IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 27 September 2024, accepted 26 October 2024, date of publication 4 November 2024,
date of current version 18 November 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3490772

RESEARCH ARTICLE

Gated Fusion Network With Progressive Modality Enhancement for Sentiment Analysis

FENG XIN[1], WENJUAN GONG[1], TINGBO SHI[1], JING ZHONG[1], KECHEN LI[1], AND JORDI GONZÀLEZ[2], (Member, IEEE)

[1]Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China), Qingdao, Shandong 266580, China
[2]Computer Vision Center, Universitat Autònoma de Barcelona, 08193 Barcelona, Spain

Corresponding author: Wenjuan Gong (wenjuan_gong@163.com)

This work was supported in part by the Natural Science Foundation of Shandong Province under Grant ZR2023MF041, and in part by Spanish Ministry of Economy and Competitiveness (MINECO) and European Regional Development Fund (ERDF) funded by Spanish Ministry of Science and Innovation (MCIN)/State Research Agency (AEI)/10.13039/501100011033 under Grant PID2020-120311RB-I00.

ABSTRACT Multi-modal sentiment analysis aims to recognize the sentiment expressed by the studied subjects utilizing textual, visual, and acoustic cues. Given the vast application prospects of multi-modal sentiment analysis in domains such as human-computer interaction and intelligent customer service systems, it has gained widespread attention in recent years. The design of fusion strategy significantly impacts analysis results. Different modalities varied on the extent of contributions to the sentiment analysis process, and the textual modality tends to play a crucial role. However, most existing fusion strategies have ignored the issue of uneven distribution of sentiment information across modalities. Therefore, in this study, we introduce a gated fusion network with progressive modality enhancement. To effectively address the challenge of imbalanced sentiment distribution, we propose a progressive modality enhancement approach that facilitates the interaction of sentiment cues. Furthermore, to overcome the limitations of insufficient inter-modal interactions observed in previous works, we introduce a gated multi-modal information interaction module. This module comprehensively exploits multi-modal data, ensuring that each modality can effectively utilize information from others. Experimental results demonstrate that the proposed method achieves highly competitive performance on the CMU-MOSI, CMU-MOSEI, and CH-SIMS datasets. Furthermore, the ablation experiments confirm the dominant role of the textual modality, while indicating that the visual modality generally performs better than the acoustic modality across our experimental settings.

INDEX TERMS Sentiment analysis, multi-modal fusion, modality enhancement.

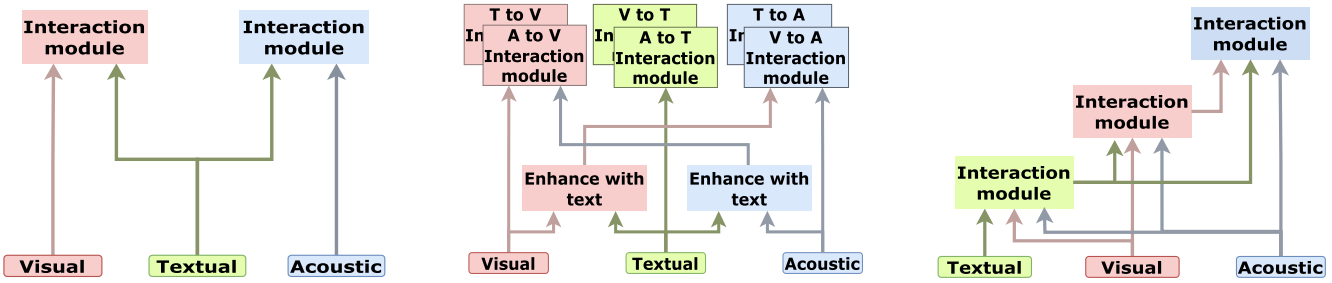
I. INTRODUCTION

In recent years, short videos have experienced tremendous growth as more individuals share their opinions on trending topics through social media platforms [1], [2], [3]. These videos integrate multiple modalities including visual content, audio, and text. Such diverse components jointly convey the subjects' sentiments. The objective of Multi-modal Sentiment Analysis (MSA) is to analyze data from these diverse modalities to accurately predict the sentiment state of a subject. Given its broad applications in human-computer interaction and intelligent customer service, MSA has garnered significant attention from the research community.

Recent years have witnessed the proposal of numerous effective methods for MSA. Nevertheless, a subset of these studies treats different modalities as equivalent, overlooking their inherent differences [1], [4], [5], [6], [7]. The textual modality often encapsulates richer information and more structure compared to non-textual modalities. Recent research has consistently demonstrated that sentiment cues are unevenly distributed across different modalities in MSA

The associate editor coordinating the review of this manuscript and approving it for publication was Shadi Alawneh.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/



(a) Fusion Schemes a [2], [8] (b) Fusion Schemes b [9] (c) Fusion Schemes c

FIGURE 1. Comparison of different fusion schemes.

tasks [2], [8], [9], [10], [11]. Notably, textual modality has been identified as important carriers of sentiment signals, playing a more significant role in sentiment analysis tasks, especially when the input textual data exhibits a high degree of reliability and quality.

Researchers have proposed several text-centered approaches to address this problem [2], [8], [9], as illustrated in Fig. 1a and Fig. 1b. These approaches enhance modal representations through interactions between the original textual modality and other modalities. While effective in using textual modalities to guide the enhancement of non-textual modalities, these methods have a significant limitation since the information in the original textual modality does not always represent the correct sentimental state of the speaker. Consider the phrase "What a brilliant idea." On its own, this text seems to convey positive sentiment. However, if spoken with a sarcastic tone and accompanied by an eye roll, the overall sentiment becomes negative, contradicting the seemingly positive text. Therefore, directly using original textual features may not effectively guide non-textual modalities in learning complementary information.

To address this limitation, this work proposes a novel gated fusion network with progressive modality enhancement (GFNPME). Specifically, this network incorporates a progressive modality enhancement approach, which prioritizes the dominant modality and leverages the enhanced modal information to process subsequent modalities, as illustrated in Fig. 1c. Initially, we prioritize the textual modality learns complementary information from other modalities, helping to disambiguate the text and better represent the speaker's sentiment state. Subsequently, this enhanced textual representation guides the sequential enhancement of visual and acoustic modalities.

In addition, these text-centered approaches primarily focus on the interaction between only two modalities, lacking comprehensive interaction among all modalities. To facilitate the learning of complementary information across different modalities, we propose a gated multi-modal information interaction module (GMIM). GMIM consists of two key components, multi-modal information aggregation unit (MIAU) and gated recurrent unit (GRU) [12], which divide the learning of complementary information into two parts. The MIAU realizes multi-modal information aggregation through an attention mechanism, while the GRU employs a gating mechanism to acquire complementary information from other modalities, ensuring that each interaction involves all three modalities.

Additionally, we introduce three single-modal tasks targeting textual, acoustic, and visual sentiment analysis to assist the model in capturing modality-specific information.

