Dominant SIngle-Modal SUpplementary Fusion (SIMSUF) for Multimodal Sentiment Analysis

Jian Huang, Yanli Ji, *Member, IEEE*, Zhen Qin, Yang Yang, *Senior Member, IEEE*, and Heng Tao Shen, *Fellow, IEEE*

***Abstract*—Multimodal sentiment analysis remains a big challenge due to the lack of effective fusion solutions. An effective fusion is expected to obtain the correct semantic representation for all modalities, and simultaneously thoroughly explore the contribution of each modality. In this paper, we propose a dominant SIngle-Modal SUpplementary Fusion (SIMSUF) approach to perform effective multimodal fusion for sentiment analysis. The SIMSUF is composed of three major components, a dominant modality supplementary module, a modality enhancement module, and a multimodal fusion module. The dominant modality supplementary module realizes dominant modality determination by estimating mutual dependence between every two modalities, and then the dominant modality is adopted to supplement other modalities for representative feature learning. To further explore the modality contribution, we propose a two-branch modality enhancement module, where one branch learns common representation distribution for multiple modalities, and simultaneously a specific modality enhancement branch is presented to perform semantic difference enhancement and distribution difference enhancement for each modality. Finally, a dominant modality leading fusion module is designed to fuse multimodal representations of two branches for sentiment analysis. Extensive experiments are evaluated on the CMU-MOSEI and CMU-MOSI datasets. Experiment results certify that our approach is superior to the state-of-the-art approaches.**

***Index Terms*—Multimodal fusion, multimodal sentiment analysis, multimodal supplementary, transformer.**

Manuscript received 15 January 2023; revised 3 August 2023 and 16 October 2023; accepted 13 December 2023. Date of publication 26 December 2023; date of current version 31 July 2024. This work was supported in part by the Science and Technology Innovation Committee of Shenzhen Municipalit Foundation under Grant JCYJ20210324132203007, and in part by the National Natural Science Foundation of China under Grant U20B2063 and in part by the Dongguan Songshan Lake Introduction Program of Leading Innovative and En-trepreneurial Talent. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Zakia Hammal. *(Corresponding author: Yanli Ji.)*

Jian Huang, Zhen Qin, and Yang Yang are with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China (e-mail: 202122080121@std.uestc.edu.cn; 202222080113@std.uestc.edu.cn; dlyyang@gmail.com).

Yanli Ji is with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China, and also with the Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen 518100, China (e-mail: yanliji@uestc.edu.cn).

Heng Tao Shen is with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China, and also with the Peng Cheng Laboratory, Shenzhen 518066, China (e-mail: shenhengtao@hotmail.com).

The source code of this work is available at https://github.com/HumanCenteredUndestanding/SIMSUF.

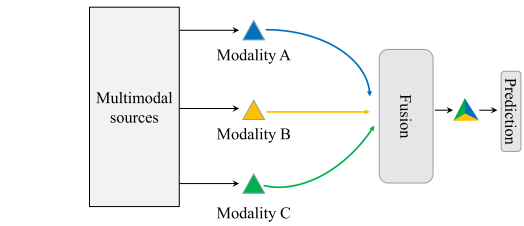
Digital Object Identifier 10.1109/TMM.2023.3344358

I. Introduction

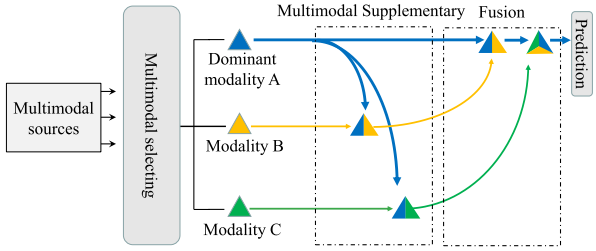
MULTIMODAL sentiment analysis has attracted increasing attention due to the booming of online multimodal source information, such as text, vision, and audio sources [26]. Multimodal sentiment analysis can help us handle user-generated online materials [32] to collect public opinions for some website applications or assist in refining human-centered systems. In contrast to using a single modality, multimodal sentiment analysis could involve sufficient modality sources, which is beneficial for correct analysis. But obviously, there still exist lots of challenges, for example, how to effectively fuse multimodal sources and obtain representative semantic features to improve the performance.

One challenge in multimodal sentiment analysis is how to effectively fuse sentiment information conveyed by different modalities. Fig. 1(a) illustrates the traditional solutions for multimodal fusion, some frameworks balanced all modalities, using the same sub-network for all modality learning, and concatenated outputs for sentiment analysis [16], [51]. However, due to the feature gap that exists in different modalities, contributions to the sentiment analysis from different modalities are much different. Some modalities may provide a larger contribution for correct analysis than others [15], [35], [55], thus traditional methods that adopted equal weights for multimodal fusion are not sufficient. Zadeh et al. [26], [53] performed a fusion operation between paired modalities to investigate multimodal interactions, which altered the priority of paired modalities. Recently, attention learning and Transformer networks are introduced to perform multimodal interaction [42], [45], [49], which provides new solutions for effective multimodal fusion. Sentiment analysis will be more effective if the most important modalities can be explored for their full contribution. Wu et al. [46], [47] focused on the text modality in multimodal fusion by designing a text-centered private and public information extraction network and a network that uses textual information to improve multimodal information. Han et al. [13] presented a text-centered pairwise Transformer fusion network. As depicted in Fig. 1(b), our distinct contribution lies in devising an algorithm that automatically selects the dominant modality. By identifying the most significant modality, we employ it to complement other modalities in representative feature learning and facilitate the multimodal fusion process.

For the sentiment analysis task, there were frameworks only using single modalities, for instance, using the video



(a) Traditional modality fusion.



(b) Supplementary and fusion with dominant modality in our approach.

Fig. 1. Motivation of our approach. (a) In traditional solutions, all modalities are treated equally for fusion. (b) In our approach, we discover the dominant modality and adopt it to perform the supplementary and fusion. It sufficiently explores the superiority of the dominant modality.

modality [38], [48], [58], the audio modality [27], and the text modality [2], [22]. It demonstrates the solid contributions of single modalities in sentiment analysis. However, if given multiple modality sources, one modality's information missing may also evoke a different sentiment perception in some cases [13]. Therefore, regarding multimodal-involved sentiment analysis, it is necessary to explore the latent relationship and correlation among various modalities, and adopt them for modality supplementary and enhancement, to obtain solid semantic representation for all modalities. For this purpose, we draw lessons from the disentanglement operation [14], [21], [23], [44] which always disassembles feature representation to a common space and a specific attribute space for detailed analysis. We design a novel modality enhancement approach that not only shares the focus with previous studies on enhancing the distribution of modality-common and modality-specific features but also places a particular emphasis on semantic enrichment. This comprehensive approach aims to achieve superior multimodal sentiment learning outcomes by effectively weakening and strengthening the representation of modality characteristics.

For multimodal fusion, according to operation strategies, existing sentiment analysis frameworks could be roughly separated into two types, early fusion and later fusion. The early fusion methods generally concatenate multiple modality features tightly following feature learning steps [5], [10], while later fusion methods always integrate prediction results of single modalities after semantic prediction [39]. Obviously, in both two fusion types, different modalities are separately processed in parallel. It completely ignores the semantic correlation among different modalities, which may lead to unstable analysis results. To avoid suffering such problems, we continue to let the dominant modality play a core role in modality fusion. Recently, Transformer performs weighted interactive fusion via attention mechanisms [16], [41], [52]. Instead of treating all modalities equally in prior works, we introduce a dominant modality driven fusion operation. Here, the dominant modality serves as the baseline, and we iteratively interact it with the features from the other two modalities using the Transformer network, aiming to achieve a more comprehensive and enriched fusion process.

In this paper, we propose a dominant SIngle-Modal SUpplementary Fusion (SIMSUF) approach to realize effective multimodal fusion. As shown in Fig. 2, the SIMSUF comprises three major modules: the dominant modality supplementary module, the modality enhancement module, and the dominant-modal driven multimodal fusion module. The dominant modality supplementary (DMS) module is proposed to automatically select the most important modality as the dominant modality, and use it to supplement other modalities, obtaining semantic supplementary representations. The dominant modality determination relies on calculating mutual dependence coefficients between every two modalities and sorting them to find the optimal modality. The modality enhancement (ME) module is presented to further explore the contribution of each modality. It is composed of two enhancement branches, where a common modality enhancement (CME) branch is set to generate common-distribution representations by drawing close multiple-modality distributions, and simultaneously a specific modality enhancement (SPME) branch is presented to perform semantic difference enhancement (SDE) and distribution difference enhancement (DDE) for each modality. Finally, a dominant-modal driven multimodal fusion (DDMF) module is designed to fuse multimodal features, where the dominant modality is used as a baseline for interactive fusion with other modalities.

