Multi-View Interactive Representations for Multimodal Sentiment Analysis

Zemin Tang, Qi Xiao, Yunchuan Qin, Xu Zhou, Joey Tianyi Zhou, and Kenli Li, *Senior Member, IEEE*

***Abstract*—Multimodal Sentiment Analysis (MSA) technology, prevalent in consumer applications and mobile edge computing (MEC), enables sentiment examination through user data collected by smart devices. Despite the focus on representation learning in MSA, current methods often prioritize recognition performance through modality interaction and fusion. However, they struggle to capture multi-view sentiment cues across different interaction states, limiting multimodal sentiment representations' expressiveness. This paper develops an innovative MSA framework, MVIR, learning multi-view interactive representations in diverse interaction states. Multiple meticulously designed sentiment tasks and an introduced self-supervised label generation algorithm (SSLGM) enable a comprehensive understanding of multi-view sentiment tendencies. The dual-view attention weighted fusion (DVAWF) module is designed to facilitate inter-modality information exchange in different interaction states. Extensive experiments on three MSA datasets affirm the efficacy and superiority of MVIR, showcasing its ability to capture sentiment information from multimodal data across various interaction states.**

***Index Terms*—Representation learning, dual-view attention weighted fusion, multi-task learning, multimodal sentiment analysis.**

I. Introduction

With the proliferation of mobile Internet and smart devices, consumers have access to a wide range of applications for shopping, communication, and entertainment. In recent years, the advancements in artificial intelligence, particularly deep learning, have facilitated enhanced products and more efficient services for consumers, thanks to mobile edge computing (MEC). One notable technology, multimodal sentiment analysis (MSA), has gained popularity. It analyzes consumers' sentiments and opinions regarding their user experiences using diverse data collected from smart devices [1], [2], [3], [4], [5], [6]. For example, by leveraging speech recognition and sentiment analysis, the system can automatically tailor music, videos, or advertisements to align with the user's sentiment, providing a personalized experience that resonates with their emotional state [2].

In the realm of MSA, effectively capturing sentiment information to learn meaningful representations poses a notable challenge due to the modality gaps persisting between heterogeneous modalities [7], [8], [9], [10], [11], [12], [13]. Existing approaches mostly address this challenge by employing sophisticated interaction and fusion mechanisms to integrate information from different modalities [7], [8], [10]. These methods include attention-based [14], [15], tensor-based [16], [17], and translation-based techniques [8], [18], among others. While these methods have made significant progress, they may fall short in capturing the multi-view shared and private sentiment cues of modalities in various interaction states. Consequently, subtle yet valuable sentiment cues from the speaker might be overlooked, thereby weakening the expressive capacity of multimodal sentiment representations.

Recent progress in multimodal learning has brought attention to the existence of shared-private information across different modalities, which holds substantial potential for enhancing the development of advanced multimodal representations [19], [20]. Shared information refers to features or representations that are common across different modalities. Through shared information, mutual complementarity and interaction between different modalities are facilitated. Private information refers to features or representations that are unique to each modality and can provide additional context or supplementary information to other modalities, thereby enhancing complementarity between modalities. By exploring these two types of information in modalities, effective multimodal representations can be learned [9], [21].

Motivation example: Figure 1 represents a motivational example, illustrating how text-image data express the shared and private information in different interaction states to collaboratively convey consumers' emotional states through their evaluations of food. The image represents a photo taken by the consumer, accompanied by a textual comment that reads: "This dish looks absolutely delicious, with great colors, aroma! However, in reality, it's disappointing as it's too salty and hard to accept!" The label for this example is negative.

In this example, the shared information conveyed by the image includes the appealing appearance and visual attractiveness of the food, while the private information includes

Manuscript received 30 June 2023; revised 4 November 2023 and 7 December 2023; accepted 18 January 2024. Date of publication 23 January 2024; date of current version 26 April 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2021YFF0901001; in part by the Natural Science Foundation for Regional Innovative Development under Grant U23A20317; in part by NSFC under Grant 62102143, Grant 62172146, and Grant 62172157; and in part by the Natural Science Foundation of Hunan Province under Grant 2023JJ30083 and Grant 2023JJ10016. *(Corresponding authors: Qi Xiao; Yunchuan Qin.)*

Zemin Tang, Qi Xiao, Yunchuan Qin, Xu Zhou, and Kenli Li are with the College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, Hunan, China (e-mail: tzm123@hnu.edu.cn; xiaoqi0909@hnu.edu.cn; qinyunchuan@hnu.edu.cn; zhxu@hnu.edu.cn; lkl@hnu.edu.cn).

Joey Tianyi Zhou is with the Centre for Frontier AI Research, Agency for Science, Technology and Research, Singapore 138632 (e-mail: zhouty@ihpc.a-star.edu.sg).

Digital Object Identifier 10.1109/TCE.2024.3357480

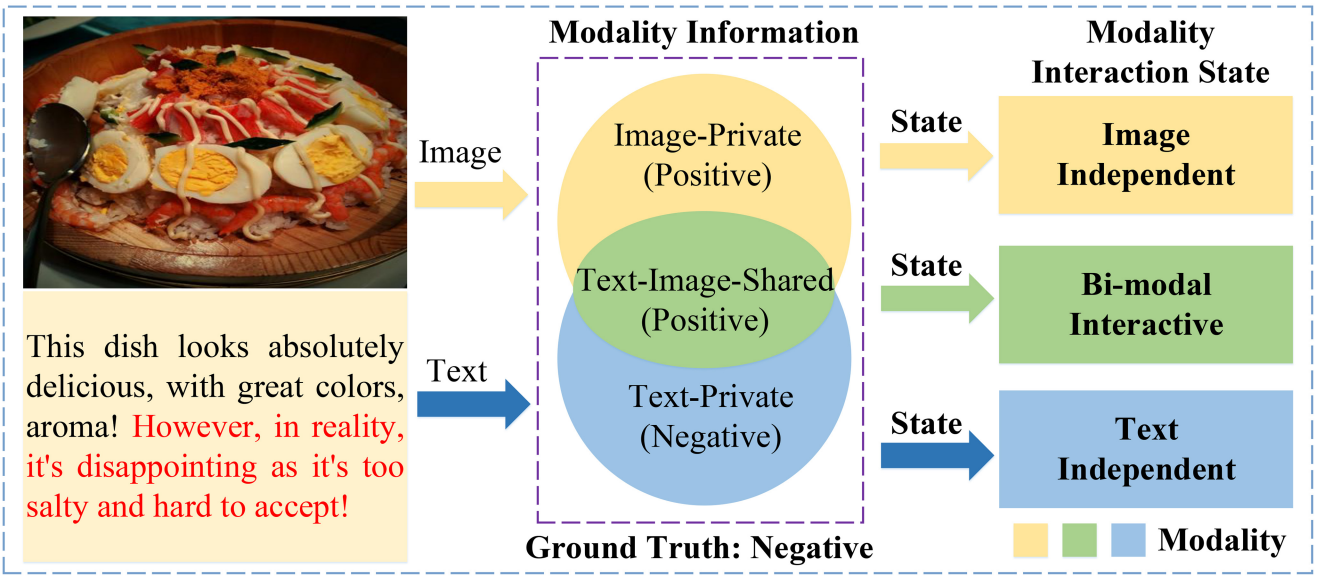


Fig. 1. An example that demonstrates how modality data can portray multi-view shared and private sentiment information across different interaction states.

specific composition details and visual aspects such as colors and lighting. The shared information conveyed by the text is the descriptive part of the comment, such as "This dish looks absolutely delicious, with great colors, aroma!" These descriptive words convey the consumer's positive evaluation of the food's appearance and expected taste. The private information includes the expression of disappointment and unacceptability, as stated in the comment: "However, in reality, it's disappointing as it's too salty and hard to accept!" When we consider the image and text separately, in a state where there is no interaction between the image and text modalities, the data mainly reflect their respective private information. We can speculate based on the food content in the image that it conveys the consumer's positive emotions, while the text comment expresses a negative sentiment.

However, when we combine the image and text, they interact synergistically. Their shared information is the evaluation of the food's appearance, which is positive. Based on their shared and private information, we can observe that the consumer has high expectations for the food, but its poor taste highlights their dissatisfaction and negativity. From the analysis of this example, it becomes evident that the shared and private information between different modalities is a reflection of the data's emotions from different perspectives. They play important roles in overall sentiment analysis, and the type of information they present is related to their interaction state.

Multi-view shared-private information extraction for tri-modal sentiment analysis: Compared to the example shown in Figure 1, the representation learning task faced in this paper is more complex. In this work, we focus on sentiment analysis in the text, audio, and video modalities [9], [21], which involve more sophisticated interaction states, including three uni-modal, three bi-modal, and one tri-modal interactions. Therefore, understanding how to effectively utilize the shared and private information in these complicated multi-view interactions to acquire expressive multimodal representations constitutes a significant challenge and represents the primary concern that needs to be addressed in our work.

We attribute this challenge to two key aspects: one pertains to the exploration and integration of multi-modal sentiment cues across diverse interaction states, while the other involves the facilitation of effective information exchange in cross-modal interactions. To this end, we propose a novel MSA framework for learning multi-view interactive representations (MVIR). As shown in Figure 2, MVIR is achieved by employing the multi-task mechanism [22], [23] to assign a separate task for each interaction state. In addition, a self-supervised label generation algorithm (SSLGM) [24] is introduced to generate state-specific labels for the representation in each interaction state. This approach facilitates the optimization of the model's multimodal common representations of various tasks with respect to the information present in different states, thereby capturing multi-view shared-private information.

Furthermore, to enhance cross-modal information exchange, we propose a novel modality interaction algorithm called dual-view attention weighted fusion (DVAWF). This algorithm utilizes graph attention networks [25] and self-attention mechanisms [26] to promote cross-modal information interaction and generate well-fused features through dynamic weighted fusion. We conduct extensive experiments on multiple public datasets, including [27], [28], and [29]. The results of these experiments provide strong evidence supporting the effectiveness of our proposed method.