To evaluate the effectiveness of our proposed method, we conducted sentiment intensity prediction tasks on three benchmark datasets: CMU-MOSI, CMU-MOSEI, and CH-SIMS. Experimental results demonstrate that our proposed model outperforms state-of-the-art (SOTA) methods across several evaluation metrics. Furthermore, ablation studies demonstrated the effect of each module, and case studies elaborated on the advantages of the model. The key contributions of the proposed method are summarized as follows:

- This study proposes a novel progressive modality enhancement approach to facilitate the exchange of sentiment information across modalities.
- We propose a gated multi-modal information interaction module comprising a GRU and a multi-modal information aggregation unit to ensure full interaction among all three modalities, thereby improving the performance of multi-modal fusion.
- Extensive experiments on three benchmark datasets for multi-modal sentiment analysis showed that the proposed GFNPME method achieved competitive results compared to SOTA methods.

II. RELATED WORK

A. MULTI-MODAL SENTIMENT ANALYSIS

Multi-modal signals offer richer sentiment information, and the combination of multiple modalities enables more accurate sentiment analysis [1], [4], [13]. Consequently, the effective fusion of information from each modality has been a central focus of research. To fully leverage multi-modal data, numerous advanced fusion methods were proposed in recent years to explore interrelationships across modalities, aiming to learn more information-rich multi-modal representations. Zadeh et al. [1] proposed the tensor fusion network (TFN),

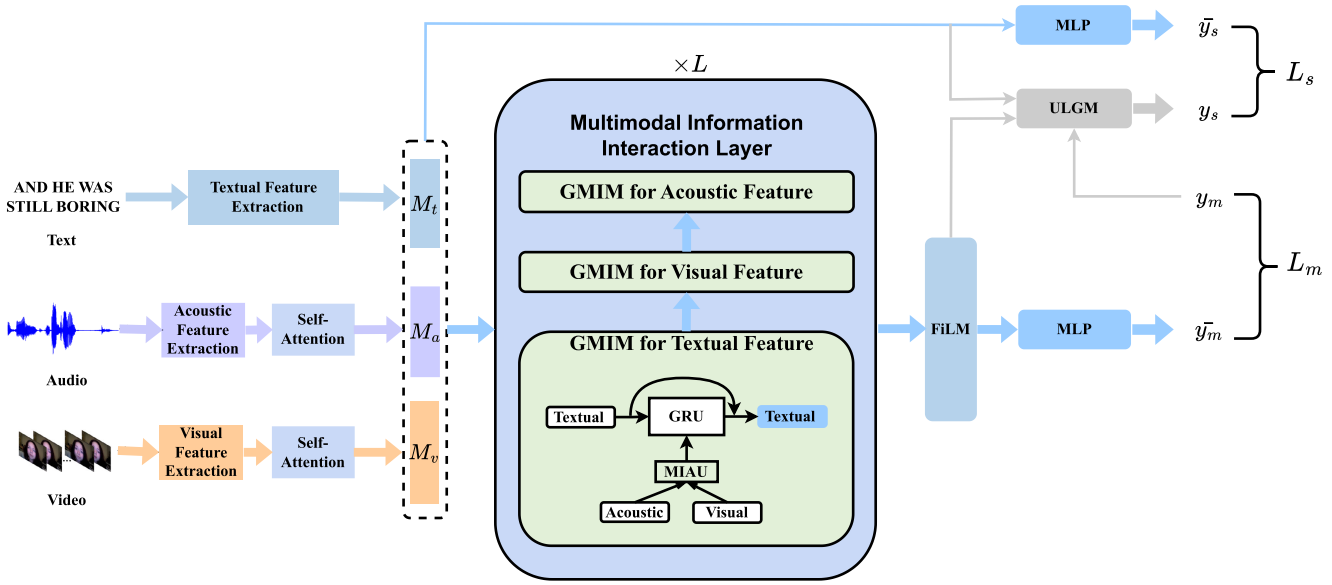


FIGURE 2. Overall architecture of the proposed GFNPME method.

using three-fold Cartesian product to explicitly model the interrelationships among unimodal, bimodal, and trimodal modalities. Sun et al. [5] proposed an MLP-based model for multimodal sentiment analysis and depression estimation, named CubeMLP, that utilized a fully-connected layer to perform multi-modal interaction.

Additionally, graph convolution was employed in some studies for modal interaction [14], [15], [16]. Mai et al. [14] introduced the graph pooling fusion network (GPFN), utilizing three single-modal graph convolution networks to examine intramodal dynamics, and for intermodal interactions a pooling strategy called link similarity pooling was designed to efficiently fuse the information from different modalities.

Related research observed that heterogeneity between modalities can significantly impact information fusion [17]. Therefore, Hazarika et al. [17] proposed the modality-invariant and modality-specific representations for multi-modal sentiment analysis model, which employed a shared feature extractor between modalities and a modality-specific encoder to extract homogeneous and heterogeneous features, respectively.

Although these previous works achieved promising results, many of them tended to treat the three modalities equally, potentially overlooking the inherent advantages of textual modalities.

In order to fully utilize the advantages of textual modalities, Han et al. [2] proposed the bi-bimodal fusion network (BBFN), which fused text with acoustic and visual modalities, respectively, and introduced layer-wise feature space separator to ensure the mutual independence of modalities in the fusion process. Wang et al. [9] proposed the text enhanced transformer fusion network (TETFN), which used text to catalyze the interaction between acoustic and visual modalities by learning text-oriented pairwise cross-modal mappings through the text-oriented multi-head attention mechanism and cross-modal transformers.

In contrast to previous text-centered approaches, we propose a progressive modality enhancement approach. This novel approach leverages enhanced textual information to guide other modalities in learning complementary features.

B. MULTI-TASK LEARNING FOR SENTIMENT ANALYSIS

Multi-task learning leverages the correlation among tasks to improve the model's generalization ability and performance [18], [19], [20]. In the field of MSA, multi-task learning models fall into two main categories. One involves conducting learning of sentiment analysis tasks from multi-modalities together with related tasks from single-modalities. This strategy aims to leverage single-modal sentiment tasks to help the model acquire modality-specific features, enhancing the accuracy of multi-modal tasks. For example, Yu et al. [21] argued that using unified multi-modal labels could restrict the model's ability to learn specific modality features. They introduced a single-modal classification task to assist the model in learning modality-specific features.

Given that most existing multi-modal datasets only have multi-modal label annotations, Yu et al. [22] creatively introduced a unimodal label generation module to generate single-modal labels for single-modal tasks.

Another category involves leveraging two or more different sentiment analysis tasks across multiple modalities. For instance, Zhang et al. [23] proposed a unified

multi-task learning framework, M2SeqSeq, which integrated multi-modal emotion polarity (positive, negative, and neutral), emotion categories (e.g., anger, disgust, fear, joy, sadness, and surprise), and sarcasm recognition into a single model. They explicitly employed a gating network to model the interdependence between different tasks. In this study, we leverage single-modal tasks of sentiment analysis to support multi-modal learning, aiding the model in capturing modality-specific information.