The main contributions can be summarized as follows:

- We propose a SIMSUF approach for the multimodal sentiment analysis, which consists of a DMS module, an ME module, and a DDMF module.
- The DMS module is designed to automatically determine the most representative modality according to mutual dependence assessment among modalities, and adopts it for modality supplementary.
- The ME module consists of one CME branch which maps multimodal features into a common distribution, and one SPME branch which performs semantic difference enhancement and distribution difference enhancement.
- The DDMF module finally performs interactive fusion among enhanced multimodal representations, where the dominant modality is used as the baseline for emphasizing semantic representations.

The arrangement of this paper is as follows. In Section III, we explain the proposed approach SIMSUF in detail. The experiment settings and datasets are illustrated in Section IV. In the following, Section V clarifies experiments, where a series of ablation studies on novel designs of our proposed approach are discussed, and results obtained by our approach SIMSUF are compared with related frameworks and a series of visualization results are illustrated for comparison. Extensive experimental results enhance the readability and completeness of our proposed approach. A conclusion of the paper is given in Section VI.

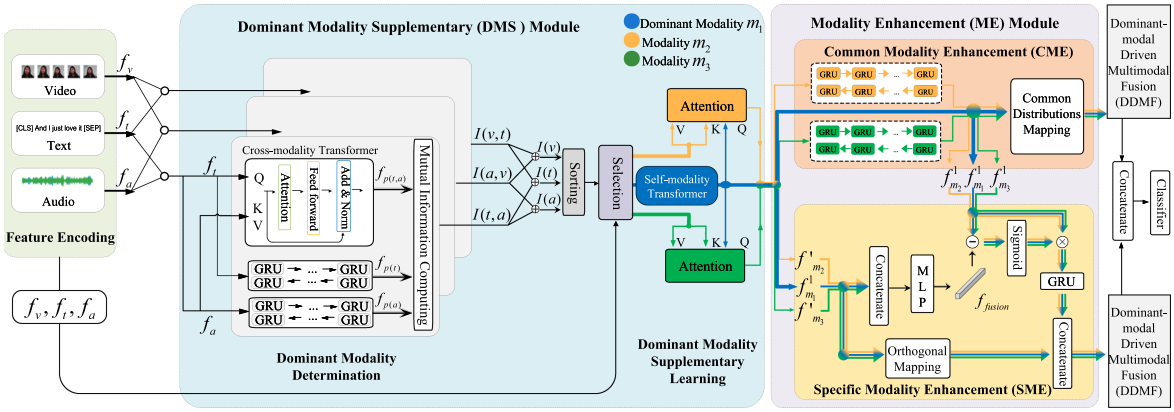


Fig. 2. Overview of our dominant SIngle-Modal SUpplementary Fusion(SIMSUF) approach. It is composed of three major components, the dominant modality supplementary module (DMS), the modality enhancement (ME) module, and the dominant-modal driven multimodal fusion (DDMF) module. We first employ the encoders to extract multimodal features from audio-visual-text sequences. Then we design the DMS module to choose the most important modality as the dominant modality of the whole model, and then perform cross-modal supplementary. Afterward, we design a two-branch ME module to extract the common representation and the unique representation of multimodal features. At last, a DDMF module is designed to realize multimodal fusion.

II. Related Work

A. Multimodal Sentiment Analysis

Multimodal fusion is the crucial problem in the multimodal sentiment analysis task [46]. Most previous frameworks focused on fusion strategies. There were attempts that performed early fusion to connect the multimodal features to generate one fused sentiment representation [5], [10], [36] and late fusion to integrate the prediction results of each single modality for the final decision [34], [39]. Although these two fusion algorithms can handle multimodal data, they are too simple and superficial to effectively deal with the features of each modality and cannot make multimodal fusion more sufficient. Over time, some frameworks attempted to assign weights to features of different modalities according to their contributions to the final sentiment analysis. Arevalo et al. [1] optimized the multimodal fusion by assigning weights to each modality with a gated network. Majumder et al. [31] achieved multimodal fusion through the employment of GRU networks to weigh different modalities. Such an algorithm can make the multimodal fusion have different emphases on different modalities according to the generated weights, but it still lacks the interaction of multimodal features in the fusion of multimodal data. With development, some recent works are mostly committed to calculating correlations among involved modalities. For example, Wang et al. [43] proposed a cross-modal enhancement method to learn the better representation of multimodal features. Although this method effectively designs a multimodal interaction algorithm, it is not an end-to-end network, and the training of the entire model is too complex. Hazarika et al. [16] proposed a representation learning approach that mapped multimodal features to a common subspace and extracted factorized representations. This method effectively designs enhancements to common and unique features of modalities, but lacks adequate fusion of multimodal features. At the same time, various algorithms try to realize the interaction of multimodal features in multimodal emotional analysis. For example, some translation-based [24], [28], [33], [41], [45] methods are proposed to perform modal interaction through cross-modal translation. Different combinations of multimodal information were adopted as vertices to design graphs for the interaction and fusion of multimodal features [20], [30], [50], [54], [60]. Nevertheless, the multimodal sentiment analysis task still remains a big challenge. In our approach, we focus on designing an effective multimodal fusion solution, and further exploring the largest contribution from single modalities.

B. Multimodal Fusion Via Transformer

1) Transformer: The Transformer, with the self-attention mechanism [42], excels in sequential data modeling, outperforming recurrent structures [51]. New Transformer-based models, such as BERT [8], ViT [9], and Swin Transformer [25], have been proposed for language and visual processing, showcasing the Transformer's potential in handling multimodal data.

Recently, Transformers have found success in multimodal tasks. Zhong et al. [19] introduced a self-adaptive neural module Transformer for multimodal integration. Feng et al. [11] devised a cross-modality interaction Transformer to capture dependencies between long-range features and enhance distinctiveness. Zhou et al. [59] presented a multimodal audio-visual Transformer to exploit inherent cues and correspondences. Zhang et al. [31] proposed a multi-stage aggregation module on the Transformer backbone for multimodal language localization.

2) Fusion via Transformer: As attention mechanisms and Transformers become increasingly widely used, several studies on attention mechanisms for multimodal fusion are presented. Wang et al. [45] introduced Transformer to multimodal sentiment analysis by designing a Transformer encoder-decoders multimodal fusion strategy. Deng et al. [7] presented a multimodal fusion algorithm through a deep co-attention Transformer network. Then enlightened by the success of BERT, Yang et al. [49] presented a fusion processing approach that fine-tuned the pre-trained BERT model and used masked multimodal attention to combine text and audio modality information. Deng et al. [6] designed a dense fusion transformer framework to integrate multimodal information. However, the crucial problem of the lack of sophisticated multimodal fusion methods still exists.

III. Proposed Approach

A. Problem Definition

In this study, sentiment analysis is conducted using three modalities: Text, Video, and Audio. The input is unimodal raw sequences $X_m \in \mathbb{R}^{l_m \times d_m}$, where l_m denotes the sequence length, and d_m is the representation vector dimension. And we have $m \in \{t, v, a\}$, where t, v, a denote the three types of modalities–text, video, and audio. The output of the model is a sentiment value $\hat{y}$, which is a real number in the range of $[-3, +3]$.

B. Feature Encoding

Initially, we perform the encoding of the multimodal sequential input X_m into unit-length representations. Particularly, we employ the two unidirectional LSTM [17] networks and one BERT [8] as encoders for multimodal feature extraction, obtaining features f_v, f_a, and f_t for video, audio, and text modalities, as shown in (1).

$$\begin{aligned} f_m &= \mathrm{LSTM}(X_m, \theta_m); m \in \{v, a\} \\ f_t &= \mathrm{BERT}(X_t, \theta_t) \end{aligned} \tag{1}$$

C. Dominant Modality Supplementary Representation

The dominant modality supplementary representation (DMS) module is designed to select the most important modality to bridge the modality gap. The dominant modality supplementary is realized in two steps. We first perform dominant modality determination and use the dominant modality to supplement the feature representations of other modalities.

1) Dominant Modality Determination: We select a modality that has the highest correlation with other modalities as the dominant modality, which is continuously used for multimodal supplementary and enhancement in our approach, assisting in improving the modality interaction efficiency and sentiment analysis performance.

We calculate the mutual dependence among modalities to determine the dominant modality. Following the estimation of mutual coefficients in [3], we first extract the mutual representation and single-modal representations of two modalities, and use them to estimate mutual dependence coefficients. Taking the video modality f_v and text modality f_t as an example, we explain the calculation of mutual dependence coefficients in detail.