The specific contributions of this study are as follows:

- We present a novel framework for MSA that focuses on learning multi-view interactive representations (MVIR) across different interaction states. This framework allows for a comprehensive understanding of sentiment information from various perspectives, enhancing the overall analysis process.
- To facilitate the exchange of information across different modalities, we propose a novel approach called the dual-view attention weighted fusion (DVAWF) algorithm.

This algorithm is designed to promote the integration of cross-modal information, thereby improving the overall effectiveness of multimodal sentiment analysis.

- The experimental results obtained from multiple datasets provide compelling evidence of the significant improvement achieved in MSA through the utilization of MVIR. These results serve to affirm the effectiveness and efficacy of our approach.

II. Related Work

A. Multi-View Learning

Multi-view learning (MVL) is a prominent machine learning approach that utilizes information from multiple perspectives or modalities to enhance predictive performance and gain a deeper understanding of complex data [30]. In many real-world scenarios, such as image recognition, bioinformatics, and social network analysis, relying solely on a single view of the data may not provide sufficient information for accurate analysis and prediction [31], [32]. The core concept of MVL is to leverage the complementary information present in different views, leading to improved learning outcomes compared to traditional single-view methods. By incorporating multiple views, we can capture diverse aspects, uncover hidden patterns, and overcome the limitations of individual views [30]. In the field of MVL, various strategies are employed. These range from early fusion techniques that merge multiple views into a single representation to late fusion methods that combine predictions from individual views [33].

B. Multimodal Sentiment Analysis

Multimodal sentiment analysis (MSA) has gained significant attention in recent years, aiming to understand people's sentiments by leveraging diverse data sources such as text, visual, and acoustic data [7], [8], [9], [10], [11]. Previous research efforts have primarily focused on integrating cross-modal information into effective multimodal representations through the design of sophisticated feature interaction and fusion modules for sentiment recognition [7], [8], [10]. These methods can be further categorized based on their interaction mechanisms, including attention-based [14], [15], tensor-based [16], [17], and translation-based methods [8], [18]. Attention-based methods employ various attention mechanisms to facilitate feature interactions within and across modalities [14], [15]. Tensor-based approaches extract utterance-level feature vectors from each modality and fuse them using the outer product of vectors [16], [17]. Translation-based methods leverage the concept of machine translation to establish robust associations across multiple modalities by translating from a source modality to a target modality [8], [18].

Recent research has highlighted the presence of shared and private information in multimodal data, which contain consistent and distinct sentiment clues between modalities, respectively [9], [21]. Explicitly mining these two types of information can significantly enhance the multimodal feature representations and improve model performance. For instance, [21] employs two separate encoders to project modalities into modality-invariant and modality-specific spaces to capture shared and private information. In another study, [9] introduces cross-modal prediction tasks to learn shared and private semantics for non-textual modalities. However, these methods can only capture a limited number of sentiment clues in a few interaction states, making it challenging to achieve a comprehensive understanding of sentiment information in complex interaction scenarios. Furthermore, their simplistic modal fusion algorithms, such as concatenation, fail to adequately facilitate cross-modal information exchange. These limitations hinder the further improvement of multimodal feature representations in the model.

In our work, we focus on learning well-expressed multimodal features as the core aspect for improving sentiment analysis performance. In one hand, current methods struggle to capture multi-view sentiment cues across different interaction states, limiting multimodal sentiment representations' expressiveness [8], [14], [15], [16], [17], [18]. We develop the innovative MVIR framework, which, through multiple meticulously designed sentiment recognition tasks related to interaction states and the introduced SSLGM algorithm, provides more comprehensive sentiment cues compared to many existing MSA algorithms. On the other hand, we propose a novel modality interaction and fusion mechanism, DVAWF, which effectively integrates modal information from dual perspectives and dynamically fuses them to learn effective sentiment representations. These innovations and uniqueness set us apart from existing works, providing a novel approach to effectively recognize sentiment.

III. Methodology

In this section, we provide a detailed explanation of our proposed MSA framework called MVIR. We start by introducing the task setup, followed by an overview of the overall framework. We then delve into the specifics of our designed DVAWF algorithm and the introduced SSLGM. Finally, we summarize the learning process of MVIR in the form of an algorithm.

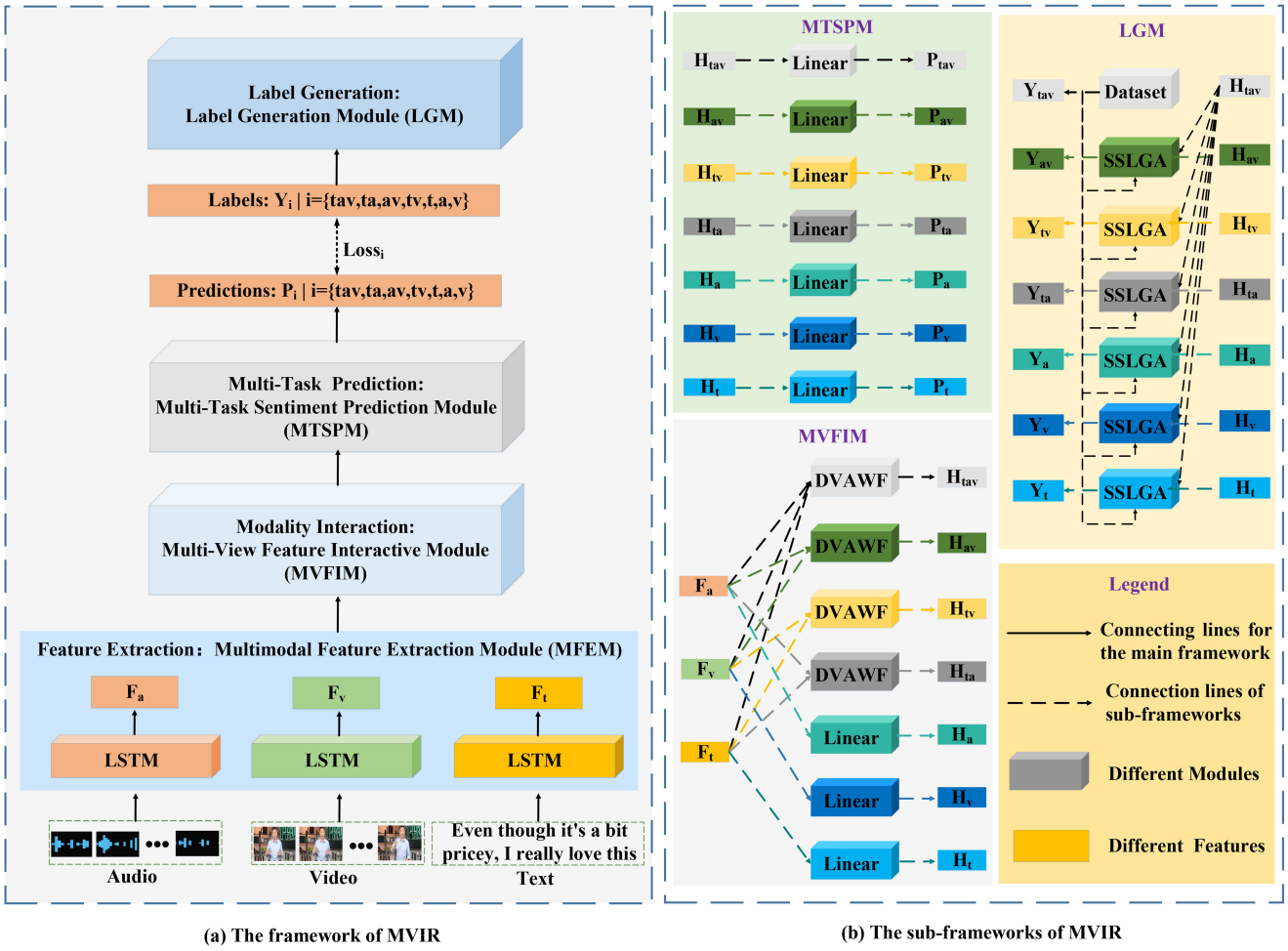
A. Task Setup

MSA aims to analyze the sentiments hidden in various modalities [7], [8], [9], [10], [11]. We consider three modalities as the input data: text data I_t, audio data I_a, and visual data I_v. The model predicts a sentiment result $P_{tav} \in \mathbb{R}$.

In addition, our proposed MVIR framework includes three uni-modal subtasks and three bi-modal subtasks. These subtasks aim to capture private information in the uni-modal independent states and shared information in the bi-modal interactive states, respectively. The auxiliary tasks output predictions $P_i \in \mathbb{R}$, where $i \in \{t, a, v, ta, tv, av\}$ corresponds to the specific uni-modal or bi-modal task. However, during the testing stage, we only use the prediction P_{tav} as the final result for sentiment analysis.

B. Overall Architecture

The overall framework of our MVIR is depicted in Figure 2. To capture the multi-view shared and private



(a) The framework of MVIR (b) The sub-frameworks of MVIR

Fig. 2. The overall framework (left) and sub modules (right) of our MVIR. MVIR consists of seven tasks, including one tri-modal task (indicated by subscript tav), three bi-modal tasks (indicated by subscript ta, tv, av), and three uni-modal tasks (indicated by subscript t, a, v). The tri-modal task serves as the main task, while the other tasks act as auxiliary tasks. During the test phase, the prediction results of the tri-modal task (P_{tav}) are utilized. The sentiment labels of the main task (Y_{tav}) are provided by the dataset, while the labels of the other tasks are obtained through the SSLGA approach corresponding to each task.

information in different interaction states of MSA, we design seven tasks to simulate the seven distinct interactive states, including uni-modal, bi-modal, and tri-modal states.