III. GATED FUSION NETWORK WITH PROGRESSIVE MODALITY ENHANCEMENT

The overall architecture of the proposed method is depicted in Fig. 2. It incorporates stacked multi-modal information interaction layers (MIILs) that enable single-modal feature representations to iteratively learn complementary information from other modalities. Initially, we extract and encode the contextual information from each modality. These encoded features are then fed into the MIILs for inter-modal complementary learning and used to generate modality-specific labels through the unimodal label generator module for single-modal sentiment analysis tasks, which facilitates the model in capturing modality-specific information. After processing through MIILs, the enhanced single-modal feature representations are concatenated and fed into a feature-wise linear modulation (FiLM) module [24] for multi-modal fusion, generating multi-modal representations for sentiment prediction. In the figure, y_s, y_m denote the single-modal and multi-modal labels, respectively, and $\bar{y_s}$ and $\bar{y_m}$ denote the predicted values for the single-modal and multi-modal tasks, respectively. The model is optimized under a multi-modal loss L_m and three single-modal losses L_s.

A. TASK FORMULATION AND FEATURE ENCODING

Suppose a video clip contains three modalities, including textual, visual, and acoustic modalities information from each modality is processed using its corresponding feature encoder. For the textual modality, we utilize the pre-trained language model, bidirectional encoder representation from transformers (BERT) [25] as the text encoder, obtaining a sequence representation with contextual information $M_T \in \mathbb{R}^{T_t \times d_t}$. Additionally, we select the first word vector in the last layer to represent the whole sentence $M_t \in \mathbb{R}^{d_t}$. For the acoustic and visual modalities, we extract the initial vector features $M_A \in \mathbb{R}^{T_a \times d_{ai}}$ and $M_V \in \mathbb{R}^{T_v \times d_{vi}}$, respectively. In practice, the initial feature dimensions of acoustic and visual modalities vary across datasets, which will be detailed in the experimental section. To better capture long-range dependencies and contextual information within these modalities, we further process them through a module incorporating a single multi-head self-attention layer. The features from the final time step after processing serve as the global representation for the entire sequence of each modality. Consequently, after feature coding, the feature representations of the textual, acoustic, and visual modalities are denoted as $M_t \in \mathbb{R}^{d_t}$, $M_a \in \mathbb{R}^{d_a}$, and $M_v \in \mathbb{R}^{d_v}$, where d_t, d_a, and d_v are 768, 16, and 32 respectively.

Thus, the objective of the MSA task is to detect the sentiment label y of the video by effectively fusing the encoded information M_t, M_a, M_v from these different modalities. Following previous works, we treat this as a regression task.

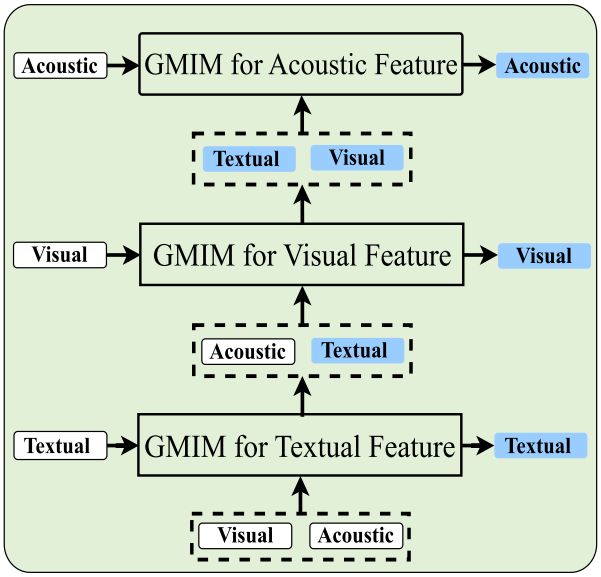


FIGURE 3. Overall architecture of a multi-modal information interaction layer.

B. MULTI-MODAL INFORMATION INTERACTION

The encoded features are further enhanced using stacked MIILs. Each MIIL, consisting of several GMIMs, processes all the considered modalities in a sequential order. This study considers three modalities. Thus, the designed MIIL comprises three GMIMs. When dealing with a different number of modalities, the module structure can be easily adjusted. For MSA, the acoustic and visual modalities are first utilized to enhance the textual modality, and the enhanced textual modality is then used to subsequently enhance the visual and acoustic modalities, as illustrated in Fig. 3. For further illustrate of the GMIM process, we use the processing unit for the textual modality in the first MIIL as an example. This unit enhances the textual modality with cues from the acoustic and visual modalities. Initially, information from the acoustic and visual modalities is aggregated using a MIAU, with its overall structure illustrated in Fig. 4. The aggregation process can be described as follows:

$$M_{av} = \sigma(W_{av}([M_v : M_a])), \tag{1}$$

$$\text{Att}_m = \sigma(W_m M_{av}), \tag{2}$$

$$M_m^* = M_m \odot \text{Att}_m, \tag{3}$$

$$M_{agg} = W_{agg}[M_v^* : M_a^*], \tag{4}$$

where the symbol : denotes concatenation, $\odot$ denotes the Hadamard product operation, σ is the activation function, $W_{av} \in \mathbb{R}^{(d_a+d_v)\times(d_a+d_v)}$, $W_m \in \mathbb{R}^{d_m\times(d_a+d_v)}$, and $W_{agg} \in \mathbb{R}^{d_t\times(d_a+d_v)}$ are learnable parameters, $M_{av} \in \mathbb{R}^{(d_a+d_v)}$ denotes the initial aggregated information, $\text{Att}_m \in \mathbb{R}^{d_m}$ denotes the attention score corresponding to each modality, and $M_{agg} \in \mathbb{R}^{d_t}$ denotes the final aggregated feature.

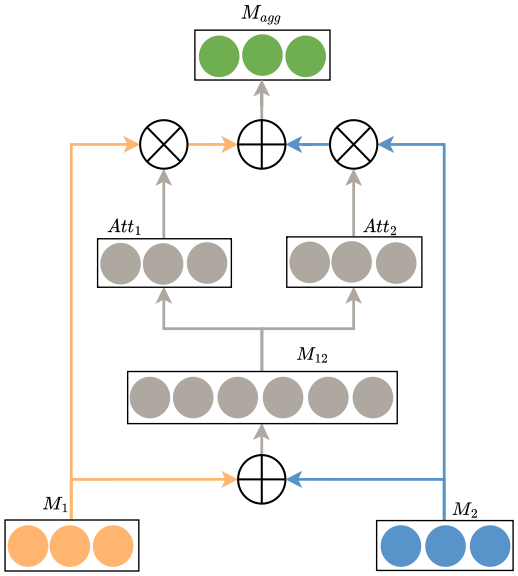


FIGURE 4. The structure of the proposed multi-modal information aggregation unit.