We employ a three-layer cross-modality Transformer to simulate the mutual representation learning of $f_{p(v,t)}$ between two modalities. The equation for general self-attention learning is illustrated in (2). For cross-attention learning inside the cross-modality Transformer, we use f_v as Query and f_t as Key and Value for the calculation of mutual representation $f_{p(v,t)}$. The cross-modality mutual representation learning is formulated by (3).

$$\mathrm{Attention}(Q, K, V) = \mathrm{Softmax}\left(\frac{Q^\top K}{\sqrt{d}}\right) V. \tag{2}$$

$$f_{p(v,t)} = \mathrm{Transformer}\left(Q = f_v, K = f_t, V = f_t\right). \tag{3}$$

Secondly, to learn semantic representations, we employ the four-layer bi-directional Gated Recurrent Unit (biGRU) to enhance single-modality features f_v and f_t. The feature enhancement is shown in (4). By setting a loss function to maximize correlation coefficients between two modalities, the biGRU will be trained to enhance the feature representation of each modality and push learned features of different modalities to become closely related.

$$\begin{aligned} f_{p(v)} &= \mathrm{biGRU}(f_v); \\ f_{p(t)} &= \mathrm{biGRU}(f_t). \end{aligned} \tag{4}$$

Finally, we use a two-layer multi-layer perceptron (MLP) to unify two input representations, and project feature representations to generate mutual dependence coefficients. The calculation process is illustrated in (5).

$$\begin{aligned} I(v,t) = \mathrm{MLP}(f_{p(v,t)}, f_{p(t)}) \\ - log(exp(\mathrm{MLP}(f_{p(v)}, f_{p(t)}))). \end{aligned} \tag{5}$$

In the same way, we can obtain mutual dependence coefficients for other modalities, $I(t,a)$ and $I(a,v)$. Because the dominant modality should have the highest correlation with other modalities, as exhibited in (6), we design the loss $\mathcal{L}_{MI}$ which controls model training to maximize mutual dependence coefficients between every two modalities. Moreover, with the constraint of the loss function, it may drive the mutual representation learning and single modality enhancement representation learning to generate consistent outputs, pushing the mutual representation to be more related to single modality enhancement representations. Therefore, the modality that is more semantic representative can significantly affect the mutual dependence coefficients. The modality should be chosen as the dominant modality to guide the multimodal fusion.

$$\mathcal{L}_{MI} = -I(a,v) - I(v,t) - I(t,a). \tag{6}$$

Due to the presence of $\mathcal{L}_{MI}$, we can estimate an upper bound for $I(v,t)$, $I(v,a)$, $I(a,t)$. Moreover, as a consequence, the disparities between the different mutual coefficients of features from the same two modalities, such as $I(v,t)$ and $I(t,v)$, can also be considered negligible.

Then we combine two mutual dependence coefficients related to the same modality, so that we define a reliability coefficient for each modality. The reliability coefficient is calculated by (7). It indicates the relevant relationship between multiple modalities and illustrates the representation ability for the sentiment analysis task. Thus, we determine the dominant modality by sorting reliability coefficients and choosing the largest one. In our approach, the sorting algorithm in the Numpy library was adopted. The dominant modality m_1 is decided to be i if the modality i corresponds to $\mathrm{Max}\{I(i), i \& \in \{a, v, t\}\}$. Other modalities are renamed to be modality m_2, and modality m_3.

$$\begin{aligned} I(i) &= I(i, j_1) + I(i, j_2); \\ i, j_1, j_2 &\in \{a, v, t\}; i \neq j_1 \neq j_2. \end{aligned} \tag{7}$$

2) Dominant Modality Supplementary Learning: To strengthen other modalities, we use the dominant modality

to supplement other modalities. Before the cross-modal supplementary, we first employ a two-layer self-modality Transformer to perform self-enhancement for the dominant modality to obtain a better feature representation, $f^1_{m_1}$. The self-enhancement operation ensures that feature representations of all modalities remain at the same feature level.

We employ two cross-modal attention learning operations to perform the dominant modality supplementary learning, as shown in Fig. 2. The cross-modal supplementary learning for the modality m_2 is performed by (8), where the self-enhancement representation of dominant modality $f^1_{m_1}$ is adopted for the cross-modal supplementary.

$$f\prime_{m_2} = \text{Attention}\left(\omega_Q * f^1_{m_1}, \omega_K * f_{m_2}, \omega_V * f_{m_2}\right). \quad (8)$$

In the same way, we perform cross-modal supplementary for modality m_3 and obtain supplemented representation $f\prime_{m_3}$.

D. Modality Enhancement Learning

With supplemented multimodal representations, we design a modality enhancement (ME) module to further exploit correlation among multiple modalities and specific characters of each modality. We perform two types of modality enhancement, common modality enhancement (CME), and specific modality enhancement (SPME). The CME is proposed to perform a common-distribution mapping for representations of multiple modalities. Unlike previous work only focusing on distribution difference, our SPME involves semantic difference enhancement (SDE) and distribution difference enhancement (DDE).

1) Common Modality Enhancement: The CME operation is designed to extract multimodal representations that fall in the common feature distribution. To transfer features to a common distribution, we employ the biGRU network to learn new representations for modality m_2, m_3, and obtain transferred representations $f^1_{m_2}$ and $f^1_{m_3}$. After the encoding, representations of less important modality m_2 and m_3, $f^1_{m_2}$ and $f^1_{m_3}$, should be mapped into a common feature distribution with $f^1_{m_1}$.

To keep the distribution consistency of multimodal representations, the Central Moment Discrepancy (CMD) [16], [57] is employed to measure the similarity of feature distributions. The CMD is used as a loss function to train the common modality enhancement module. Let A and B be bounded sets with respective probability distributions in the interval $[a, b]$. The definition of CMD is shown in (9).

$$\text{CMD}_k(A, B) = \frac{1}{|b-a|}\|\text{E}(A) - \text{E}(B)\|_2 + \sum_{k=2}^{K} \frac{1}{|b-a|^k} \|C_k(A) - C_k(B)\|_2. \quad (9)$$

where, $\text{E}(A) = \frac{1}{|A|}\sum_{x\in A} x$ calculates the expectation value of A, and $C_k(A) = \mathbf{E}((x - \mathbf{E}(A))^k)$ calculates central values for all k^{th} order sample central moments of A. In our approach, we simplify the measurement CMD_k to only consider the first order, $k = 1$. The definition of the CMD loss function of our approach is described in (10). Once the expectation and variance values of features are similar, feature distributions are regarded to be common.

$$\mathcal{L}_{\text{sim1}}\text{CMD}_1\left(f^1_{m_2}, f^1_{m_1}\right) + \text{CMD}_1\left(f^1_{m_3}, f^1_{m_1}\right). \quad (10)$$

2) Specific Modality Enhancement: We design two types of modality-specific feature learning solutions to enhance representation for each modality, the SDE, and the DDE. The solution thoroughly covers both semantic and feature distribution two feature learning aspects.

Semantic Difference Enhancement: Firstly, we learn each modality's unique information from the semantic aspect and use it to enhance the representation of each modality. To extract semantic modality-specific information, we first learn an average semantic representation f_{fusion}. We employ a two-layer MLP network to integrate three features f'_{m_2}, f'_{m_3} and $f^1_{m_1}$, and project integrated feature to a lower dimension. The generation process of the f_{fusion} is shown in (11).

$$f_{fusion} = \text{MLP}\left(f^1_{m_1} \oplus f'_{m_2} \oplus f'_{m_3}\right) \quad (11)$$

where $\oplus$ refers to the concatenation operation.

Following that, we adopt a classifier to assign the semantic attribute to the f_{fusion}. The classifier is applied to predict the sentiment category y' with f_{fusion}. A Huber loss with delta=1, as shown in the (12), is set to train our model to learn the semantic-oriented average representation f_{fusion}.

$$y' = \text{Classifier}(f_{fusion});$$
$$\mathcal{L}_{\text{sem}} = \begin{cases} 0.5 * |y - y'|^2, & \text{if } |y - y'| < 1. \\ |y - y'| - 0.5, & \text{others.} \end{cases} \quad (12)$$

To keep consistency with dominant modalities, we update the feature distribution of f_{fusion} to make it fall into the common feature distribution with the dominant modality m_1. The loss $\mathcal{L}_{\text{sim2}}$ in (13) is set for common-distribution feature learning. The operation ensures the leadership contribution from the dominant modality.

$$\mathcal{L}_{\text{sim2}} = \text{CMD}_1\left(f_{fusion}, f^1_{m_1}\right). \quad (13)$$

The semantic enhancement is realized by a semantic differential calculation and an enhancement operation. The detailed calculation for modality m_1 is exhibited in (14). As illustrated, the difference value is calculated between f_{fusion} and $f^1_{m_1}$, and we also use $f^1_{m_1}$ and the difference value to perform enhancement operation. It needs to be noted that f_{fusion} is obtained using three features f'_{m_2}, f'_{m_3} and $f^1_{m_1}$, which are not enhanced by the common modality enhancement operation. The differences between f_{fusion} and $f^1_{m_1}$, $f^1_{m_2}$ and $f^1_{m_3}$ are able to significantly exhibit the discrepancy between representations of single modalities and the multimodal average representation. Thus difference values $f_{fusion} - f^1_m$ can be used as bias filters to obtain semantic modality-specific representations.

$$f^2_{m_1} = \text{GRU}\left(\text{Sigmoid}\left(\text{FC}\left(f_{fusion} - f^1_{m_1}\right)\right) \cdot f^1_{m_1}\right). \quad (14)$$

In the same way, we may obtain semantic modality-specific representations $f^2_{m_2}, f^2_{m_3}$ for modality m_2, m_3.