MVIR consists of four main components: feature extraction, modality interaction, multi-task prediction, and label generation. Through the mechanism of multi-task learning, sentiment information from different interaction states can be jointly optimized to enhance the representation of multimodal features. This approach allows us to capture sentiment cues from multiple perspectives and learn effective multimodal sentiment representations. Note that the tri-modal prediction task serves as the primary perspective employed by this model for sentiment analysis.

Feature Extraction. First, we extract the multimodal features from three input modality data I_t, I_a and I_v through the multimodal feature extraction module (MFEM). For the feature extraction of the audio and vision modalities, we adopt a similar preprocessing approach as previous work [8], which will be explained in detail in the experimental section. Specifically, we preprocess the audio and vision data to obtain the original data.

After preprocessing, we acquire sequential features from both vision and audio data. Long Short-Term Memory (LSTM) [34], a key technology in deep learning, excels in modeling sequential data by effectively capturing long-term dependencies, enhancing tasks like language processing and time series analysis. In this research, we utilize a single directional LSTM (sLSTM) to model the temporal information of sentiments in both modalities. The end-state hidden vectors generated by the sLSTM model serve as the representations of the entire sequences. We set the feature dimension d to be the same for both audio and vision modalities, F_a and F_v are the extracted audio and vision features by sLSTM.

$$F_a = \text{sLSTM}(I_a) \in R^d, \tag{1}$$

$$F_v = \text{sLSTM}(I_v) \in R^d. \tag{2}$$

In its capacity as a pre-trained language model, BERT exhibits robust semantic comprehension capabilities. It is adept at converting textual data into high-quality embedding representations, thereby furnishing profound insights into the deep semantic information contained within the text. Thus, for text feature extraction, we utilize the pre-trained 12-layer BERT model [35] to extract sentence representations. Subsequently,

we employ a linear layer to unify the feature dimensions of the text with the other modalities, and use an averaging function to obtain the global features F_t of text modality.

$$F_t = \text{Average}(\text{Linear}(\text{BERT}(I_t))) \in R^d. \tag{3}$$

Modality Interaction. As shown in Figure 2 (a) and (b), in this step we develop a multi-view feature interaction module (MVFIM) to extract sentiment features from multiple perspectives during different interaction states. In the MVFIM, we design a novel approach called DVAWF for cross-modal interaction in the tri-modal and bi-modal tasks. For the three uni-modal tasks, linear layers are utilized to extract deeper features:

$$\begin{aligned}
H_t &= \text{Linear}_t(F_t) \in R^d, \\
H_a &= \text{Linear}_a(F_a) \in R^d, \\
H_v &= \text{Linear}_v(F_v) \in R^d, \\
H_{ta} &= \text{DVAWF}_{ta}(F_t, F_a) \in R^d, \\
H_{tv} &= \text{DVAWF}_{tv}(F_t, F_v) \in R^d, \\
H_{av} &= \text{DVAWF}_{av}(F_a, F_v) \in R^d, \\
H_{tav} &= \text{DVAWF}_{tav}(F_t, F_a, F_v) \in R^d,
\end{aligned} \tag{4}$$

where H_i represents the learnt features of task i, $i \in \{tav, ta, tv, av, t, a, v\}$.

Multi-Task Prediction. As illustrated in Figure 2 (a) and (b), following the MVFIM, we pass the learned features to the multi-task sentiment prediction module (MTSPM) for generating the final predictions, denoted as P_i, for each task. This is accomplished using linear layers that leverage the features acquired from the previous module:

$$P_i = \text{Linear}_i(H_i) \in R, \tag{5}$$

where $i \in \{tav, ta, tv, av, t, a, v\}$.

Label Generation. Since most popular MSA datasets only provide labels that represent multimodal sentiments (i.e., Y_{tav} in this paper), it is challenging to learn sentiment knowledge in other states. Therefore, we design a label generation module (LGM) that introduce SSLGM [24] to construct labels for bi-modal interactive states and uni-modal independent states, which utilizes a self-supervised learning strategy. The SSLGM generates pseudo labels Y_i, $i \in \{ta, tv, av, t, a, v\}$ for each task, using the features and labels from the tri-modal interactive state, along with the features of the corresponding task i, as shown in Figure 2 (b):

$$Y_i = \text{SSLGM}(H_{tav}, Y_{tav}, H_i). \tag{6}$$

Optimization Objectives. The optimization objective comprises seven task regression losses, which are optimized using the smooth L1 loss [36]. For bi-modal and uni-modal tasks, we calculate the weight of each task using the difference between the dataset labels and the pseudo labels generated by the auxiliary task. This weight encourages the model to focus more on the labels of larger differences, thereby facilitating the mining of sentiment knowledge from different interactive states that cannot be provided by the dataset labels:

$$\begin{aligned}
Loss &= 1/N \sum_{i}^{N} \sum_{j} Loss_j \\
&= 1/N \sum_{i}^{N} \left(\text{SL}\left(P_{tav}^i, Y_{tav}^i\right) + \sum_{s}^{\{t,a,v,ta,tv,av\}} W_s^i * \text{SL}\left(P_s^i, Y_s^i\right) \right),
\end{aligned} \tag{7}$$

where $W_s^i = \tanh|Y_{tav}^i - Y_s^i|$ is the weight of i_{th} sample for auxiliary task s, $\text{SL}(\cdot)$ represents smooth L1 loss [36], $j \in \{t, a, v, ta, tv, av, tav\}$.

C. DVAWF

The DVAWF plays a crucial role in facilitating knowledge transfer between different modalities in tri-modal and bi-modal interactive states. This module constructs cross-modal information exchange through multi-head attention (MHA) [26] and multi-head graph attention network (MHG) [25]. It enables feature interaction and fusion from these two branches separately, and learns robust dual-view attention features through dynamic weighting.

On one hand, we treat different modality features as nodes in MHG and employ the message passing mechanism to model the interactions between different modalities. On the other hand, we leverage the attention scores from MHA to determine the degree of mutual attention between modalities and update modal fusion features accordingly. Both mechanisms utilize a multi-head design, each head can focus on different contextual information and learn feature representations at different feature spaces, thereby improving model's representation capability. Figure 3 illustrates the framework of DVAWF in the tri-modal state task. It's worth noting that the DVAWF in the bi-modal state tasks shares the same framework as the tri-modal state task, with the only difference being the input of the respective task.

First, we construct a three-node graph $< V, E >$ to represent the modality features and their interactions. In this graph, the nodes $V = [V_t, V_a, V_v]$ correspond to the text, audio, and vision modalities, while the edges $E = [e_{ij}|i, j \in \{t, a, v\}]$ denote the weights of the connections between modalities. The features $F = [F_t, F_a, F_v]$ extracted from MFEM are considered as the features of the corresponding nodes in the graph.

Then, we calculate the weight of modality connections E:

$$e_{ij} = \text{LeakyReLU}\left(\overrightarrow{\mathbf{a}}^T \left[\mathbf{W}F_i || \mathbf{W}F_j\right]\right), \tag{8}$$

where $\text{LeakyReLU}(\cdot)$ represents the activate function LeakyReLU [37], $\overrightarrow{\mathbf{a}}$ represents the weight parameter of a linear layer, $||$ represents the concatenation operation, T represents transposition. W is a learnable shared parameter matrix initialized using a fully connected layer, with a matrix size of (D, D1), where D equals D1, which corresponds to d in to 4.

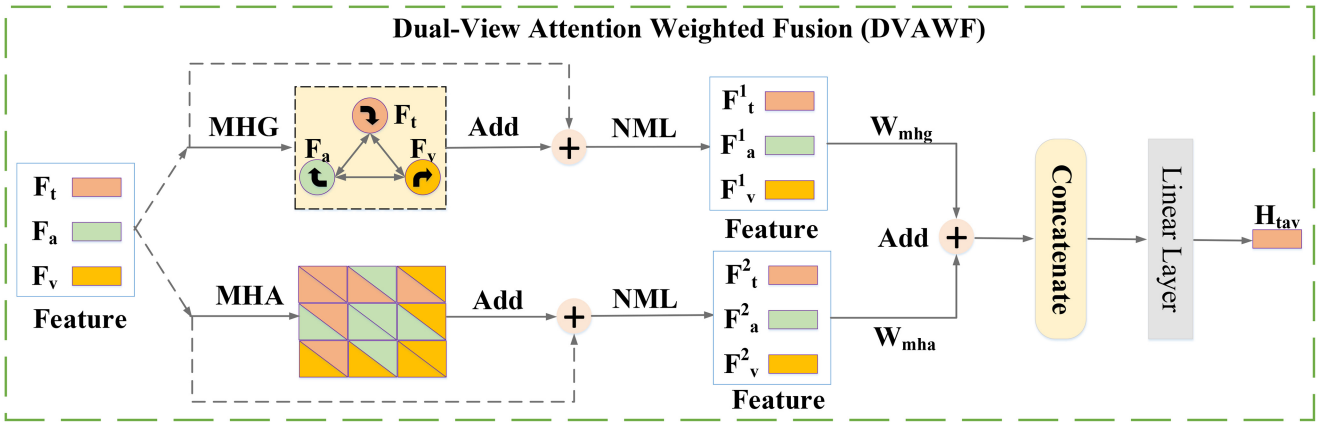


Fig. 3. The framework of DVAWF in tri-modal interactive state task.

Next, we use softmax to calculate the weight coefficients α_{ij} to make weight E easily comparable across different nodes:

$$\alpha_{ij} = \text{softmax}_j(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{m\in\mathcal{N}_i}\exp(e_{im})}, \tag{9}$$

where $\mathcal{N}_i$ is the neighbor nodes of modality node i.

Then, we perform multi-head attention to get the features of modality interactions $h^{mhg} = \{h_t^{mhg}, h_a^{mhg}, h_v^{mhg}\}$:

$$h^{mhg} = \frac{1}{K}\sum_{k=1}^{K}\sum_{j\in\mathcal{N}_i}\alpha_{ij}^k \mathbf{W}^k F_j, \tag{10}$$

where K is the number of attention head, $i \in \{t, a, v\}$.