Then, a GRU is employed to generate the complementary information that needs to be passed to the textual modality. The aggregated acoustic and visual modalities are fed into the GRU, and the textual modality is utilized to update the aggregated features. This procedure is formulated as follows:

$$M_{msg} = \text{GRU}\left(M_t, M_{agg}\right). \tag{5}$$

The gating mechanism of the GRU allows the aggregated cue from the acoustic and visual modalities, M_{agg}, to reset and update its information. As specified in Eq. 5, the GRU takes the textual modality features, M_t, as input and M_{agg}, as the hidden state. First, the candidate information $M_h \in \mathbb{R}^{d_t}$ is acquired by utilizing the reset gate signal R, followed by obtaining the multi-modal complementary feature M_{msg} through the update gate signal Z. The formulated representations are as follows:

$$R = \sigma\left(W_{r1}M_{agg} + W_{r2}M_t\right), \tag{6}$$
$$Z = \sigma\left(W_{z1}M_{agg} + W_{z2}M_t\right), \tag{7}$$
$$M_h = \tanh(W_{h1}(R \odot M_{agg}) + W_{h2}M_t), \tag{8}$$
$$M_{msg} = Z \odot M_{agg} + (1 - Z) \odot M_h, \tag{9}$$

where W_* are learnable parameters.

Using a residual connection, we combine the original textual modality features, M_t, with the multi-modal complementary feature from the processing unit for the textual modality in the first MIIL, M_{msg}, to obtain the enhanced feature representation for the textual modality:

$$M_t^{i+1} = M_t^i + M_{msg}, \tag{10}$$

where i denotes the i-th layer, and i equals to one for the first layer.

Following the same procedures, the original visual modality feature representation, M_v, and the original acoustic modality feature representation, M_a, are also enhanced.

After L layers of iterative processing of MIILs, we concatenate the enhanced single-modal feature representations to maximize the preservation of sentiment information from each modality:

$$M_{tav}^L = [M_t^L : M_a^L : M_v^L], \tag{11}$$

where L denotes the number of MIILs and is a hyperparameter.

The enhanced multi-modal feature representation, $M_{tav}^L \in \mathbb{R}^{d_t+d_a+d_v}$, is fed into the FiLM module so that all modalities can further interact with each other and exchange information:

$$\gamma = f(M_{tav}^L), \tag{12}$$
$$\beta = h(M_{tav}^L), \tag{13}$$
$$M' = \gamma \odot M_{tav}^L + \beta, \tag{14}$$

where f and h are fully connected layers, their outputs are $\gamma \in \mathbb{R}^{d_t+d_a+d_v}$ and $\beta \in \mathbb{R}^{d_t+d_a+d_v}$, respectively, and $M' \in \mathbb{R}^{d_t+d_a+d_v}$ denotes the final multi-modal feature representation.

C. MULTI-TASK RE-FORMULATION AND OBJECTIVES

Even though all modalities in a video collectively express the same sentiment, single-modal and multi-modal labels might differ from each other. A single task of multi-modal sentiment analysis is not capable of capturing the inconsistencies among modalities. Introducing single-modal tasks and formulating the problem as a multi-task learning problem tend to boost the performance of multi-modal recognition by retaining modality-specific information. Therefore, we introduce single-modal sentiment analysis tasks, including textual, acoustic, and visual sentiment analysis, to assist the multi-modal sentiment analysis task.

The single-modal feature representations before applying the MIILs are taken as inputs to the respective single-modal sentiment analysis tasks. Multilayer perceptrons (MLPs) are employed to predict sentiment labels for each single-modal feature representation:

$$\bar{y}_s = MLP(M_s), \tag{15}$$

where M_s is equal to single-modal feature representation M_m and the symbol is modified for the purpose of clarity.

Considering that most existing multi-modal sentiment analysis datasets lack single-modality labels, we employ the unimodal label generation module (ULGM) proposed in the self-supervised multi-task multimodal sentiment analysis

network (Self-MM) [22]. For single-modal and multi-modal features, the ULGM computes the relative distances of the features to the centers of their positive and negative sample clusters, respectively. Subsequently, based on these relative distances, ULGM derives a mapping between the single-modal labels and the multi-modal labels. Thereby, the single-modal labels are computed from the multi-modal labels (i.e., polarity values) and relative distances to the centers. The single-modal tasks are exclusively included in the training process for optimizing model parameters but are not used in the validation and inference phases.

After introducing single-modal tasks, the sentiment analysis problem can be reformulated as a multi-task regression problem with multiple objectives, and the mean absolute error is used as the loss measure. The final loss is obtained by combining the loss values of the multi-modal and single-modal tasks:

$$\text{Loss} = \frac{1}{N}\sum_{1}^{N}\left(\left|\bar{y}_m^i - y_m^i\right| + \sum_{s}^{t,a,v} W_s^i \left|\bar{y}_s^i - y_s^i\right|\right), \tag{16}$$

$$W_s^i = \tanh(|\bar{y}_s^i - y_m^i|), \tag{17}$$

where N is the total number of training samples, and W_s^i is the weight of the i-th sample for the single-modal task s and represents the degree of difference between single-modal labels and multi-modal labels.

IV. EXPERIMENTS

In this section, we comprehensively evaluated the performance of the proposed method. First, we performed a quantitative comparison with several SOTA methods. Subsequently, we demonstrated the effectiveness of the multi-modal information interaction layer and the progressive modality enhancement approach through ablation studies. Finally, we visualized and qualitatively analyzed several representative results.

A. DATASETS

The proposed GFNPME method was evaluated on three widely-used multi-modal sentiment analysis datasets: CMU-MOSI [26], CMU-MOSEI [27] and CH-SIMS [21].

1) CMU-MOSI

The CMU-MOSI is a dataset of YouTube videos featuring speakers expressing opinions on various topics. The dataset comprises videos from 89 distinct speakers with a total of 2, 199 video clips. Of these, 1, 284 samples were allocated for training, 229 for validation, and 686 for testing. Each video clip was manually annotated with a sentiment score ranging from −3 (strongly negative) to +3 (strongly positive), indicating the intensity of the speaker's sentiment.

2) CMU-MOSEI

The CMU-MOSEI is an extension of CMU-MOSI, consisting of 23, 453 samples from 1, 000 speakers across 250 topics. For CMU-MOSEI, 16, 326 samples were designated for training, 1, 871 for validation, and 4, 659 for testing. The annotation approach for sentiment scores is identical to CMU-MOSI.

3) CH-SIMS

The CH-SIMS dataset is a Chinese MSA benchmark containing individual unimodal labels. The dataset consists of 2,281 refined video clips collected from different movies and TV serials. 1368 samples were used for training, 456 samples for validation and 457 samples for testing. Each video clip was manually annotated with a sentiment score ranging from −1 (strongly negative) to +1 (strongly positive).