Distribution Difference Enhancement: The DDE is realized by an orthogonal mapping method, which is employed to extract

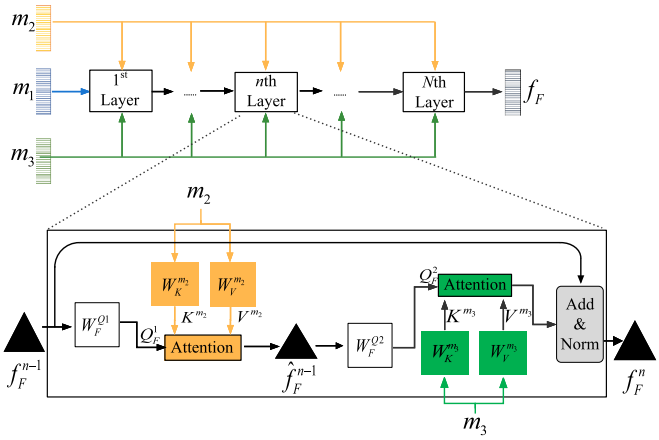


Fig. 3. Flowchart of the dominant-modal driven multimodal fusion module. It is composed of N layers, and each layer involves two cross-modal attention calculations. The modality m_1 is continuously used as the dominant modality, interacting with other modalities.

modality-specific representations that contain less common distribution regions.

For the orthogonal mapping, we adopt a two-layer MLP network to map modality features $f_{m_1}^1$, f_{m_2}' and f_{m_3}' into orthogonal spaces to obtain $f_{m_1}^3$, $f_{m_2}^3$ and $f_{m_3}^3$ with the orthogonal mapping loss $\mathcal{L}_{\text{diff}}$. We perform the orthogonal mapping following the Squared Frobenius Norm theory [16], and define the orthogonal mapping loss $\mathcal{L}_{\text{diff}}$ in (16). Training networks with the loss function, we may generate modality-specific representations that have the largest differences in distribution. In this way, we learn modality-specific representations from the distribution aspect.

$$f_{m_1}^3 = \text{ML\&P}(f_{m_1}^1); f_m^3 = \text{MLP}(f_m'); \&m \in \{m_2, m_3\}. \tag{15}$$

$$\mathcal{L}_{\text{diff}} = \left\|f_{m_1}^{3\top} \cdot f_{m_2}^3\right\|_F + \left\|f_{m_2}^{3\top} \cdot f_{m_3}^3\right\|_F + \left\|f_{m_1}^{3\top} \cdot f_{m_3}^3\right\|_F. \tag{16}$$

Finally, we connect two types of modality-specific features and use a one-layer MLP network for fusion again to obtain final modality-specific representations. Taking the modality m_1 as an example, the fusion operation is illustrated in (17).

$$f_{m_1}^4 = \text{MLP}\left(f_{m_1}^2 \oplus f_{m_1}^3\right). \tag{17}$$

In the same way, we can get $f_{m_2}^4$ and $f_{m_3}^4$.

E. Dominant-Modal Driven Multimodal Fusion Module

The dominant-modal driven multimodal fusion (DDMF) module is designed to perform a fully interactive fusion on multimodal features. As illustrated in Fig. 3, this module is an iterative multi-layer Transformer network. Different from prior works treating multimodal features equally while fusing, we use modality m_1 as the dominant modality, interacting with other modalities in every layer. We use this module to fuse enhanced multimodal representations after the common modality enhancement operation and specific-modality enhancement operation, respectively.

As shown in Fig. 3, the modal-guided fusion module consists of total N fusion layers. We take the nth layer as an example to explain the fusion operation in detail. Firstly, we denote three inputs of the nth layer as f_F^{n-1}, f_{m_2} and f_{m_3}, where f_F^{n-1} is the $(n-1)$th layer's output whose initial input belonging to the modality m_1. The interactive fusion between f_F^{n-1} and the feature of modality m_2 is explained in (18). The fusion operation of the first-step result $\hat{f}_F^{n-1}$ and the feature of modality m_3 is illustrated in (19). After finishing one layer's interactive fusion, we obtain the fused representation f_F^n, which will serve as the dominant modality input for the next layer.

$$\begin{aligned} Q_F^1 &= W_F^{Q1} * f_F^{n-1}; K^{m_2} = W_K^{m_2} * f_{m_2}; \\ V^{m_2} &= W_K^{m_2} * f_{m_2}; \\ \hat{f}_F^{n-1} &= \text{Attention}\left(Q_F^1, K^{m_2}, V^{m_2}\right). \end{aligned} \tag{18}$$

$$\begin{aligned} Q_F^2 &= W_F^{Q2} * \hat{f}_F^{n1}; K^{m_3} = W_K^{m_3} * f_{m_3}; \\ V^{m_3} &= W_V^{m_3} * f_{m_3}; \\ f_F^n &= \text{Norm}(\text{Attention}(Q_F^2, K^{m_3}, V^{m_3}) + f_F^{n-1}). \end{aligned} \tag{19}$$

After N-layer iterations, we finally obtain a fusion representation f_F. Given common modality enhancement representations, $f_{m_1}^1$, $f_{m_2}^1$ and $f_{m_3}^1$, we fuse them using the modal-guided fusion module, and we can obtain a fused representation f_C. In the same way, with modality-specific representations, $f_{m_1}^4$, $f_{m_2}^4$ and $f_{m_3}^4$, we can obtain the fused representation f_S.

F. Sentiment Prediction

We perform sentiment prediction by (20).

$$\hat{y} = \text{Classifier}(f_C \oplus f_S). \tag{20}$$

We employ the Huber loss with delta=1 to compose the loss $\mathcal{L}_{\text{task}}$ for sentiment recognition.

$$\mathcal{L}_{\text{task}} = \begin{cases} 0.5 * |y - \hat{y}|^2, & \text{if } |y - \hat{y}| < 1. \\ |y - \hat{y}| - 0.5, & \text{others.} \end{cases} \tag{21}$$

The overall loss of our SIMSUF is shown in (22).

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \alpha\mathcal{L}_{\text{sim1}} + \beta\mathcal{L}_{\text{sim2}} + \gamma\mathcal{L}_{\text{diff}} + \delta\mathcal{L}_{\text{sem}} + \epsilon\mathcal{L}_{\text{MI}}. \tag{22}$$

where α, β, γ, δ, and ϵ are weights that determine the contribution of each loss component to the overall loss.

IV. Experiment Settings

A. Datasets

The **CMU-MOSI** dataset [56] is one of the most popular benchmark datasets for multimodal sentiment analysis. It comprises 2,199 short monologue video clips taken from 93 Youtube movie review videos. Human annotators label each sample with a sentiment score from -3 (strongly negative) to 3 (strongly positive).

The **CMU-MOSEI** dataset [54] is an upgraded version of the CMU-MOSI dataset. It is also enriched in terms of the versatility of speakers and covers a broader scope of topics. The dataset contains 23,453 video segments, which are annotated in the

TABLE I
RESULT COMPARISON WITH THE STATE-OF-THE-ART APPROACHES IN TWO CHALLENGING DATASETS, THE CMU-MOSEI AND THE CMU-MOSI DATASETS

Approaches	CMU-MOSEI					CMU-MOSI				
	MAE↓	Corr↑	Acc-7↑	Acc-2↑	F1↑	MAE↓	Corr↑	Acc-7↑	Acc-2 ↑	F1↑
DFF-ATMF [4]	-	-	-	77.10	78.30	-	-	-	80.90	81.20
MISA [16]	0.557	0.748	51.70	84.90	84.80	0.817	0.748	41.40	82.10	82.00
ICCN [40]	0.565	0.713	51.58	84.18	84.15	0.862	0.714	39.01	83.07	83.02
Self-MM [51]	0.530	0.765	53.87	85.30	85.17	0.713	0.798	**46.67**	85.98	**85.98**
TCSP [46]	0.576	0.715	-	82.80	82.70	0.908	0.710	-	80.90	81.00
MMIM [14]	**0.526**	**0.772**	54.24	85.97	85.94	**0.700**	**0.800**	**46.65**	**86.06**	85.98
BBFN [13]	0.529	0.767	**54.80**	**86.20**	**86.10**	0.776	0.755	45.00	84.30	84.30
CubeMLP [18]	0.529	0.760	**54.90**	85.10	84.50	0.770	0.767	45.50	85.60	85.50
MMLatch [12]	0.582	0.704	52.10	82.80	82.90	-	-	-	-	-
SIMSUF (Ours)	**0.529**	**0.772**	53.68	**86.23**	**86.12**	**0.709**	**0.802**	45.72	**86.08**	**85.98**

We bolded the top two results of each metric. According to the comparison, our approach outperforms SOTA approaches in both two large-scale datasets.

same way as CMU-MOSI. These segments are extracted from 5,000 videos involving 1,000 distinct speakers and 250 different topics.