For MHA, we map the multimodal features to query, key, and value in the self-attention mechanism. This allows us to calculate attention scores and perform information exchange between the modality features. We calculate the features of each head as:

$$\text{Head} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \tag{11}$$

where $Q = FW^Q$, $K = FW^Q$, $V = FW^Q$, $F = [F_t, F_a, F_v]$. Then, multi-head features h^{mha} can be learned as:

$$h^{mha} = \text{Concat}(\text{Head}_1, \ldots, \text{Head}_n)W^O. \tag{12}$$

After that, we adopt the structural design inspired by the Transformer model [26]. We utilize residual connections (Add) and layer normalization (NML) techniques to optimize the features F^1 and F^2 from MHG and MHA:

$$\begin{aligned} F^1 &= \text{NML}\left(F + h^{mhg}\right) \\ F^2 &= \text{NML}\left(F + h^{mha}\right). \end{aligned} \tag{13}$$

Next, we dynamically weight and fuse the features of two branches through learnable parameters W_{mhg}, W_{mha} to obtain the fusion feature F':

$$F' = W_{mhg} * F^1 + W_{mha} * F^2. \tag{14}$$

Finally, we concatenate features F' internally, scale the dimension of the concatenated feature through a linear layer to obtain the output feature H_{tav}:

$$H_{tav} = \text{Linear}\big(\text{Concat}(F')\big) \in R^d. \tag{15}$$

D. SSLGM

We introduce the SSLGM algorithm [24] to aid in the construction of sentiment labels for auxiliary tasks. The SSLGA is a self-supervised approach, which is designed based on two key insights. Firstly, considering that the features of each interaction state are derived from the shared three modalities, there should be a high correlation among the sentiment labels of different tasks [24]. In other words, the labels of other interactive states can be constructed based on the labels of the tri-modal interactive state provided by the datasets. Secondly, the label difference of different modality is positively correlated with the difference in distances between modality representations and category centers in each state [24]. The pseudo labels are learned through the following steps: relative distance calculation, shifting value calculation, and momentum-based label update

Relative Distance Calculation. This step calculates the relative distance RD_i from the modality representation to the positive center PC_i and the negative center NC_i in each modality interactive state.

First, we compute the two sentiment class centers in each state:

$$\begin{aligned} PC_i &= \frac{\sum_{j=1}^{N} I(Y_i(j) > 0)\cdot F_{ij}^g}{\sum_{j=1}^{N} I(Y_i(j) > 0)}, \\ NC_i &= \frac{\sum_{j=1}^{N} I(Y_i(j) < 0)\cdot F_{ij}^g}{\sum_{j=1}^{N} I(Y_i(j) < 0)}, \end{aligned} \tag{16}$$

where $i \in \{tav, ta, tv, av, t, a, v\}$ represents different modality interactive states, j represents the index of samples, $I(\cdot)$ is an indicator function, F_{ij}^g and $Y_i(j)$ is the global representation and labels of the j_{th} sample in modality state i respectively. Note that F_{ij}^g is used to store multi-view feature in formula (4) that initialized as zero and $Y_i(j)$ is initialized as the labels of tri-modal interactive state that offered by datasets, at the beginning of training.

Next, we calculate the absolute distance PD_i and ND_i, from modality representation to the positive center PC_i and the negative center NC_i in each state:

$$PD_i = \frac{||H_i - PC_i||_2^2}{\sqrt{d}},$$

TABLE I
DATASET STATISTICS AND TRAINING PARAMETERS IN MOSI, MOSEI, AND SIMS

Description	CMU-MOSI	CMU-MOSEI	SIMS
Number of utterances	2199	22856	2281
Data splits (train/dev/test)	1284/229/686	16326/1871/4659	1368/456/457
Batchsize	32	32	32
Epoch	Early stopping	Early stopping	Early stopping
Hidden size	300	300	300
Number of attention head in DVAWF	12	11	7
Learning rate (BERT/a-lstm/v-lstm/others)	5e-5/1e-3/1e-4/1e-3	5e-5/5e-3/1e-4/1e-3	5e-5/1e-4/1e-4/1e-4

$$ND_i = \frac{||H_i - NC_i||_2^2}{\sqrt{d}}, \tag{17}$$

where H_i is the feature of formula (4), $||\cdot||_2^2$ represents L2 normalization [38], and d is the feature dimension.

Finally, we can obtain the relative distance RD_i in each state:

$$RD_i = \frac{ND_i - PD_i}{PD_i + \epsilon}, \tag{18}$$

where ϵ is a small number to prevent zero exceptions.

Shifting Value Calculation. This step is responsible for calculating the sentiment label shift value of different modality states, i.e., the label value difference between the tri-modal labels and other states. The shift value calculation using relative distance RD_i based on the following two relationships:

$$\begin{aligned} &\frac{Y_i}{Y_{tav}} \propto \frac{P_i}{P_{tav}} \propto \frac{RD_i}{RD_{tav}} \Rightarrow Y_i = \frac{RD_i * Y_{tav}}{RD_{tav}}, \\ &Y_i - Y_{tav} \propto P_i - P_{tav} \propto RD_i - RD_{tav} \\ &\Rightarrow Y_i = Y_{tav} + RD_i - RD_{tav}. \end{aligned} \tag{19}$$

We can get the labels of other states by equal-weight summation:

$$\begin{aligned} Y_i &= \frac{Y_{tav} * RD_i}{2RD_{tav}} + \frac{Y_{tav} + RD_i - RD_{tav}}{2} \\ &= Y_{tav} + \frac{RD_i - RD_{tav}}{2} * \frac{Y_{tav} + RD_{tav}}{RD_{tav}} \\ &= Y_{tav} + \delta_{shift}, \end{aligned} \tag{20}$$

where $i \in \{ta, tv, av, t, a, v\}$, $\delta_{shift} = \frac{RD_i - RD_{tav}}{2} * \frac{Y_{tav} + RD_{tav}}{RD_{tav}}$ represents the offset value of other modality state labels to tri-modal modality state labels.

Momentum-based Label Update. To allow the model to converge as quickly as possible, a momentum based label update policy is used to mitigate the adverse effects:

$$Y_i^{(j)} = \begin{cases} Y_{tav} & j = 1 \\ \frac{j-1}{j+1} Y_i^{(j-1)} + \frac{2}{j+1} Y_i^j & j > 1, \end{cases} \tag{21}$$

where j is the index of training epoch, $Y_i^{(j)}$ is the generated pseudo labels using momentum based label update policy. This strategy gives a dynamically increased weight to the historical value of the pseudo labels. With the increase of training epochs, the pseudo label values of each sample in different modality states will gradually tend to stabilize.

Differing from the unimodal label generation module (ULGM) in [24], our SSLGM generates corresponding sentiment labels for both bimodal and unimodal data, enabling the capture of more comprehensive and nuanced emotional cues. Specifically, the labels generated for the unimodal data can learn private information without modality interaction, whereas the labels generated for the bimodal data can learn shared information in interactive states.

Algorithm 1 Multi-View Interactive Representations (MVIR)

Input:
Training set $X = \{I_t, I_a, I_v\}$, batch size b, number of iterations E, learning rate η, task type $T \in \{t, a, v, ta, tv, av, tav\}$, number of sample N.
Output:
Prediction P_{tav}, multi-modal representations H_{tav}.
Initialize model parameters: $\theta \leftarrow$ Random initialization
Initialize labels $Y_i | i \in \{t, a, v, ta, tv, av\}$ as Y_{tav}
Initialize features $F_{ij}^g | i \in \{t, a, v, ta, tv, av, tav\}, j = N$ as 0
for $e \in \{1, E\}$ **do**
  Sample a mini-batch $X_j | j \in \{1, \ldots, b\}$ with size b
  Extract modality features F^t, F^a, F^v in Equation (1)∼(3)
  if $T \in \{t, a, v\}$ **do**
    Obtain uni-modal features H_t, H_a, H_v through linear layers in Equation (4)
  end if
  if $T \in \{ta, tv, av, tav\}$ **do**
    Obtain bi-modal and tri-modal features H_{ta}, H_{av}, H_{tv}, H_{tav} through DVAWF in Equation (8∼15)
  end if
  Store features $H_{ij} | i \in \{t, a, v, ta, tv, av, tav\}, j \in \{1, \ldots, b\}$ to F_{ij}^g
  if $e > 1$ **do**
    Update labels $Y_i | i \in \{t, a, v, ta, tv, av\}$ using Equation (16∼21)
  end if
  Obtain predictions of all tasks using Equation (5)
  Compute loss using Equation (7)
  Compute gradient and update parameters $\theta = \theta - \eta \frac{\vartheta L}{\vartheta \theta}$
end for

E. Algorithm Summary

In this subsection, we present a summarized algorithmic form of MVIR to provide a clearer understanding of the representation learning and sentiment prediction process introduced in Sections III-B–III-D, i.e., Algorithm 1.

IV. EXPERIMENTS

A. Datasets

Our experiments are validated on three popular datasets, including CMU-MOSI [27], CMU-MOSEI [28] and SIMS [29]. We show the statistics of the three datasets and the according parameters of training models in Table I.

The following is a detailed introduction to relevant datasets.

- CMU-MOSI [27] consists of 93 Youtube movie review videos. It provides sentiment intensity annotations on a scale ranging from −3 (highly negative) to +3 (highly positive).

- CMU-MOSEI [28] is made for multimodal sentiment and emotion analysis. It is an extension of CMU-MOSI with a larger number of samples. In our study, we evaluate our model's performance on the sentiment analysis task.
- The SIMS dataset [29] is a Chinese benchmark for multimodal sentiment analysis. It provides detailed annotations for each modality, which is collected from various movies, TV series, and variety shows, capturing spontaneous expressions and diverse lighting conditions. Each sample in the dataset is assigned a sentiment score ranging from -1 (strongly negative) to $+1$ (strongly positive).