The statistical information of the datasets is shown in Table 1.

B. FEATURE EXTRACTION

To ensure a fair comparison and isolate the impact of our proposed fusion strategy, we adopted the same feature extraction methodologies as those used in previous works.

For the CMU-MOSI and CMU-MOSEI datasets, we utilized the pre-trained BERT model as the text encoder. Visual features were extracted by the Facet tool.[1] Acoustic features were extracted by the collaborative voice analysis repository for speech technologies (COVAREP) [28]. For the CMU-MOSI dataset, the feature dimensions of acoustic, visual, and textual modalities are 5, 20, and 768, respectively. For the CMU-MOSEI dataset, the feature dimensions are 74, 35, and 768 for acoustic, visual, and textual modalities, respectively.

For the CH-SIMS dataset, we used the pre-trained Chinese BERT model for text encoding to obtain 768-dimensional features. Acoustic features were extracted using the LibROSA speech processing toolkit [29], yielding 33-dimensional features. Visual features were obtained using the MultiComp OpenFace 2.0 toolkit [30], resulting in 709-dimensional features.

TABLE 1. Statistics of the CMU-MOSI, CMU-MOSEI and CH-SIMS datasets.

Dataset	Train	Valid	Test	Total
CMU-MOSI	1,284	229	686	2,199
CMU-MOSEI	16,326	1,871	4,659	22,856
CH-SIMS	1,368	456	457	2,281

C. EXPERIMENTAL SETTING

We implemented our model using PyTorch 1.13.0 and CUDA 11.7. The Adam optimizer is employed with initial learning rates of 5e-5 for BERT and 1e-3 for other modules. We set the batch size to 32 for the CMU-MOSI and CH-SIMS datasets, and 256 for the MOSEI dataset. In the context of hyperparameter L, CMU-MOSI, CMU-MOSEI, and CH-SIMS are respectively set to 4, 3, and 2. To ensure fair comparison, all experiments and baseline methods utilize

[1] https://imotions.com/platform/

BERT as the text feature extractor. Since the CH-SIMS dataset provides unimodal labels, we directly utilized these labels in our experiments.

We use five different metrics to evaluate model performance:

- Mean absolute error (MAE), which directly calculates the error between predicted and ground truth sentiment scores. For all metrics except MAE, higher values indicate better performance.
- Pearson's correlation (Corr), which measures the linear correlation between predicted and ground truth sentiment scores.
- Seven-class accuracy (Acc7), where the ground truth sentiment scores ranging from -3 to 3 are evenly divided into seven intervals, and this metric indicates the proportion of predictions within the correct interval.
- Two-class accuracy (Acc2), which measures the overall correct prediction rate for binary tasks of negative/non-negative(non-exclude zero) and negative/positive(exclude zero) sentiment.
- F1 scores, which evaluates the performance of binary sentiment classification tasks negative/non-negative and negative/positive by calculating the harmonic mean of precision and recall.

For the SIMS dataset, we report four metrics: Corr, MAE, Acc2, and F1 score. Both Acc2 and F1 score are calculated for negative/non-negative classification.

D. BASELINE AND SOAT MODELS

We compared our results with several advanced multi-modal fusion frameworks:

- TFN [1]. Tensor Fusion Network captures unimodal, bimodal, and trimodal interactions by three-fold Cartesian products.
- LMF [4]. Low-rank Multimodal Fusion performs multimodal fusion using low-rank tensors to improve efficiency.
- BBFN [2]. Bi-Bimodal Fusion Network performs fusion and separation of paired modal representations by stacked cross-modal complementary layers.
- CubeMLP [5]. This method uses three MLP units for sequential-mixing, modality-mixing, and channel-mixing, respectively, thereby generating fused features for multi-modal analysis.
- Self-MM [22]. Self-Supervised Multi-task Multimodal Sentiment Analysis Network designs a unimodal label generation module to automatically generate unimodal supervised signals and jointly trains multi-modal and unimodal tasks to learn inter-modal consistency and dissimilarity.
- MISA [17]. This method learns effective modal representations by projecting each modality onto modality-invariant and modality-specific subspaces.
- TETFN [9]. Text Enhanced Transformer Fusion Network achieves effective multi-modal fusion representations by employing text-driven multi-head attention mechanisms to facilitate interaction between acoustic and visual modalities.
- HFNGC [31]. Heterogeneity Fusion Network Based on Graph Convolutional Neural uses graph convolution for multi-modal fusion, which overcomes the problem of semantic differences between modalities through a shared convolutional aggregation mechanism and dynamic routing techniques.
- AOBERT [32]. All-modalities in One BERT achieves joint representation learning for textual, acoustic and visual modalities by simultaneously performing pre-training tasks for multi-modal masked language modeling and alignment prediction.
- MIM [33]. Multimodal Interaction Model use subspaces to learn specific feature representations for each modality to improve the efficiency of information exchange and reduce the redundancy of modal information.
- DTN [34]. Disentanglement Translation Network employs a encoder-decoder based disentanglement framework to decouple the unimodal representation into modality-common and modality-specific subspaces, and also proposes a slack reconstruction method to capture desirable information properties.

E. RESULT AND DISCUSS

1) QUANTITATIVE RESULTS

To evaluate the overall performance of the proposed method, we conducted extensive comparisons with several SOTA approaches in the current field on three widely-used public datasets: CMU-MOSI, CMU-MOSEI and CH-SIMS. Specifically, we assessed the model's performance across a range of metrics, including MAE, Corr, Acc7, Acc2, and F1 scores. These metrics quantify different aspects of performance for multi-modal sentiment analysis tasks. The results are summarized in Table 2 and 3.

We observed that for the CMU-MOSI dataset, the proposed model achieved the best performance on four metrics: MAE, Corr, Acc7, and Acc2 (negative/positive). While our model showed comparable performance to DTN on the Corr metric, it consistently outperformed DTN across all other evaluation criteria. Notably, our method achieved improvements of 0.2% in the Acc7 metric and 0.23% in the Acc2 (negative/positive) metric compared to DTN.

