.

[34] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[35] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How does batch normalization help optimization?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[36] H. Shi, S. Gao, Y. Tian, X. Chen, and J. Zhao, "Learning bounded context-free-grammar via lstm and the transformer: Difference and the explanations," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 8, pp. 8267–8276.

[37] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1839–1848.

[38] S. Mai, Y. Zeng, and H. Hu, "Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations," *IEEE Trans. Multimedia*, vol. 25, pp. 4121–4134, 2022.

[39] J. Nam, H. Cha, S. Ahn, J. Lee, and J. Shin, "Learning from failure: De-biasing classifier from biased classifier," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 20673–20684.

[40] C. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Res.*, vol. 30, no. 1, pp. 79–82, 2005.

[41] H. Sun, H. Wang, J. Liu, Y.-W. Chen, and L. Lin, "CubeMLP: An MLP-based model for multimodal sentiment analysis and depression estimation," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 3722–3729.

[42] F. Zhang, X.-C. Li, C. P. Lim, Q. Hua, C.-R. Dong, and J.-H. Zhai, "Deep emotional arousal network for multimodal sentiment analysis and emotion recognition," *Inf. Fusion*, vol. 88, pp. 296–304, Dec. 2022.

[43] H. Lin, P. Zhang, J. Ling, Z. Yang, L. K. Lee, and W. Liu, "PS-mixer: A polar-vector and strength-vector mixer model for multimodal sentiment analysis," *Inf. Process. Manag.*, vol. 60, no. 2, Mar. 2023, Art. no. 103229.

[44] L. Sun, Z. Lian, B. Liu, and J. Tao, "Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis," *IEEE Trans. Affect. Comput.*, vol. 15, no. 1, pp. 309–325, Jan./Mar. 2024.

[45] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, "Pre-training with whole word masking for Chinese BERT," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 3504–3514, 2021.

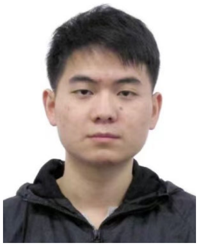
**LEI MA** received the master's degree from Monash University, Australia, in 2005. He is currently an Associate Professor with Kunming University of Science and Technology. His research interests include data mining, machine learning, image processing, and the application of artificial intelligence technology.

**JINGTAO LI** was born in Kunming, Yunnan, in 2000. He received the Bachelor of Science degree from Kunming University of Science and Technology, in 2022, where he is currently pursuing the master's degree in software engineering. His research interests include multi-modality, sentiment analysis, and debiased learning.

**DANGGUO SHAO** (Member, IEEE) received the Ph.D. degree in computer science from Sichuan University, in 2012. Since 2013, he has been with Kunming University of Science and Technology, Kunming, China, where he is currently an Associate Professor. His research interests include image processing, text processing, machine learning, and data mining.

**JIANGKAI YAN** was born in Bijie, Guizhou, in 2000. He received the bachelor's degree in engineering from Guizhou Institute of Technology, in 2022. He is currently pursuing the master's degree in computer application technology with the Kunming University of Science and Technology. His research interests include computer vision, natural language processing, and sentiment analysis.

**JIAWEI WANG** was born in Leshan, Sichuan, in 1999. He received the degree in engineering from Chengdu University, in 2021. He is currently pursuing the master's degree in software engineering with the School of Information Engineering and Computer, Kunming University of Science and Technology. His research interests include natural language processing, sentiment analysis, knowledge graph, and relationship extraction.



**YUKUN YAN** was born in Ziyang, Sichuan, in 1999. He received the bachelor's degree from Shanxi University, in 2022. He is currently pursuing the master's degree in software engineering with Kunming University of Science and Technology. His research interests include machine learning and computer-aided drug design.