B. Experiment Settings

The proposed SIMSUF is trained by using Adamax as the optimizer. We set the batch size to 180 for the CMU-MOSEI dataset and 120 for the CMU-MOSI dataset. The learning rate is initialized as 0.001 and decreases by a factor of 10 in every 10 epochs. The weights of loss functions α, β, γ, δ and ϵ are set to 1/3, 1/6, 1/5, 1/4 and 1, respectively. Then we initialize fully connected layer parameters to 0 for efficient training. We set the dropout of LSTM and GRU to 0.1. The dimension of the feed-forward layer in all Transformers is set to 2560, and the dropout of feed feed-forward layer is set to 0.1. The gradient clipping value is set to 1.0. We set the RELU as the activation function for all MLPs we used in SIMSUF. For the CMU-MOSEI dataset, the early stopping threshold is set to 10, and it is set to 20 for the CMU-MOSI dataset.

C. Evaluation Metrics

We use the same metric set that has been consistently presented and compared in prior works.

Mean absolute error (MAE) refers to the average value of the distance between the model prediction values and the truth values. It measures the error of the model's predictions.

Pearson correlation (Corr) measures the degree of prediction skew. Corr can be used to measure the linear correlation between the predicted values and the truth values.

Seven classification accuracy (Acc-7(%)) indicates the proportion of predictions that correctly fall into seven intervals between -3 and +3 as the corresponding truths. It measures the accuracy of the model on fine-grained sentiment analysis

Binary classification accuracy (Acc-2(%)) indicates the proportion of predictions that correctly fall into positive/negative as the corresponding truths. It measures the accuracy of the model on coarse-grained sentiment analysis

F1 score (F1) are results for non-negative/negative classification. It takes into account both the precision and recall of the prediction model.

V. EXPERIMENT RESULTS

A. Comparison With the SOTA Approaches

We compare with a variety of baseline methods in multimodal sentiment analysis. The results are listed in Table I.

CMU-MOSEI dataset: As illustrated, our model significantly outperforms the State-of-the-Art (SOTA) approaches in metrics of Corr, Acc-2, and F1 scores in the CMU-MOSEI. For the metrics of Acc-7 and MAE, our method achieves a close performance to SOTA. Compared with the MISA approach [16], we obtain a significant improvement. Our approach obtains a result improvement of more than 1.33% in Acc-2 and more than 1.98% in Acc-7 in the CMU-MOSEI dataset. In addition, our SIMSUF improves the results of Corr and Acc-2 by 0.5% and 0.03% respectively compared with the SOTA approach, BBFN [13] in the CMU-MOSEI dataset. Compared to recent approaches, our method demonstrates superiority with an improvement of 1.13% in ACC2 and 1.62% in F1 over CubeMLP [18]. Additionally, in comparison to MMLatch [12], our method exhibits achieves an enhancement of 3.43% in ACC2 and 3.22% in F1. These results underscore the effectiveness of our approach, showcasing the benefits of modality enhancement and dominant modality-driven fusion.

CMU-MOSI dataset: The experiment results show that our model also outperforms the SOTA approaches in metrics of Corr, Acc-2, and F1 scores in the CMU-MOSI. For the metrics of Acc-7 and MAE, our method achieves a close performance to SOTA. Our SIMSUF improves the results of Corr and Acc-2 by 0.2% and 0.02% respectively compared with the SOTA approach, MMIM [14] in the CMU-MOSI dataset. Compared to recent approaches, our method demonstrates superiority with an improvement of 0.48% in ACC2 and 0.48% in F1 over CubeMLP [18].

Overall, these results certify the effectiveness of our proposed SIMSUF approach for multimodal sentiment analysis.

B. Ablation Study

1) Evaluation on Components in SIMSUF: We evaluate three modules in our SIMSUF model and some minor components in these modules. The experiment results are listed in Table II.

TABLE II
EVALUATION ON COMPONENTS OF OUR MODEL IN THE CMU-MOSEI DATASET

Description	CMU-MOSEI				
	MAE↓	Corr↑	Acc-7↑	Acc-2↑	F1↑
w/o DMS module	0.540	0.766	53.48	85.52	85.43
w/o SME module	0.551	0.763	52.98	85.29	85.24
w/o DDMF module	0.532	0.768	53.55	85.95	85.85
w/o biGRU in DMS module	0.534	0.767	53.58	86.04	85.96
w/o dominant modality supplementary learning	0.536	0.763	53.25	85.65	85.55
w/o Transformer in DMS	0.536	0.769	53.56	86.03	85.96
w/o selection of dominant modality + random select	0.535	0.765	53.48	85.77	85.71
w/o CME operation	0.537	0.768	53.60	85.88	85.79
w/o SPME operation	0.548	0.761	53.16	85.33	85.26
w/o SDE operation	0.542	0.767	53.56	85.70	85.62
w/o DDE operation	0.545	0.762	53.25	85.58	85.48
replacing two enhancement operations with projection layers	0.543	0.766	53.47	85.65	85.56
classifying directly with the pre-extracted features	0.575	0.716	49.58	79.35	79.28
SIMSUF (Ours)	**0.529**	**0.772**	**53.68**	**86.23**	**86.12**

The results certify that three major modules give their contributions to sentiment analysis.

The experiment results significantly decrease if any one of the three modules is removed. They all provide solid contributions to multimodal sentiment analysis. Results also certify the most important module of the whole network is the ME module.

For the DMS module, we evaluate the contributions of the biGRU, Transformer, supplementary learning, and dominant modality selection methods, respectively. The experiment results certify the most crucial block in the dominant modality supplementary module is the dominant modal supplementary learning, which helps to make the representations of each modality more balanced and make the other modality closer to the most important modality. It also can be seen that when other parts in this module are eliminated, the results also decrease. It demonstrates the necessity of biGRU, Transformer, and dominant modality selection.

For the ME module, we analyze the necessity of two branches, and two difference enhancement operations in the SPME branch. The results decrease when any one of these two branches and two difference enhancement operations is eliminated, which certifies that these blocks and operations are indispensable. The experiment results also demonstrate the most important part is the SPME, which helps to learn the unique representation of each modality. We replace the CME with a shared projection layer, and also replace the SPME module with three disjoint projection layers. The experiment result is worse than the experiment result of SIMSUF. It shows that our enhancement operations effectively extract modality-common and modality-specific information.

We further analyze the effectiveness of our designed model. We perform the multimodal sentiment analysis directly with the pre-extracted features. The experiment result is far worse than SIMSUF's, which shows all our modules play positive roles in sentiment analysis.

2) Evaluation on Feature Dimension: What is the optimal setting of feature dimension in our model? We perform the experiment to evaluate the most suitable dimension setting in our model. Recognition results are shown in Table III. The feature dimension refers to the dimension of $\{f_m\}$, $\{f_m^1\}$, $\{f_m^2\}$, $\{f_m^3\}$, $\{f_m^4\}$, f'_{m_2} and f'_{m_3}.

TABLE III
EVALUATION ON FEATURE DIMENSION OF OUR MODEL IN THE CMU-MOSEI DATASET

Feature dimension	CMU-MOSEI				
	MAE↓	Corr↑	Acc-7↑	Acc-2↑	F1↑
32	0.535	0.769	53.62	86.15	86.06
64	**0.529**	**0.772**	**53.68**	**86.23**	**86.12**
128	0.532	0.768	53.65	86.08	86.00

The feature dimension denotes the dimensions of $\{f_m\}$, $\{f_m^1\}$, $\{f_m^2\}$, $\{f_m^3\}$, $\{f_m^4\}$, f'_{m_2} AND f'_{m_3}.

The experiments are performed by setting the feature dimensions as 32, 64, and 128, respectively. When the feature dimension is set to 32 or 128, the experiment results decrease. When the hidden dimension equals 64, the result achieves the best. Hence, it demonstrates that setting the feature dimension of the whole model to 64 is the most appropriate. The dimension's setting of 64 makes the length of the feature not too short to lose information, and also avoids obtaining inaccurate feature representation due to too long dimension.