For the CMU-MOSEI dataset, the proposed model achieved the best results so far on Acc2 (negative/non-negative) and the corresponding F1 scores. Compared to the AOBERT method, which currently holds the highest published Acc2 for negative/non-negative classification, our method achieved an Acc2 that was 0.68% higher. Additionally, our F1 score for negative/non-negative classification surpassed AOBERT by 0.54%. Compared to DTN, the

TABLE 2. Performance of sentiment analysis on the CMU-MOSI and CMU-MOSEI datasets. For Acc2 and F1 score, the number on the left of the "/" corresponds to "negative/non-negative" classification and the right denotes "negative/positive" classification. †represents the result is from [35] and its corresponding GitHub page.[2]

Model	CMU-MOSI					CMU-MOSEI				
	MAE	Corr	Acc7 (%)	Acc2 (%)	F1 (%)	MAE	Corr	Acc7 (%)	Acc2 (%)	F1 (%)
TFN [1]	0.901	0.698	-	-/80.8	-/80.7	0.593	0.700	-	-/82.5	-/82.1
LMF [4]	0.917	0.695	-	-/82.5	-/82.4	0.623	0.677	-	-/82.0	-/82.1
BBFN [2]	0.776	0.755	45.0	-/84.3	-/84.3	0.529	0.767	54.8	-/86.2	-/86.1
Self-MM † [22]	0.708	0.796	46.7	83.44/85.46	83.13/85.33	0.531	0.765	53.9	83.76/85.15	83.82/84.90
MISA [17]	0.783	0.761	42.3	81.3/83.4	81.7/83.6	0.555	0.756	52.2	83.6/85.5	83.8/85.3
CubeMLP [5]	0.770	0.767	45.5	-/85.6	-/85.5	0.529	0.760	**54.9**	-/85.1	-/84.5
TETFN † [9]	0.708	0.798	45.8	83.24/85.37	83.13/85.33	0.537	0.770	53.9	84.12/86.21	84.35/86.11
HFNGC [31]	0.706	0.800	46.2	-/84.40	-/84.27	0.543	0.755	51.5	-/84.10	-/84.19
AOBERT [32]	0.856	0.700	40.2	**85.2**/85.6	**85.4/86.4**	**0.515**	0.763	54.5	84.9/86.2	85.0/85.9
MIM [33]	0.734	0.785	-	84.11/86.28	84.44/86.37	0.549	0.768	-	83.88/85.44	83.83/85.71
DTN [34]	0.714	**0.807**	48.1	-/86.2	-/86.2	0.579	**0.788**	52.5	-/86.3	-/**86.3**
GFNPME	**0.680**	**0.807**	**48.3**	84.43/**86.43**	84.30/86.39	0.528	0.771	53.9	**85.58/86.52**	**85.54**/86.26

current SOTA on Acc2 for negative/positive classification, our method outperformed it by 0.22% on this metric.

Additionally, our method has achieved SOTA results across all metrics on the CH-SIMS dataset. In comparison with TETFN, our method demonstrated substantial improvements. It achieved a 1.97% increase in Acc2 and a 2.86% enhancement in the corresponding F1 score.

The experimental results indicate that GFNPME demonstrates competitive performance across multiple datasets.

TABLE 3. Performance of sentiment analysis on the CH-SIMS dataset.

Model	CH-SIMS			
	MAE	Corr	Acc2 (%)	F1 (%)
TFN †	0.437	0.582	77.07	76.94
LMF †	0.438	0.578	77.42	77.35
MISA †	0.447	0.563	76.54	76.59
Self-MM †	0.425	0.595	80.04	80.44
TETFN †	0.420	0.577	81.18	80.24
GFNPME	**0.402**	**0.616**	**83.15**	**83.10**

2) ABLATION STUDIES

In this section, we conducted experiments by removing the model components individually and compared their performance against the complete model, which demonstrated their respective contributions. Furthermore, we designed experiments involving varying fusion orders to assess the merits of the progressive modality enhancement approach.

a: IMPACT OF EACH MODEL COMPONENT

To validate the effects of all components in the GFNPME model, we conducted extensive ablation studies on the CMU-MOSI dataset, including the complete GFNPME model and variant models with certain components removed. The evaluated models included the following.

- The GFNPME model, the complete gated fusion model with progressive modality enhancement.
- The GFNPME w/o PME, the GFNPME model without progressive modality enhancement approach, i.e., all modalities are enhanced concurrently.
- The GFNPME w/o MIAU, the GFNPME without the multi-modal information aggregation unit, i.e., the two input modalities are simply concatenated.
- The GFNPME w/o MIILs, the GFNPME model without the stacked multi-modal information interaction layers, i.e., visual, acoustic, and textual features are simply concatenated to form multi-modal feature.
- The GFNPME w/o FiLM, the GFNPME model without the FiLM module, i.e., after the MIILs, the features are concatenated together for further prediction.
- The GFNPME w/o singleTs, the GFNPME model without single-modal tasks.

TABLE 4. Performance of ablation experiments of model components on the CMU-MOSI dataset.

Model	CMU-MOSI				
	MAE	Corr	Acc7 (%)	Acc2 (%)	F1 (%)
GFNPME	**0.68**	**0.81**	**48.3**	**84.4/86.4**	**84.3/86.4**
-PME	0.77	0.79	44.5	83.1/84.0	83.1/84.1
-MIAU	0.74	0.77	43.4	83.2/85.1	83.2/85.1
-MIILs	0.73	0.79	46.4	83.1/84.9	83.0/84.9
-FiLM	0.70	0.80	48.3	82.4/83.5	82.4/83.6
-singleTs	0.74	0.79	44.3	82.4/84.0	82.3/84.0

TABLE 5. The impact of different modules on model training time on the MOSI dataset; "s" denotes seconds.

Model	GFNPME	-MIILs	-FiLM	-singleTs
Time(s)	166	85	103	87

Experimental results of ablation studies on model components are listed in Table 4. Symbol "-" in the table indicates "GFNPME w/o". The performance of these variant models highlighted the role of each module within the proposed model. From the results, we observe that, by the metric of

[2]https://github.com/thuiar/MMSA/tree/master

MAE, the most effective module in the GFNPME was the progressive modality enhancement approach, which boosted the performance of the GFNPME by 0.09. Treating all modalities equally resulted in a decrease in accuracies across metrics, with Acc7 experiencing the most significant drop of 3.8%.

TABLE 6. Performance of various modal fusion orders on the CMU-MOSI dataset.

Order	CMU-MOSI				
	MAE	Corr	Acc7 (%)	Acc2 (%)	F1 (%)
TVA	**0.68**	**0.81**	**48.3**	**84.4/86.4**	**84.3/86.4**
TAV	0.73	0.79	46.2	82.9/84.3	83.0/84.4
VTA	0.73	0.79	45.6	82.9/84.9	82.9/84.9
VAT	0.72	0.80	43.7	83.7/85.1	83.7/85.1
ATV	0.71	0.79	45.5	82.2/84.0	82.2/84.0
AVT	0.71	0.80	45.0	82.5/84.6	82.4/84.5

TABLE 7. Performance of sentiment analysis from single modalities on the CMU-MOSI dataset.

Modality	CMU-MOSI				
	MAE	Corr	Acc7 (%)	Acc2 (%)	F1 (%)
T	0.93	0.67	36.4	79.6/81.4	79.6/81.4
V	1.43	0.13	17.4	53.2/52.6	52.2/51.9
A	1.45	0.16	17.5	51.0/49.4	45.3/43.8

Single-modal task incorporation also played a crucial role in the proposed GFNPME model, enhancing the performance by 4.0% according to Acc7. Based on the F1 score, by introducing three single-modal tasks and formulating the problem as a multi-task learning problem, the GFNPME model improved the performance by 2.0% and 2.4% for negative/non-negative and positive/negative classification, respectively. This suggested that optimizing the entire model using only multi-modal labels from fused features might hinder the capture of modality-specific features, thus deteriorating the overall performance.