B. Parameter Settings

The training parameters are presented in Table I. The hidden size refers to the feature dimension of different modalities after extracting modality-specific features, denoted as d in formula (1∼15). The learning rate (LR) consists of four components: the LR for the BERT model [35] applied to the text modality, the LR for the sLSTM model [34] applied to the audio modality, the LR for the sLSTM model applied to the vision modality, and the LR for the remaining part of our model.

We utilize grid search and empirically set parameter ranges to perform multiple experiments in search of potentially optimal parameters. The batch size is chosen from {24, 32, 40}, the hidden size is selected from {128, 256, 300}, the number of attention heads in DVAWF is explored within the range of 1 to 12, the LR of BERT is set to $\{1e-5, 2e-5, 5e-5\}$, and LR for the two LSTMs and the remaining modules is chosen from $\{1e-4, 1e-3, 1e-2\}$. We employ the early stopping method used in [24] for the epoch value.

C. Data Preprocessing

In our experiments involving MOSI and MOSEI datasets, we utilize COVAREP [39] and P2FA [40] for extracting the audio features. Each segment is then represented by a 74-dimensional feature vector. For the vision modality, we employ Facet [41] to capture 35 facial action units for recording facial muscle movements, and serve as indicators of per-frame basic and advanced emotions. In the case of SIMS, we directly utilize the data provided by the published dataset [29].

D. Baselines for Comparison

The following baseline models are used for comparison on MOSI and MOSEI datasets: (1) TFN [16]; (2) LMF [17]; (3) MFN [42]; (4) MFM [7]; (5) MulT [8]; (6) RAVEN [43]; (7) MAG-BERT [14]; (8) MISA [21]; (9) TCSP [9]; (10) Self-MM [24]; (11) MTAG [44]; (12) MoNIG [45]; (13) MCMulT [10]; (14) EF-HEMT [15]; (15) AOBERT [46]. In these baselines, we select several recent works and replicate them under the same experimental conditions to compare with our MVIR model, including MAG-BERT [14], Self-MM [24], and AOBERT [46]. MAG-BERT [14] enables BERT to incorporate multimodal nonverbal data during fine-tuning by modifying BERT's internal representation. Self-MM [24] extracts similarity and differentiated information from multimodal and unimodal tasks, enhancing feature representations. AOBERT [46] is pre-trained on two tasks simultaneously: Multimodal Masked Language Modeling (MMLM) and Alignment Prediction (AP), determining dependencies and relationships between modalities.

TABLE II
EXPERIMENTAL RESULTS ON THE MOSI DATASET. [1] IS FROM [24]; [2] IS FROM [9]; [3] IS FROM [10]; [4] IS FROM [15]; [5] IS FROM [44]. MODELS WITH * ARE REPRODUCED UNDER THE SAME CONDITIONS

MOSI				
Model	Acc2^h	F1^h	MAE^l	Corr^h
TFN[1] [16]	-/80.8	-/80.7	0.901	0.698
LMF[1] [17]	-/82.5	-/82.4	0.917	0.695
MFN[1] [42]	77.4/-	77.3/-	0.965	0.632
RAVEN[1] [43]	78.0/-	76.6/-	0.915	0.691
MFM[1] [7]	-/81.7	-/81.6	0.877	0.706
MulT[1] [8]	81.5/84.1	80.6/83.9	0.861	0.711
MISA[1] [21]	80.8/82.1	80.8/82.0	0.804	0.764
TCSP[2] [9]	-/80.9	-/81.0	0.908	0.710
MoNIG[4] [45]	80.6/-	80.6/-	0.951	0.680
MTAG[5] [44]	-/82.3	-/82.1	0.866	0.722
MCMulT[3] [10]	83.9/-	83.2/-	0.866	0.701
EF-HEMT[4] [15]	82.3/-	82.5/-	0.901	0.701
MAG-BERT* [14]	82.9/84.3	83.0/84.1	0.729	0.774
Self-MM* [24]	82.4/84.2	82.4/84.3	0.733	0.791
AOBERT* [46]	82.8/84.5	82.6/84.5	0.727	0.782
MVIR (Ours)*	**84.3/85.5**	**83.9/85.5**	**0.714**	**0.799**
MVIR (w/o MHA)*	83.7/85.2	83.8/85.3	0.736	0.786
MVIR (w/o MHG)*	83.2/85.1	83.5/85.2	0.744	0.776

Due to the close relationship between model training performance and its initialization parameters, and the fact that different random seeds result in varying initialization parameters, we conduct experiments with several different random seed initializations, and the performance averaged across these runs serves as the final result, where the random seed values are {1111, 1112, 1113, 1114, 1115}. To ensure a fair comparison, on MOSI and MOSEI, we conduct five separate runs of our model (MVIR) as well as three top-performing baselines, namely MAG-BERT [14], Self-MM [24], and AOBERT [46]. We then calculate the average performance across these five runs and report it as the final result for each model. For SIMS, we select the comparison methods as those in [11].

E. Evaluation Metrics

Consistent with previous related work [11], we evaluate the performance of our models using four metrics. For the binary sentiment classification task, we utilize the 2-class accuracy (Acc2) and F1 score (F1), which are based on percentage statistics. For the regression task, we employ the mean absolute error (MAE) and Pearson correlation (Corr), with a value between 0 and 1. In general, higher scores indicate better model performance, except for the mean absolute error (MAE) where a lower score indicates better performance. Note that we calculate Acc2 and F1 in two ways on MOSI and MOSEI: negative / non-negative (non-exclude zero) and negative / positive (exclude zero). These metrics provide a comprehensive evaluation of the effectiveness and accuracy of our models in MSA.

F. Experimental Results

Tables II, III, and IV present the performance of our proposed MVIR model on the popular MSA datasets: MOSI,

TABLE III
EXPERIMENTAL RESULTS ON THE MOSEI DATASET. [1] IS FROM [24]; [2] IS FROM [9]; [3] IS FROM [10]; [4] IS FROM [15]; [5] IS FROM [44]. MODELS WITH * ARE REPRODUCED UNDER THE SAME CONDITIONS

MOSEI				
Model	Acc2^h	F1^h	MAE^l	Corr^h
TFN[1] [16]	-/82.5	-/82.1	0.593	0.700
LMF[1] [17]	-/82.0	-/82.1	0.623	0.677
MFN[1] [42]	76.0/-	76.0/-	-	-
RAVEN[1] [43]	79.1/-	79.5/-	0.614	0.66
MFM[1] [7]	-/84.4	-/84.3	0.568	0.717
MulT[1] [8]	-/82.5	-/82.3	0.58	0.703
MISA[1] [21]	82.6/84.2	82.7/84.0	0.568	0.724
TCSP[2] [9]	-/82.8	-/82.7	0.576	0.715
MoNIG[4] [45]	81.0/-	81.5/-	0.600	0.688
MCMulT[3] [10]	83.1/-	82.8/-	0.582	0.706
EF-HEMT[4] [15]	81.9/-	82.15/-	0.597	0.699
MAG-BERT* [14]	83.2/84.4	83.1/84.4	0.541	0.757
Self-MM* [24]	82.7/84.9	82.5/84.8	0.544	0.761
AOBERT* [46]	83.1/84.9	82.9/84.7	0.553	0.747
MVIR (Ours)*	**83.9/85.8**	**84.2/85.6**	**0.531**	**0.770**
MVIR (w/o MHA)*	83.5/85.5	83.6/85.4	0.544	0.760
MVIR (w/o MHG)*	83.3/85.4	83.3/85.3	0.549	0.757

MOSEI, and SIMS, respectively. Models with * means that we rerun the method under the same conditions. We highlight the best results in bold. Note that ^{h} means higher score is better, ^{l} means lower score is better.

In all three datasets, MVIR consistently achieves superior results across nearly all metrics, surpassing the baselines. For instance, when compared to Self-MM, our MVIR demonstrates performance improvements of 1.9%/1.3% and 1.5%/1.2% in Acc2 and F1, respectively, on the MOSI dataset. In comparison to AOBERT, MVIR yields performance gains of 1.5%/1% and 1.3%/1%. Compared to MAG-BERT, MVIR achieves 1.4%/1.2% and 0.9%/1.4% improvements in performance. These results signify that, thanks to the learning of multi-view shared and private information through SSLGM, and the fusion and interaction of cross-modal information via DVAWF, MVIR can capture more comprehensive and subtle emotional cues compared to existing methods. This, in turn, aids in learning richer multimodal features with more nuanced emotional information, ultimately enhancing the recognition performance.

G. Study on DVAWF

In the realm of modality interaction, DVAWF serves as a platform for information exchange and comprises primarily two multi-head mechanisms, namely MHG and MHA. To probe the inner workings and effectiveness of DVAWF, we devise three experiments for validation: ablation experiments concerning MHA and MHG, assessment of the impact of the number of attention heads on performance, and visualization experiments for attention mechanisms.

Study on the impact of MHA and MHG in DVAWF on multimodal representation ability: Tables II, III, and IV present the performance of the model without using MHA and MHG on different datasets, indicated as "w/o MHA" and "w/o MHG" respectively.

For example, in the experiments conducted on the MOSI dataset, as shown in Table II, the performance without MHA decreased by 0.6%/0.3%/0.1%/0.20%/0.022/0.013 compared to the complete MVIR model, while the performance without MHG decreased by 1.1%/0.4%/0.4%/0.3%/0.030/0.023. We can observe that the model's performance decreases when MHA or MHG is not used, indicating that both can assist in better sentiment analysis. Additionally, we find that WHG and WHA contribute similarly to the model, with WHG providing slightly greater assistance. This phenomenon is also observed in the performance on the other two datasets.