3) Evaluation on Layer Numbers of Transformers: The cross-modality Transformer is in the dominant modality determination block to learn the mutual representation between different modalities. The self-modality information is in the dominant modality supplementary learning block to enhance the representation of the dominant modality. As each Transformer is comprised of several sequential encoder layers, we further analyze the optimal layer-number settings of two Transformers. Recognition results are shown in Table IV.

For the cross-modality Transformer in the dominant modality determination, we perform experiments by setting layer numbers from 1 to 5, respectively. Observing results in Table IV, when the layer number is set to 3, the best result is achieved. For the self-modality Transformer in the dominant modality supplementary learning block, we perform the experiments by setting layer numbers from 1 to 4, respectively. As illustrated in Table IV, when the layer number equals 2, the best results are achieved. Therefore, in our SIMSUF model, the optimal layer number for the cross-modality Transformer is 3, and we set 2 layers for the

TABLE IV
EVALUATION ON OPTIMAL LAYER NUMBERS OF TRANSFORMERS IN THE CMU-MOSEI DATASET

Transformers	Layers	CMU-MOSEI				
		MAE↓	Corr↑	Acc-7↑	Acc-2↑	F1↑
Cross-modality Transformer	1	0.530	0.772	53.65	86.20	86.11
	2	0.530	0.772	53.68	86.22	86.12
	3	**0.529**	**0.772**	**53.68**	**86.23**	**86.12**
	4	0.530	0.772	53.62	86.18	86.10
	5	0.532	0.771	53.62	86.18	86.10
Self-modality Transformer	1	0.533	0.768	53.56	86.15	86.08
	2	**0.529**	**0.772**	**53.68**	**86.23**	**86.12**
	3	0.532	0.769	53.55	86.16	86.08
	4	0.537	0.765	53.54	86.09	86.02

The cross-modality transformer and the self-modality transformer are both employed in the DMS module.

TABLE V
EVALUATION ON LOSS SETTINGS IN THE CMU-MOSEI DATASET

Loss functions	CMU-MOSEI				
	MAE↓	Corr↑	Acc-7↑	Acc-2↑	F1↑
w/o $\mathcal{L}_{\text{sim1}}$	0.547	0.766	53.57	85.17	85.12
w/o $\mathcal{L}_{\text{sim2}}$	0.542	0.769	53.45	85.25	85.15
w/o $\mathcal{L}_{\text{diff}}$	0.548	0.762	53.05	86.08	85.88
w/o $\mathcal{L}_{\text{sem}}$	0.541	0.767	53.55	85.62	85.56
w/o $\mathcal{L}_{\text{MI}}$	0.533	0.770	53.32	85.88	85.79
All loss	**0.529**	**0.772**	**53.68**	**86.23**	**86.12**

Experiment results certify the contribution from all loss functions in the SINMSUF.

self-modality Transformer to obtain optimal performance. Such settings of the number of layers enable the two Transformers to learn features from the sequence effectively. The layer number of the Transformers also is not set too much to increase Flops and cause inaccurate encoding.

4) Evaluation on Loss Settings: We evaluate the contribution of five loss functions. Results are shown in Table V. When we remove the loss $\mathcal{L}_{\text{sim1}}$, the experiment results drop sharply. It demonstrates that the loss $\mathcal{L}_{\text{sim1}}$ is rather important because it helps to obtain the similar feature distributions of multimodal features in common modality enhancement operation, and also contributes to providing features as inputs to specific modality enhancement. It can also be seen that when $\mathcal{L}_{\text{sim2}}$ is eliminated, the results become worse. It demonstrates that the $\mathcal{L}_{\text{sim2}}$ helps to make the fused feature have similar feature distribution with the dominant modality, so that we can extract the semantic difference information effectively. When we remove $\mathcal{L}_{\text{diff}}$, all results also decrease. It certifies the loss $\mathcal{L}_{\text{diff}}$ can make the distributions of multimodal features have the largest difference in the modality-specific enhancement operation. Worse results caused by the absence of $\mathcal{L}_{\text{sem}}$ or $\mathcal{L}_{\text{MI}}$ certify that $\mathcal{L}_{\text{sem}}$ pushes our model to fuse semantic representation for representative feature learning, and $\mathcal{L}_{\text{MI}}$ helps to maximize the degree of correlation between the dominant modality and other modalities. According to these analyses, obviously, results decrease when any one of these loss functions is removed. It certifies all loss settings are essential in our model.

5) Evaluation on the Mutual Coefficients: We have performed experimental verification to assess the disparities in mutual coefficients among different modalities. After normalizing the estimated mutual coefficients of the best epoch to the range of 0 to 1, we calculate the absolute differences between them. The corresponding results are shown in Table VII. The mutual coefficients are mathematically defined by (5) and subject to constraints imposed by the loss function $\mathcal{L}_{\text{MI}}$, aiming to approach the upper bound. Our observations reveal that the differences in mutual coefficients between the two modalities in question are exceedingly small, rendering them negligible in comparison to the disparities observed among distinct modalities. This behavior arises from the mutual coefficients' inclination to approximate the upper limit, leading to an exceedingly minute difference between the two variables. These significant findings underscore the effectiveness of our dominant modality determination operation, which successfully selects the most influential modality based on the mutual coefficients.

6) Evaluation on the Weights of Loss Functions: We conduct ablation experiments on the weight parameters of the loss function to determine the optimal combination. These parameters, denoted as α, β, γ, δ, and ϵ, are defined by (22). The experimental results are presented in Table VI. Notably, the effectiveness of the model was assessed under various weight combinations for α, β, γ, δ, and ϵ. We observed that when the settings of parameters α, β, γ, δ, and ϵ approach values close to 0 or 1, the experimental outcomes exhibit suboptimal performance. This observation implies that the settings of parameters should be kept at a distance from extreme values. The results of the experiments revealed that the best performance is achieved when the weight values were set as follows: $\alpha = 1/3$, $\beta = 1/6$, $\gamma = 1/5$, $\gamma = 1/4$, and $\epsilon = 1$. This particular combination led to the most favorable outcomes in the evaluation metrics. Conversely, when employing alternative weight combinations, the experimental performance experienced a decline. The observed superiority of the specific weight values can be attributed to their appropriate balance in capturing the contributions of various components in the loss function.

7) Evaluation on the Order of CMD: In our investigation, we delve deeper into the impact of the order of CMD, as defined by (9). As revealed in Table VIII, the experimental performance reaches its peak when the order is set to 1. This favorable outcome can be attributed to the direct constraint of mean and variance consistency between the two features when the CMD loss order is set to 1. Such a constraint maximizes the shared representation among the three modalities through the Common Distributions Mapping operation. The features f_{m1}^1, f_{m2}^1, and f_{m3}^1, originating from different modalities, undergo the Specific Modality Enhancement operation, wherein their differences with the simply fused feature f_{fusion} are computed. When a higher order is employed, the CMD-constrained features will exhibit certain similarities in higher dimensions. Consequently, the calculation of differences between f_{m1}^1, f_{m2}^1, f_{m3}^1, and f_{fusion} along a single dimension introduces significant redundant information in the high-dimensional space. As a result, the effectiveness of the Semantic Difference Enhancement operation may be compromised.

TABLE VI
EVALUATION OF THE WEIGHTS OF LOSS FUNCTIONS IN THE CMU-MOSEI DATASET

α	β	γ	δ	ϵ	MAE↓	CORR↑	ACC-7↑	ACC-2↑	F1↑
1/3	**1/6**	**1/5**	**1/4**	**1**	**0.529**	**0.772**	**53.68**	**86.23**	**86.12**
1/9	1/6	1/5	1/4	1	0.546	0.766	53.57	85.25	85.16
1/4	1/6	1/5	1/4	1	0.535	0.768	53.61	86.02	85.9
1/2	1/6	1/5	1/4	1	0.532	0.769	53.58	85.99	85.86
1	1/6	1/5	1/4	1	0.544	0.762	53.12	85.04	84.99
1/3	1/9	1/5	1/4	1	0.542	0.770	53.42	85.22	85.12
1/3	1/7	1/5	1/4	1	0.532	0.769	53.55	86.02	85.89
1/3	1/5	1/5	1/4	1	0.534	0.767	53.62	85.97	85.85
1/3	1	1/5	1/4	1	0.546	0.760	53.08	85.12	85.02
1/3	1/6	1/9	1/4	1	0.546	0.761	53.02	86.00	85.93
1/3	1/6	1/6	1/4	1	0.530	0.77	53.52	85.99	85.9
1/3	1/6	1/4	1/4	1	0.532	0.768	53.5	85.95	85.82
1/3	1/6	1	1/4	1	0.551	0.758	52.89	84.99	84.89
1/3	1/6	1/5	1/9	1	0.542	0.765	53.53	85.58	85.50
1/3	1/6	1/5	1/5	1	0.533	0.767	53.6	86.08	85.94
1/3	1/6	1/5	1/3	1	0.53	0.770	53.59	86.05	85.92
1/3	1/6	1/5	1	1	0.549	0.762	53.19	85.222	85.18
1/3	1/6	1/5	1/4	1/9	0.535	0.770	53.34	85.88	85.78
1/3	1/6	1/5	1/4	10/9	0.529	0.770	53.65	86.15	86.02

$\alpha, \beta, \gamma, \delta$ and ϵ are the weights of the loss functions.