The introduction of the MIAU significantly boosted the performance of the GFNPME model. Directly merging information from two modalities (i.e., GFNPME w/o MIAU) led to a substantial drop of 4.9% compared to the classification accuracy of the complete model (48.3%) for the Acc7 task. This underscores the importance of the information aggregation module for achieving accurate sentiment classification, especially for finer-grained categories.

The proposed MIILs also have a great impact on model accuracy, resulting in an improvement of 1.9% on the Acc7 task. However, we observe that removing certain components from the model's interaction layers causes a greater decrease in performance than removing all interaction layers. This indicates that only a specific configuration of the fusion process can effectively combine multi-modal information, while a randomly designed interaction scheme might have a negative impact, causing modalities with initially distinguishable sentiments to lose decisive cues during the fusion process.

To investigate the impact of different modules on model training time, we conducted experiments on several components within the model, with the results presented in Table 5. Removal of the MIILs reduced the training time from 166 seconds (GFNPME) to 85 seconds, resulting a decrease of 81 seconds. This indicates that the MIILs introduces a certain level of computational complexity during the process of multi-modal information interaction.

Removal of the FiLM reduced the training time to 103 seconds. Notably, eliminating the MIILs saved an additional 18 seconds compared to removing the FiLM. Given the significant contributions of the MIAU and the progressive modality enhancement strategy within the MIILs to the model's performance, we believe this additional computational cost is justifiable.

b: IMPACT OF THE MODAL PRIORITY

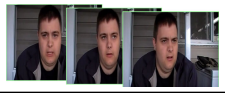
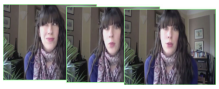
To validate the effectiveness of prioritizing the dominant modality and guiding the fusion process with it, we conducted experiments with six different fusion orders, including sequentially processing textual, visual, and acoustic modalities (TVA), sequentially processing the textual, acoustic and visual modalities (TAV), etc.

Results of the ablation experiments are listed in Table 6. The results demonstrated that the textual modality priority scheme yielded more accurate performance compared to other schemes. For example, it can be observed that when the relative positions of the visual and acoustic modalities were fixed, and the visual modality was processed prior to the acoustic modality, i.e., the processing orders were TVA, VTA, and VAT, Acc7 dropped as the processing order of the textual modality decreased. Specifically, when the textual modality was processed last (VAT), Acc7 was only 43.7%, with a margin of 4.6% lower than the best-performing configuration of TVA. When the processing order of the textual modality was fixed, i.e., the orders were TVA, TAV, VTA, ATV, AVT, and VAT, prioritizing the processing of the visual modality generally performed better than prioritizing the processing of the acoustic modality.

From these experiments, we concluded that within our study's experimental settings, prioritizing the textual modality played the most crucial role in the sentiment analysis process, effectively enhancing the performance of the multi-modal sentiment analysis task. Furthermore, prioritizing visual modalities proved more influential than prioritizing acoustic modalities.

The argument for prioritizing textual modality, followed by visual and acoustic modalities sequentially, is further supported by single-modality sentiment analysis experiments. We conducted sentiment analysis experiments using textual, visual, and acoustic modalities separately. Table 7 presents

TABLE 8. Case study of the modal priority strategy for sentiment analysis.

Case No.	Multi-modal Inputs			GT	Predictions			
	Text	Video	Audio		ATV	VTA	TETFN [9]	GFNPME
Case 1	Sigh it probably would have helped if i went with someone if at least one other one other person went with me		(Hesitant tone)	-0.6	0.1	0.1	-0.3	-0.6
Case 2	Where the problem with that lies not entirely sure		(Impatient tone)	-1.2	-0.9	-0.9	-1.1	-1.2
Case 3	I think um the movies are phenomenal especially that first one		(Emphasized tone)	2.8	2.2	2.2	2.3	2.6
Case 4	Um the people seemed sort of dead		(Emphasized tone)	-1.6	-1.4	-1.3	-1.5	-1.5
Case 5	This credit sequence i think is the best credit sequence um since watchman		(Mildly interested tone)	2.4	2.1	2.0	2.3	2.4
Case 6	Marlon Wayans had a lot of fun.		(Calm expression)	1.4	1.4	1.5	1.1	1.4

the results of these experiments. From the table, we observe that the textual modality achieved the best performance compared to the other two modalities. Additionally, the visual modality generally performed better than the acoustic modality. These observations align with the results from our previous ablation studies.

To further demonstrate the effectiveness of the progressive modality enhancement approach for sentiment analysis, we compared it with the text-centered method TETFN, the visual prior (VTA) and the acoustic prior (ATV) strategies on six examples. The experimental results are illustrated in Table 8. The contents in parentheses represent our subjective understanding of the cases. In the first and second examples, while the textual information might not overtly convey sentiment, the speaker's visual expression and acoustic cues clearly expressed sentiment polarity. The predicted values of GFNPME align closely with the ground truth, whereas TETFN correctly predicted the polarity of the samples but showed larger deviations in its predicted values compared to the ground truth. In the third and fourth examples, all modalities exhibited distinctive sentiment cues. In case 4, both GFNPME and TETFN achieved comparable performance scores. However, in case 3, GFNPME's predicted value demonstrated superior accuracy, aligning more closely with the ground truth compared to other methods. For the fifth and sixth examples, when the textual modality contained explicit sentiment-laden language, unsurprisingly, the GFNPME aligned closer with the ground truth of the samples. Across all six cases, the GFNPME demonstrated superior performance. These results substantiate that modal priority in our progressive modality enhancement approach effectively facilitates the integration of the three modalities.

V. CONCLUSION AND FUTURE WORKS

In this study, we proposed a gated fusion network with progressive modality enhancement for sentiment analysis. Incorporated a novel progressive modality enhancement approach to solve the problem of uneven distribution of sentiment, and proposed a gated multi-modal information interaction module to ensure comprehensive interaction among the three modalities during the interaction process.

Quantitative results demonstrate that the proposed approach achieves competitive results on the three benchmark datasets CMU-MOSI, CMU-MOSEI and CH-SIMS. In particular, on the CH-SIMS dataset, our method attained an Acc2 of 83.15% and an F1 score of 83.10%, which represented significant improvements of 1.97% and 2.68%, respectively, compared to the next best method. Through comprehensive ablation studies, we demonstrated the effectiveness of each component module. Notably, removing the progressive modality enhancement approach decreased Acc7 by 3.8%, while eliminating the multi-modal information aggregation unit caused a more substantial 4.9% reduction in this metric. The ablation experiments with varying fusion sequences and single-modality sentiment analysis experiments indicated that the visual modality generally performed better than the acoustic modality in the MSA tasks across our experimental settings. Additionally, a case study comparing our progressive modality enhancement approach with various fusion sequences and existing methods further substantiates its efficacy.