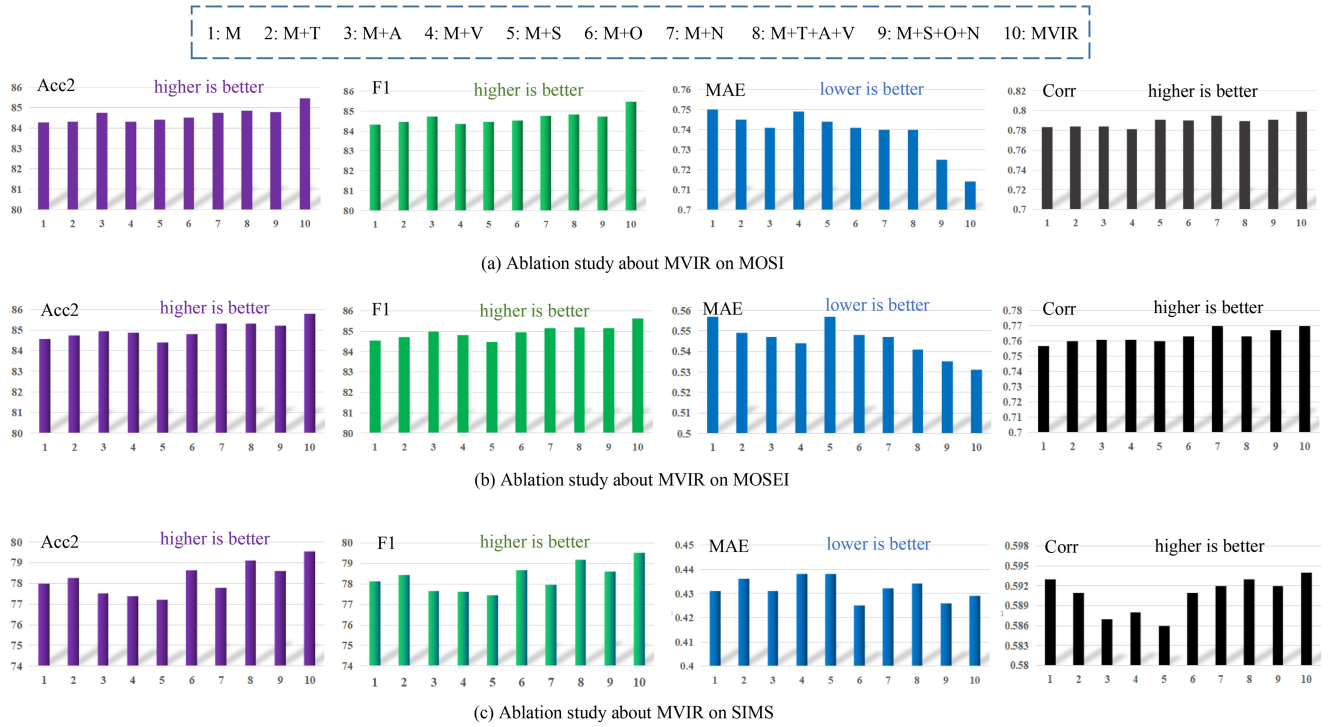
TABLE IV
EXPERIMENTAL RESULTS ON THE SIMS DATASET. MODELS WITH * ARE REPRODUCED UNDER THE SAME CONDITIONS

SIMS				
Model	Acc2^h	F1^h	MAE^l	Corr^h
LF-DNN [11]	76.68	76.48	0.446	0.567
EF-LSTM [11]	69.37	56.82	0.591	0.380
TFN [16]	77.07	76.94	0.437	0.582
LMF [17]	77.42	77.35	0.438	0.578
MFM [7]	75.06	75.58	0.477	0.525
MFN [42]	78.55	78.23	0.442	0.575
GMFN [28]	78.77	78.21	0.445	0.578
Self-MM* [24]	78.69	78.71	**0.422**	0.584
MVIR (Ours)*	**79.56**	**79.52**	0.429	**0.594**
MVIR (w/o MHA)*	79.41	79.38	0.436	0.581
MVIR (w/o MHG)*	79.33	79.31	0.442	0.577

Study on different number of attention heads in DVAWF: In our DVAWF module, the number of attention heads can be adjusted to obtain diverse features. In order to understand the impact of the number of attention heads on the overall model performance, we conduct experiments on MOSI, MOSEI, and SIMS datasets. The results are shown in Figure 5, and we have made the following observations:

- The optimal number of attention heads for achieving the best performance varies across different datasets, and having more attention heads does not necessarily result in better performance. Specifically, the best number of attention heads for achieving the highest performance on most metrics are 12, 11, and 7 for the MOSI, MOSEI, and SIMS datasets, respectively.
- Even within the same dataset, there is no single perfect number of attention heads that consistently yields the best results across all metrics. For example, on the MOSI dataset, while the scheme with 12 attention heads achieves the best performance in terms of Acc2^h, F1^h, and Corr^h metrics, the schemes with 11 and 6 attention heads perform the best in terms of MAE^l metric. Similar observations are also found in the MOSEI and SIMS datasets.

Visualizations study on attention mechanisms within DVAWF: To examine modality interactions within DVAWF, we randomly select a sample from each dataset and visualize the attention matrices of MHA and MHG on the main task, as shown in Fig. 6. Sub-figure (a) features a sample from the MOSI dataset with the text "THEY SHOULD'VE LIVED UP WITH MORE." Sub-figure (b) presents a MOSEI example with the text "In fact, it's sort of how I got my job," while (c) showcases a Chinese sample from the SIMS dataset with the text "What do you think he should do?" (translated into English for clarity). As both MHA and MHG are multi-headed,

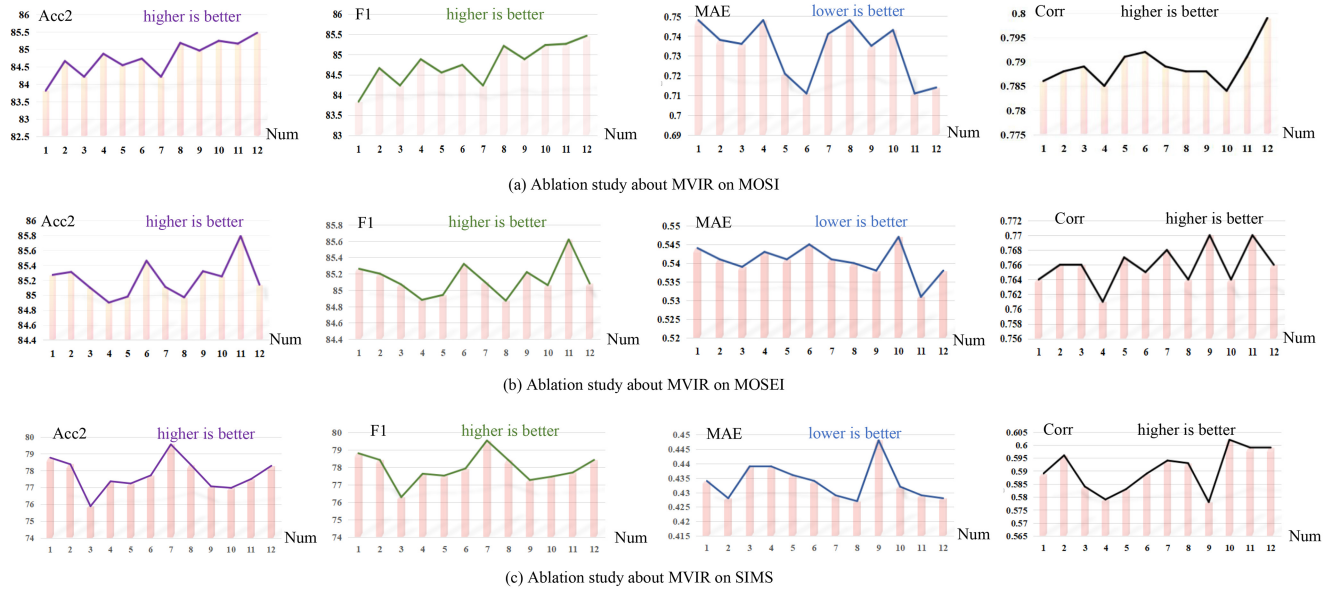


(a) Ablation study about MVIR on MOSI

(b) Ablation study about MVIR on MOSEI

(c) Ablation study about MVIR on SIMS

Fig. 4. Results of ablation experiments to explore the effect of different tasks on MVIR. M, T/A/V, S/O/N represent the tri-modal interactive task, text/audio/vision uni-modal independent tasks, text+audio/text+vision/audio+vision bi-modal interactive tasks, respectively.



(a) Ablation study about MVIR on MOSI

(b) Ablation study about MVIR on MOSEI

(c) Ablation study about MVIR on SIMS

Fig. 5. Results of ablation experiments to explore the effect of different number of attention head on MVIR. The index of the horizontal coordinate represents the number of attention heads.

we average the element values of all attention heads according to each position in attention matrices to obtain the overall attention scores.

Our observations across different datasets reveal two key findings. Firstly, regardless of the dataset, text consistently receives the highest attention across all modalities. This suggests that text provides the most significant information for overall sentiment analysis and representation learning, consistent with prior research [9], [14]. Secondly, when MHA and MHG engage in feature interaction within samples, they tend to demonstrate distinct attention patterns. This divergence in attention effectively provides two different perspectives for feature updates. For example, in the case of MOSI (sub-figure (a)), MHA predominantly focuses on the text modality, largely ignoring audio and video modalities, while MHG, allocates a great deal of attention to the text and audio modality. Thus, it can be seen that DVAWF offers a way to learn features from multiple perspectives, and integrating such features provides us

with a strategy to learn more superior and robust multimodal features.