TABLE VII
EVALUATION ON THE DIFFERENCE OF COEFFICIENTS BETWEEN TWO MODALITIES

$\lvert I(a,v)-I(a,t)\rvert$	$\lvert I(a,t)-I(t,v)\rvert$	$\lvert I(t,v)-I(a,v)\rvert$	$\lvert I(v,t)-I(t,v)\rvert$
0.402702567	0.023183851	0.379519316	0.000237294

TABLE VIII
EVALUATION ON THE ORDER OF CMD IN THE CMU-MOSEI DATASET

CMD Order	MAE ↓	CORR ↑	ACC-7 ↑	ACC-2 ↑	F1 ↑
1	**0.529**	**0.772**	**53.68**	**86.23**	**86.12**
2	0.533	0.769	53.62	85.89	85.78
3	0.538	0.768	53.60	85.42	85.30
4	0.539	0.768	53.57	85.25	85.10

Experiment results certify the best choice is 1.

TABLE IX
COMPARISON OF THE PARAMETER QUANTITY AND FLOPS WITH SOTA METHODS IN THE CMU_MOSEI DATASET

methods	Parameter(M)↓	FLOPs(M)↓	Acc-2↑
Mag-BERT[37]	86.9*	17.3*	84.70
Self-MM[51]	85.9*	17.0*	85.17
MMIM[14]	109.8*	47.5*	85.97
SIMSUF(Ours)	**109.9**	**49.2**	**86.23**

* denotes the parameter and FLOPs are estimated by our experiments.

C. Visualization Results

We further visualize feature distributions of hidden representations in our SIMSUF model via the tSNE projections [29]. Fig. 4 illustrates feature distributions of $\{f_m\}$, $\{f_m^1\}$, $\{f_m^2\}$, $\{f_m^3\}$, $\{f_m^4\}$, f'_{m_2} and f'_{m_3}, where the color blue represents modality m_1, the color orange represents modality m_2 and the color green represents modality m_3.

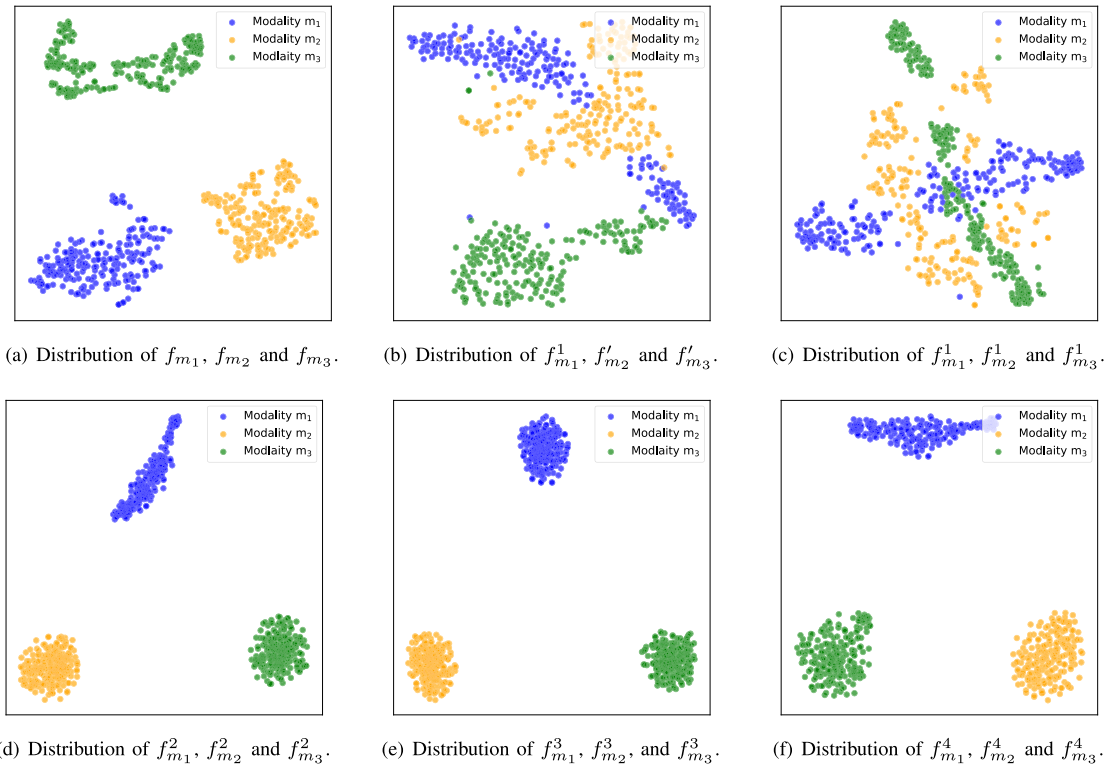
Features $\{f_m\}$ are inputs of the DMS. Feature $f_{m_1}^1$, feature f'_{m_2} and feature f'_{m_3} are outputs of the DMS. Comparison between Fig. 4(a) and (b) illustrates that all three features are driven closer to modality m_1. It certifies that the DMS performs the modality supplementary with modality m_1.

Features $\{f_m^1\}$ are outputs of the CME operation. As shown in Fig. 4(c), features of three modalities appear in a type of interactive distribution, and multimodal features mostly overlap in the figure. It demonstrates that the CME operation can learn the common representation for different modalities.

Features $\{f_m^2\}$ are outputs of the SDE operation, features $\{f_m^3\}$ are outputs of the DDE operation, and features $\{f_m^4\}$ are outputs of SPME module. Fig. 4(d), (e), and (f) illustrate the feature distributions of three modalities that are separated by the SPME operation. The results certify that our method can extract modality-specific information.

D. Model Complexity Analysis

To conduct a comprehensive analysis of the proposed model but not just accuracy, we evaluate the number of parameters for the proposed method alongside two state-of-the-art approaches, as detailed in Table IX. As we can see, our algorithm has a parameter value of 109.9 M, while the MMIM [14] algorithm, considered as SOTA, has 109.8 M. It is evident that the parameter quantity for MMIM [14] and our SIMSUF are practically similar, suggesting a negligible difference in model complexity. We conduct experiments to estimate the Floating Point Operations per Second (FLOPs) of our approach. We further estimate the FLOPs for related models from the literature and compare them with our SIMSUF model. The results reveal that our FLOPs count is a little higher than existing SOTA models, primarily due to the utilization of more complex architectures like Transformers. Compared to MMIM [14], our FLOPs count exhibits a marginal increase, 1.7 M higher than the MMIM, but the result of our approach is relatively increased, which reaffirms the effectiveness of our model.



(a) Distribution of f_{m_1}, f_{m_2} and f_{m_3}. (b) Distribution of $f^1_{m_1}$, f'_{m_2} and f'_{m_3}. (c) Distribution of $f^1_{m_1}$, $f^1_{m_2}$ and $f^1_{m_3}$.

(d) Distribution of $f^2_{m_1}$, $f^2_{m_2}$ and $f^2_{m_3}$. (e) Distribution of $f^3_{m_1}$, $f^3_{m_2}$, and $f^3_{m_3}$. (f) Distribution of $f^4_{m_1}$, $f^4_{m_2}$ and $f^4_{m_3}$.

Fig. 4. Visualization results of $\{f_m\}$, $\{f^1_m\}$, $\{f^2_m\}$, $\{f^3_m\}$, $\{f^4_m\}$, f'_{m_2} and f'_{m_3}, $m \in \{m_1, m_2, m_3\}$. Different colors represent different modalities. The features $\{f_m\}$ refer to the inputs of the DMS. $f^1_{m_1}$, f'_{m_2} and f'_{m_3} are the outputs of the DMS. Features $\{f^1_m\}$ denote the outputs of the CME operation. Features $\{f^2_m\}$ denote the outputs of the SDE operation. Features $\{f^3_m\}$ are outputs of the DDE operation. Features $\{f^4_m\}$ refer to the outputs of the SPME operation. Comparisons between (a), (b), (c), (d), (e), and (f) indicate that each part of the model gives its contribution to effective feature learning.