In future work, we will explore the application of gated fusion networks with progressive modality enhancement to other multi-modal tasks, and continue to investigate more streamlined and computationally efficient methods for multi-modal interactions.

REFERENCES

[1] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Copenhagen, Denmark, M. Palmer, R. Hwa, and S. Riedel, Eds., 2017, pp. 1103–1114. [Online]. Available: https://aclanthology.org/D17-1115

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

[3] P. Du, Y. Gao, L. Li, and X. Li, "SGAMF: Sparse gated attention-based multimodal fusion method for fake news detection," *IEEE Trans. Big Data*, early access, Jun. 13, 2024, doi: 10.1109/TBDATA.2024.3414341.

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

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

[6] L. Zhao, Y. Yang, and T. Ning, "A three-stage multimodal emotion recognition network based on text low-rank fusion," *Multimedia Syst.*, vol. 30, no. 3, p. 142, Jun. 2024.

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

[8] 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.*, vol. 34, no. 5, 2020, pp. 8992–8999.

[9] D. Wang, X. Guo, Y. Tian, J. Liu, L. He, and X. Luo, "TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis," *Pattern Recognit.*, vol. 136, Apr. 2023, Art. no. 109259.

[10] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," *Inf. Fusion*, vol. 91, pp. 424–444, Mar. 2023.

[11] Q. Qi, L. Lin, R. Zhang, and C. Xue, "MEDT: Using multimodal encoding-decoding network as in transformer for multimodal sentiment analysis," *IEEE Access*, vol. 10, pp. 28750–28759, 2022.

[12] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Doha, Qatar, 2014, pp. 1724–1734. [Online]. Available: https://aclanthology.org/D14-1179

[13] P. Du, Y. Gao, and X. Li, "Bi-attention modal separation network for multimodal video fusion," in *Proc. Int. Conf. Multimedia Modeling*. Cham, Switzerland: Springer, 2022, pp. 585–598.

[14] S. Mai, S. Xing, J. He, Y. Zeng, and H. Hu, "Multimodal graph for unaligned multimodal sequence analysis via graph convolution and graph pooling," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 19, no. 2, pp. 1–24, May 2023.

[15] L. Xiao, X. Wu, W. Wu, J. Yang, and L. He, "Multi-channel attentive graph convolutional network with sentiment fusion for multimodal sentiment analysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 4578–4582. [Online]. Available: https://api.semanticscholar.org/CorpusID

[16] J. Huang, Y. Pu, D. Zhou, J. Cao, J. Gu, Z. Zhao, and D. Xu, "Dynamic hypergraph convolutional network for multimodal sentiment analysis," *Neurocomputing*, vol. 565, Jan. 2024, Art. no. 126992, doi: 10.1016/j.neucom.2023.126992.

[17] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-invariant and -specific representations for multimodal sentiment analysis," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1122–1131. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/3394171.3413678

[18] D. S. Chauhan, A. Ekbal, and P. Bhattacharyya, "Sentiment and emotion help sarcasm? A multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 4351–4360.

[19] Y. Zhang and Q. Yang, "An overview of multi-task learning," *Nat. Sci. Rev.*, vol. 5, no. 1, pp. 30–43, Jan. 2018.

[20] 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.

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

[22] 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.*, vol. 35, no. 12, 2021, pp. 10790–10797.

[23] Y. Zhang, J. Wang, Y. Liu, L. Rong, Q. Zheng, D. Song, P. Tiwari, and J. Qin, "A multitask learning model for multimodal sarcasm, sentiment and emotion recognition in conversations," *Inf. Fusion*, vol. 93, pp. 282–301, May 2023.

[24] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/11671

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, J. Burstein, C. Doran, and T. Solorio, Eds., Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[26] 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, Nov. 2016.

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

[28] 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)*, May 2014, pp. 960–964.

[29] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "Librosa: Audio and music signal analysis in Python," in *Proc. Python Sci. Conf.*, 2015, pp. 18–24.

[30] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 59–66.

[31] T. Zhao, J. Peng, Y. Huang, L. Wang, H. Zhang, and Z. Cai, "A graph convolution-based heterogeneous fusion network for multimodal sentiment analysis," *Appl. Intell.*, vol. 53, no. 24, pp. 30455–30468, Dec. 2023.

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

[33] Y. Luo, R. Wu, J. Liu, and X. Tang, "Balanced sentimental information via multimodal interaction model," *Multimedia Syst.*, vol. 30, no. 1, pp. 1–9, Feb. 2024.

[34] Y. Zeng, W. Yan, S. Mai, and H. Hu, "Disentanglement translation network for multimodal sentiment analysis," *Inf. Fusion*, vol. 102, Feb. 2024, Art. no. 102031.

[35] H. Mao, Z. Yuan, H. Xu, W. Yu, Y. Liu, and K. Gao, "M-SENA: An integrated platform for multimodal sentiment analysis," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics, Syst. Demonstrations*, V. Basile, Z. Kozareva, and S. Stajner, Eds., Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 204–213. [Online]. Available: https://aclanthology.org/2022.acl-demo.20

FENG XIN received the B.S. degree from Qilu University of Technology, China, in 2022. He is currently pursuing the master's degree with Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China). His research interests include deep learning and multi-modal sentiment analysis.

WENJUAN GONG received the Ph.D. degree (cum laude) from Universitat Autònoma de Barcelona, Barcelona, in 2013. She was a Postdoctoral Research Assistant with Oxford Brookes University, Oxford, in 2014. She is currently an Associate Professor with China University of Petroleum (East China). Her research interests include computer vision and machine learning.

TINGBO SHI received the bachelor's degree from the College of Computer Science and Technology, Nanjing Audit University, in 2021. He is currently pursuing the master's degree with China University of Petroleum (East China). His research interests include machine learning and federated semi-supervised learning.

JING ZHONG is currently pursuing the degree with the College of Computer Science and Technology, China University of Petroleum (East China). Her research interest includes multi-modal sentiment analysis.

KECHEN LI is currently pursuing the degree in computer science and technology with China University of Petroleum (East China). His research interests include machine learning, multimodal algorithms, and federated learning.

JORDI GONZÀLEZ (Member, IEEE) received the Ph.D. degree in computer engineering from Universitat Autònoma de Barcelona (UAB), Barcelona, in 2004. He is currently a Professor of computer science with the Department of Computer Science, UAB. He is also a Research Fellow with the Computer Vision Center, where he has cofounded three spin-offs, namely Cloud Size Services, Visual Tagging, and Care Respite, and the Image Sequence Evaluation (ISE Lab) Research Group. His research interest includes machine learning techniques for the computational interpretation of social images or visual hermeneutics.