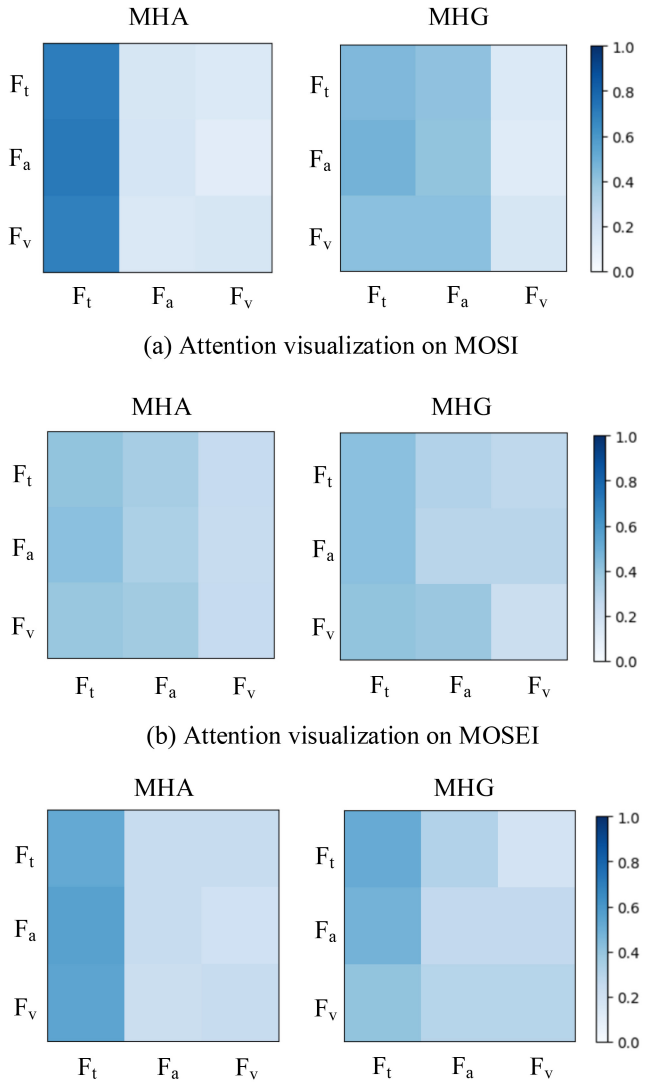
H. Study on SSLGM

In this section, we primarily investigate the effectiveness of the SSLGM algorithm. This portion comprises two experiments: one focusing on the study of different task combinations and the other on the examination of different sources of supervised information.

Study on different task combinations: In order to evaluate the effectiveness of the seven modality state tasks in learning multi-view interactive sentiment knowledge, as well as the impact of different task combinations on the main tri-modal interactive task, we conduct ablation experiments. We explore various combinations of auxiliary modality interactive tasks and investigate their effects on the main task. As all auxiliary tasks of MVIR are generated using the SSLGM algorithm, this experiment can also explore whether SSLGM can provide effective sentiment information for various tasks.

The results for MSA with different task combinations are presented in Figure 4. Here are the questions (Q) and corresponding answers (A) regarding the findings:

- Q: Is sentiment knowledge from different modality interactive states truly useful?
  A: Yes, the knowledge from different modality interactive states is indeed beneficial. As shown in Figure 4, across all datasets, incorporating tasks from different states through multi-task learning generally improves performance compared to using only the tri-modal main task.
  However, it is important to note that on the SIMS dataset, while incorporating all auxiliary tasks improves performance, using a single auxiliary task can sometimes result in performance degradation compared to using only the main task. We hypothesize that this phenomenon may be attributed to the presence of noise in the data. These noises can potentially hinder the SSLGM of the specific auxiliary task from generating accurate pseudo labels. Consequently, the incorrect label information might adversely affect the performance of the main task, leading to a negative impact when adding certain auxiliary tasks. However, when all tasks are combined, the valuable information contained in the data compensates for some of the noise interference. The diverse perspectives and complementary information provided by multiple auxiliary tasks collectively contribute to enhancing the overall performance of MVIR on the SIMS dataset.
- Q: Among the three modalities (text/audio/vision), which modality's auxiliary task is most helpful for performance improvement?
  A: The effectiveness of different auxiliary tasks depends on the specific dataset. For MOSI and MOSEI, the schemes involving the addition of audio (A) and audio with text (N) tasks achieve the best results in most metrics when only one auxiliary task is added. Hence, we can infer that the auxiliary task of the audio modality may be most helpful in these cases. On the other hand, for the SIMS dataset, the schemes involving the addition of text (T) and text with vision (O) tasks achieve the best results in most metrics when only one auxiliary task is added. This suggests that the auxiliary task of the text modality may be most helpful in this case.
- Q: When all uni-modal auxiliary tasks and all bi-modal auxiliary tasks are added respectively, which one leads to greater performance improvement?
  A: We compare the schemes of M+T+A+V (adding all uni-modal tasks) and M+S+O+N (adding all bi-modal tasks) and found that they exhibit similar performances. However, in general, the uni-modal scheme tends to yield better results in most cases, while the bi-modal scheme shows some advantages in regression performances.



(a) Attention visualization on MOSI

(b) Attention visualization on MOSEI

(c) Attention visualization on SIMS

Fig. 6. Visualizations of attention mechanisms within DVAWF, encompassing MHA and MHG components. Sub-figure (a) depicts an instance with the data "THEY SHOULD'VE LIVED UP WITH MORE" sourced from the MOSI dataset, sub-figure (b) presents an example with the data "In fact, it's sort of how I got my job" from the MOSI dataset, and (c) showcases an instance with the Chinese data "What do you think he should do?" from the SIMS dataset. Note that SIMS is a Chinese dataset that the example in sub-figure (c) is translated to English for clarity and exposition.

TABLE V
EXPERIMENTS FOR EXPLORING WHETHER SSLGM CAN PROVIDE EFFECTIVE MULTI-VIEW SENTIMENT INFORMATION. GT REFERS TO THE SCENARIO WHERE ALL TASKS IN MVIR UTILIZE THE MULTIMODAL LABELS PROVIDED BY THE DATASET. SSLGM REPRESENTS THE UNALTERED MVIR MODEL

MOSI				
Model	$Acc2^h$	$F1^h$	MAE^l	$Corr^h$
MVIR (GT)	83.8/84.6	83.2/85.1	0.729	0.787
MVIR (SSLGM)	**84.3/85.5**	**83.9/85.5**	**0.714**	**0.799**
MOSEI				
Model	$Acc2^h$	$F1^h$	MAE^l	$Corr^h$
MVIR (GT)	83.4/85.1	83.0/84.9	0.546	0.760
MVIR (SSLGM)	**83.9/85.8**	**84.2/85.6**	**0.531**	**0.770**
SIMS				
Model	$Acc2^h$	$F1^h$	MAE^l	$Corr^h$
MVIR (GT)	79.18	79.26	0.447	0.571
MVIR (SSLGM)	**79.56**	**79.52**	**0.429**	**0.594**

TABLE VI
EXAMPLES OF MULTIMODAL SENTIMENT ANALYSIS UTILIZING OUR MVIR FRAMEWORK. THE GROUND TRUTH SENTIMENT LABELS (GT) RANGE FROM STRONGLY NEGATIVE (-3) TO STRONGLY POSITIVE (+3). ID REPRESENTS THE INDEX OF EACH EXAMPLE. FOR EACH EXAMPLE, WE DISPLAY THE GT, AS WELL AS THE PREDICTION OUTPUTS OF MVIR USING ONLY THE TRI-MODAL INTERACTIVE TASK (MVIR(M)) AND THE COMPLETE SET OF TASKS (MVIR(C))

ID	Sentence	GT	MVIR(M)	MVIR(C)
1	And he was still boring.	-1.8	-2.0	-1.8
2	Youre like oh theyre still kind of cool	1.0	1.8	1.3
3	Um kato kicks ass handsome guy	1.2	1.6	1.2
4	Just like look away what he can do lots of stuff	-1.4	0.3	-1.2

Study on different supervised information: In order to investigate whether SSLGM can facilitate the acquisition of effective multi-view shared and private information by the model, we conduct a comparative experiment. In this experiment, we utilize the multimodal labels provided by the dataset as the supervision for all tasks in the MVIR model (denoted as GT), replacing the pseudo-labels generated by SSLGM. As shown in Table V, through a comprehensive performance comparison of various metrics, we observe a clear superiority in both regression and classification metrics when employing the MVIR model with SSLGM. For example, on MOSI, the model with SSLGM can improve 0.5%/0.9%, and 0.7%/0.4% than that with GT in Acc2 and F1. This demonstrates that SSLGM is indeed capable of extracting more effective emotional information from the data, thus furnishing the model with multi-view emotional cues across various interaction states.

I. Qualitative Analysis

In Table VI, we provide examples that demonstrate the effectiveness of MVIR(M) and MVIR(C) in adjusting sentiment intensity by considering either only the tri-modal interactive task and the complete set of auxiliary tasks. These examples highlight the importance of incorporating sentiment knowledge from different modality interaction states to improve model performance.

In most cases, MVIR(M) accurately predicts the polarity of the displayed sentiment, except for the last example. However, MVIR(M) is unable to utilize additional sentiment information present in other modality interaction states. On the other hand, MVIR(C) can better predict the magnitude of the displayed sentiment by leveraging the available information from multiple modalities. For instance, in Example 4, the sentiment is negative, but MVIR(M) mistakenly predicts a positive sentiment. However, MVIR(C) correctly predicts the emotional tendency by considering the complete set of auxiliary tasks. These examples demonstrate that our model, which incorporates multi-view sentiment knowledge from different modality interaction states, achieves superior performance compared to relying solely on the tri-modal interactive task.