VI. Conclusion

In this paper, we proposed a dominant SIngle-Modal SUpplementary Fusion (SIMSUF) approach for multimodal sentiment analysis. The SIMSUF included three major contributions, the dominant modality supplementary module, the modality enhancement module, and the dominant-modal driven multimodal fusion module. Our model performed the multimodal supplementary, enhancement and fusion with a dominant modality for multimodal sentiment analysis. The proposed SIMSUF was evaluated in two commonly used datasets, CMU-MOSEI and CMU-MOSI. Extensive experiments and ablation studies demonstrated the effectiveness of our approach. By analyzing the results, we find that the representations of text modality are concise accurate, and well-quantified. The results illustrate that the processing of extracting information and generating feature vectors of the text modality is more effective compared to other modalities. In our future research, we will focus on extracting reliable and compact representations of all modalities.

References

[1] J. Arevalo, T. Solorio, M. M. Y. Gómez, and F. González, "Gated multimodal units for information fusion," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–17.

[2] J. Barnes, R. Kurtz, S. Oepen, L. Ovrelid, and E. Velldal, "Structured sentiment analysis as dependency graph parsing," in *Proc. 59th Ann. Meeting Assoc. Comput. Linguistics, 11th Int., Joint Conf., Natural Lang. Process.*, 2021, pp. 3387–3402.

[3] M. Belghazi et al., "Mutual information neural estimation," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, vol. 80, pp. 531–540.

[4] F. Chen, Z. Luo, and Y. Xu, "Complementary fusion of multi-features and multi-modalities in sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 1–18.

[5] M. Chen et al., "Multimodal sentiment analysis with word-level fusion and reinforcement learning," in *Proc. Int. Conf. Multimodal Interact.*, 2017, pp. 163–171.

[6] H. Deng, Z. Yang, T. Hao, Q. Li, and W. Liu, "Multimodal affective computing with dense fusion transformer for inter- and intra-modality interactions," *IEEE Trans. Multimedia*, vol. 25, pp. 6575–6587, 2023.

[7] J. Deng and C. Leung, "Towards learning a joint representation from transformer in multimodal emotion recognition," in *Proc. Int. Conf. Brain Informat.*, 2021, pp. 179–188.

[8] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Int. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2019, pp. 4171–4186.

[9] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–22.

[10] S. D'Mello and J. Westlund, "A review and meta-analysis of multimodal affect detection systems," *ACM Comput. Surv.*, vol. 47, pp. 1–36, 2015.

[11] Y. Feng et al., "Visible-infrared person re-identification via cross-modality interaction transformer," *IEEE Trans. Multimedia*, vol. 25, pp. 7647–7659, 2023.

[12] G. Paraskevopoulos, E. Georgiou, and A. Potamianos, "Mmlatch: Bottom-up top-down fusion for multimodal sentiment analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 4573–4577.

[13] W. Han et al., "Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis," in *Proc. Int. Conf. Multimodal Interact.*, 2021, pp. 6–15.

[14] W. Han, H. Chen, and S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 9180–9192.

[15] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, "Icon: Interactive conversational memory network for multimodal emotion detection," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 2594–2604.

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

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, pp. 1735–1780, 1997.

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

[19] Z. Huasong et al., "Self-adaptive neural module transformer for visual question answering," *IEEE Trans. Multimedia*, vol. 23, pp. 1264–1273, 2021.

[20] Z. Jia et al., "HeteMotionnet: Two-stream heterogeneous graph recurrent neural network for multi-modal emotion recognition," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 1047–1056.

[21] D. Klindt et al., "Towards nonlinear disentanglement in natural data with temporal sparse coding," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–16.

[22] Y. Lal, V. Kumar, M. Dhar, M. Shrivastava, and P. Koehn, "De-mixing sentiment from code-mixed text," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics: Student Res. Workshop*, 2019, pp. 371–377.

[23] X. Li et al., "Image-to-image translation via hierarchical style disentanglement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8635–8644.

[24] Y. Ling, J. Yu, and R. Xia, "Vision-language pre-training for multimodal aspect-based sentiment analysis," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistic*, 2022, pp. 2149–2159.

[25] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.

[26] Z. Liu et al., "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2247–2256.

[27] Z. Luo, H. Xu, and F. Chen, "Audio sentiment analysis by heterogeneous signal features learned from utterance-based parallel neural network," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 80–87.

[28] F. Lv, X. Chen, Y. Huang, L. Duan, and G. Lin, "Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2554–2562.

[29] L. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

[30] S. Mai, H. Hu, and S. Xing, "Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 164–172.

[31] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowl. Based Syst.*, vol. 161, pp. 124–133, 2018.

[32] H. Mao et al., "M-SENA: An integrated platform for multimodal sentiment analysis," in *Proc. 60th Annu. Meet. Assoc. Comput. Linguistics*, 2022, pp. 204–213.

[33] H. Pham, P. Liang, T. Manzini, L. Morency, and B. Póczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 6892–6899.

[34] S. Poria, E. Cambria, and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 2539–2544.

[35] S. Poria et al., "Context-dependent sentiment analysis in user-generated videos," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 873–883.

[36] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional MKL based multimodal emotion recognition and sentiment analysis," in *Proc. IEEE Int. Conf. Data Mining*, 2016, pp. 439–448.

[37] W. Rahman et al., "Integrating multimodal information in large pretrained transformers," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 2359–2369.

[38] D. She et al., "WSCNet: Weakly supervised coupled networks for visual sentiment classification and detection," *IEEE Trans. Multimedia*, vol. 22, pp. 1358–1371, 2020.

[39] E. Shutova, D. Kiela, and J. Maillard, "Black holes and white rabbits: Metaphor identification with visual features," in *Proc. Int. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2016, pp. 160–170.

[40] Z. Sun, P. Sarma, W. Sethares, and Y. Liang, "Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 8992–8999.

[41] J. Tang et al., "CTFN: Hierarchical learning for multimodal sentiment analysis using coupled-translation fusion network," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics, 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 5301–5311.

[42] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 6000–6010.

[43] D. Wang et al., "Cross-modal enhancement network for multimodal sentiment analysis," *IEEE Trans. Multimedia*, vol. 25, pp. 4909–4921, 2023.

[44] S. Wang, J. Zhang, N. Lin, and C. Zong, "Probing brain activation patterns by dissociating semantics and syntax in sentences," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 9201–9208.

[45] Z. Wang, Z. Wan, and X. Wan, "Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis," in *Proc. Int. Conf. World Wide Web*, 2020, pp. 2514–2520.

[46] Y. Wu et al., "A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis," in *Proc. Findings Assoc. Comput. Linguistics*, 2021, pp. 4730–4738.

[47] Y. Wu et al., "Sentiment word aware multimodal refinement for multimodal sentiment analysis with ASR errors," in *Proc. Findings Assoc. Comput. Linguistics*, 2022, pp. 1397–1406.

[48] L. Xu, Z. Wang, B. Wu, and S. Lui, "MDAN: Multi-level dependent attention network for visual emotion analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9469–9478.

[49] K. Yang, H. Xu, and K. Gao, "CM-BERT: Cross-modal bert for text-audio sentiment analysis," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 521–528.

[50] X. Yang, S. Feng, Y. Zhang, and D. Wang, "Multimodal sentiment detection based on multi-channel graph neural networks," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics, 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 328–339.

[51] 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.*, 2021, pp. 10790–10797.

[52] Z. Yuan, W. Li, H. Xu, and W. Yu, "Transformer-based feature reconstruction network for robust multimodal sentiment analysis," in *Proc. ACM Int. Conf. Multimedia*, 2021, pp. 4400–4407.

[53] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1103–1114.

[54] A. Zadeh, P. Liang, S. Poria, E. Cambria, and L. 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.

[55] A. Zadeh et al., "Multi-attention recurrent network for human communication comprehension," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 5642–5649.

[56] A. Zadeh, R. Zellers, E. Pincus, and L. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intell. Syst.*, vol. 31, no. 6, pp. 82–88, Nov./Dec. 2016.

[57] W. Zellinger, T. Grubinger, E. D. Lughofer, T. Natschläger, and S. Platz, "Central moment discrepancy (CMD) for domain-invariant representation learning," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–13.

[58] H. Zhang and M. Xu, "Weakly supervised emotion intensity prediction for recognition of emotions in images," *IEEE Trans. Multimedia*, vol. 23, pp. 2033–2044, 2021.

[59] X. Zhou, X. Song, H. Wu, J. Zhang, and X. Xu, "MAVT-FG: Multimodal audio-visual transformer for weakly-supervised fine-grained recognition," in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 3811–3819.

[60] T. Zhu, L. Li, J. Yang, S. Zhao, and X. Xiao, "Multimodal emotion classification with multi-level semantic reasoning network," *IEEE Trans. Multimedia*, vol. 25, pp. 6868–6880, 2023.