V. CONCLUSION

Contemporary approaches often employ intricate interaction fusion mechanisms for effective multimodal representation. However, they struggle to capture shared-private modality information across varying interaction states, limiting the acquisition of refined multimodal sentiment features. To address this, we develop the MVIR framework, leveraging multi-task learning with the introduced SSLGA to acquire comprehensive multi-view interactive representations. Our proposed DVAWF module enhances feature interaction and fusion. We conducted rigorous comparisons with established baselines across three popular datasets and validated the core modules SSLGM and DVAWF through extensive ablation and comparative experiments. Findings show that incorporating SSLGM in the multi-task framework enables effective learning of shared and private multi-view information under varying interaction states. Notably, the optimal number of attention heads is data-dependent. Ablation and visualization experiments highlight DVAWF's effectiveness in providing multiple modal interaction perspectives for robust emotional information acquisition. Qualitative analysis further demonstrates improved recognition performance with the utilization of multi-view emotional cues. These experiments collectively confirm the efficacy and superiority of our approach. In the future, we aim to extend our methods to other tasks [47], [48], [49], [50], [51], [52], [53], [54].

REFERENCES

[1] G. A. Prabhakar, B. Basel, A. Dutta, and C. V. R. Rao, "Multichannel CNN-BLSTM architecture for speech emotion recognition system by fusion of magnitude and phase spectral features using DCCA for consumer applications," *IEEE Trans. Consum. Electron.*, vol. 69, no. 2, pp. 226–235, May 2023.
[2] H.-G. Kim, G. Y. Lee, and M.-S. Kim, "Dual-function integrated emotion-based music classification system using features from physiological signals," *IEEE Trans. Consum. Electron.*, vol. 67, no. 4, pp. 341–349, Nov. 2021.
[3] R. Chatterjee, S. Mazumdar, R. S. Sherratt, R. Halder, T. Maitra, and D. Giri, "Real-time speech emotion analysis for smart home assistants," *IEEE Trans. Consum. Electron.*, vol. 67, no. 1, pp. 68–76, Feb. 2021.
[4] J. Wang, S. Wang, M. Lin, Z. Xu, and W. Guo, "Learning speaker-independent multimodal representation for sentiment analysis," *Inf. Sci.*, vol. 628, pp. 208–225, May 2023.
[5] Z. Li et al., "Multi-level correlation mining framework with self-supervised label generation for multimodal sentiment analysis," *Inf. Fusion*, vol. 99, Nov. 2023, Art. no. 101891.

[6] J. Tang et al., "BAFN: Bi-direction attention based fusion network for multimodal sentiment analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 4, pp. 1966–1978, Apr. 2022.

[7] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," 2018, *arXiv:1806.06176*.

[8] 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. Conf. Assoc. Comput. Linguist. Meet.*, 2019, pp. 1–12.

[9] Y. Wu, Z. Lin, Y. Zhao, B. Qin, and L.-N. Zhu, "A text-Centered shared-private framework via cross-modal prediction for multimodal sentiment analysis," in *Proc. Find. Assoc. Comput. Linguist, (ACL-IJCNLP)*, 2021, pp. 4730–4738.

[10] L. Ma, Y. Yao, T. Liang, and T. Liu, "Multi-scale cooperative multimodal transformers for multimodal sentiment analysis in videos," 2022, *arXiv:2206.07981*.

[11] H. Mao, Z. Yuan, H. Xu, W. Yu, Y. Liu, and K. Gao, "M-SENA: An integrated platform for multimodal sentiment analysis," 2022, *arXiv:2203.12441*.

[12] S. Mai, S. Xing, and H. Hu, "Analyzing multimodal sentiment via acoustic-and visual-LSTM with channel-aware temporal convolution network," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1424–1437, Mar. 2021.

[13] J. Wu, S. Mai, and H. Hu, "Interpretable multimodal capsule fusion," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1815–1826, May 2022.

[14] W. Rahman et al., "Integrating multimodal information in large pre-trained transformers," in *Proc. Conf. Assoc. Comput. Linguist. Meet.*, 2020, pp. 1–11.

[15] Y. Ma and B. Ma, "Multimodal sentiment analysis on unaligned sequences via holographic embedding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2022, pp. 8547–8551.

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

[17] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," 2018, *arXiv:1806.00064*.

[18] J. Tang, K. Li, X. Jin, A. Cichocki, Q. Zhao, and W. Kong, "CTFN: Hierarchical learning for multimodal sentiment analysis using coupled-translation fusion network," in *Proc. 59th Annu. Meet. Assoc. Comput. Linguist.*, 2021, pp. 5301–5311.

[19] B. Van Amsterdam, A. Kadkhodamohammadi, I. Luengo, and D. Stoyanov, "ASPnet: Action segmentation with shared-private representation of multiple data sources," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 2384–2393.

[20] L. Zhu, Z. Zhu, C. Zhang, Y. Xu, and X. Kong, "Multimodal sentiment analysis based on fusion methods: A survey," *Inf. Fusion*, vol. 95, pp. 306–325, Jul. 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*, 2020, pp. 1122–1131.

[22] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 12, pp. 5586–5609, Dec. 2022.

[23] M. Crawshaw, "Multi-task learning with deep neural networks: A survey," 2020, *arXiv:2009.09796*.

[24] 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.

[25] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. ICLR*, 2018, pp. 1–12.

[26] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

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

[28] A. B. 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. Meet. Assoc. Comput. Linguist.*, 2018, pp. 2236–2246.

[29] W. Yu et al., "CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in *Proc. 58th Annu. Meet. Assoc. Comput. Linguist.*, 2020, pp. 3718–3727.

[30] X. Yan, S. Hu, Y. Mao, Y. Ye, and H. Yu, "Deep multi-view learning methods: A review," *Neurocomputing*, vol. 448, pp. 106–129, Aug. 2021.

[31] Z. Han, C. Zhang, H. Fu, and J. T. Zhou, "Trusted multi-view classification with dynamic evidential fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2551–2566, Feb. 2023.

[32] G. Wen, P. Cao, H. Bao, W. Yang, T. Zheng, and O. Zaiane, "MVS-GCN: A prior brain structure learning-guided multi-view graph convolution network for autism spectrum disorder diagnosis," *Comput. Biol. Med.*, vol. 142, p. 105239, 2022.

[33] Q. Zheng, J. Zhu, Z. Li, Z. Tian, and C. Li, "Comprehensive multi-view representation learning," *Inf. Fusion*, vol. 89, pp. 198–209, Jan. 2023.

[34] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[35] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[36] C. Liu et al., "Adaptive smooth l1 loss: A better way to regress scene texts with extreme aspect ratios," in *Proc. IEEE Symp. Comput. Commun. (ISCC)*. IEEE, 2021, pp. 1–7.

[37] J. Xu, Z. Li, B. Du, M. Zhang, and J. Liu, "Reluplex made more practical: Leaky ReLU," in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, 2020, pp. 1–7.

[38] C.-P. Lee and C.-J. Lin, "A study on L2-loss (squared hinge-loss) multiclass SVM," *Neural Comput.*, vol. 25, no. 5, pp. 1302–1323, 2013.

[39] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP—A collaborative voice analysis repository for speech technologies," in *Proc. IEEE Int. Conf. Acoust., speech signal Process. (ICASSP)*, 2014, pp. 960–964.

[40] J. Yuan and M. Liberman, "Speaker identification on the SCOTUS corpus," *J. Acoust. Soc. America*, vol. 123, no. 5, p. 3878, 2008.

[41] P. Dente, D. Küster, L. Skora, and E. Krumhuber, "Measures and metrics for automatic emotion classification via FACET," in *Proc. Conf. Study Artif. Intell. Simul. Behaviour (AISB)*, 2017, pp. 160–163.

[42] 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, pp. 1–9.

[43] 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, pp. 7216–7223.

[44] J. Yang et al., "MTAG: Modal-temporal attention graph for unaligned human multimodal language sequences," in *Proc. Conf. North Am. Chapter Assoc. Comput. Linguist., Human Lang. Technol.*, 2021, pp. 1009–1021.

[45] H. Ma, Z. Han, C. Zhang, H. Fu, J. T. Zhou, and Q. Hu, "Trustworthy multimodal regression with mixture of normal-inverse gamma distributions," in *Proc. 35th Conf. Neural Inf. Process. Syst.*, 2021, pp. 6881–6893.

[46] K. Kim and S. Park, "AOBERT: All-modalities-in-one BERT for multimodal sentiment analysis," *Inf. Fusion*, vol. 92, pp. 37–45, Apr. 2023.

[47] Y. Han, L. Li, and J. Zhang, "A coordinated representation learning enhanced multimodal machine translation approach with multi-attention," in *Proc. Int. Conf. Multimedia Retrieval*, 2020, pp. 571–577.

[48] J. Ma, S. Qin, L. Su, X. Li, and L. Xiao, "Fusion of image-text attention for transformer-based multimodal machine translation," in *Proc. Int. Conf. Asian Lang. Process. (IALP)*, 2019, pp. 199–204.

[49] S. Yao and X. Wan, "Multimodal transformer for multimodal machine translation," in *Proc. 58th Annu. Meet. Assoc. Comput. Linguist.*, 2020, pp. 4346–4350.

[50] X. Zhou et al., "Information theoretic learning-enhanced dual-generative adversarial networks with causal representation for robust OOD generalization," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Nov. 17, 2023, doi: 10.1109/TNNLS.2023.3330864.

[51] X. Zhou et al., "Digital twin enhanced federated reinforcement learning with lightweight knowledge distillation in mobile networks," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 10, pp. 3191–3211, Oct. 2023.

[52] X. Zhou et al., "Hierarchical federated learning with social context clustering-based participant selection for Internet of Medical Things applications," *IEEE Trans. Comput. Social Syst.*, vol. 10, no. 4, pp. 1742–1751, Aug. 2023.

[53] X. Zhou, W. Liang, I. Kevin, K. Wang, and L. T. Yang, "Deep correlation mining based on hierarchical hybrid networks for heterogeneous big data recommendations," *IEEE Trans. Comput. Soc. Syst.*, vol. 8, no. 1, pp. 171–178, Feb. 2021.

[54] X. Zhou et al., "Decentralized P2P federated learning for privacy-preserving and resilient mobile robotic systems," *IEEE Wireless Commun.*, vol. 30, no. 2, pp. 82–89, Apr. 2023.