Information Processing and Management 61 (2024) 103675

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

A cross modal hierarchical fusion multimodal sentiment analysis method based on multi-task learning

Lan Wang [a], Junjie Peng [a,b,*], Cangzhi Zheng [a], Tong Zhao [a], Li'an Zhu [a]

[a] School of Computer Engineering and Science, Shanghai University, Shanghai, China
[b] Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai, China

ARTICLE INFO

Keywords:
Multi-modality
Sentiment analysis
Cross-modal interactions
Multi-task learning

ABSTRACT

Humans often express affections and intentions through multiple forms when communicating, involving text, audio, and vision modalities. Using a single modality to determine the sentiment state may be biased, but combining multiple clues can fully explore more comprehensive information. Effective fusion of heterogeneous data is one of the core problems of multimodal sentiment analysis. Most cross-modal fusion strategies inevitably bring noisy information, resulting in low-quality joint feature representations and impacting the accuracy of sentiment classification. Considering the unique cues of modality-specific, common information between modalities, and sentiment variability among different layers, we introduce multi-task learning and propose a cross-modal hierarchical fusion method for multimodal sentiment analysis. The model combines unimodal, bimodal, and trimodal tasks to enhance multimodal feature representation for the final sentiment prediction. We conduct extensive experiments on CH-SIMS, CMU-MOSI, and CMU-MOSEI, where the first one is in Chinese and the last two are in English. The results demonstrate the generalizability of the proposed method. It effectively improves the accuracy of sentiment analysis while reducing the adverse impact of the noise compared to the existing models.

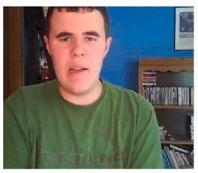
1. Introduction

Sentiment analysis is one of the typical research areas in artificial intelligence, mainly categorized into text sentiment analysis, audio sentiment analysis, and video sentiment analysis. These three sentiment analysis tasks process and analyze textual, acoustic, and facial expression data, respectively, and then predict the corresponding sentiment states (Chen, Peng et al., 2023; Lin et al., 2023; Zhao et al., 2023). In most real-world speech scenarios, people express feelings in multiple forms. They detect the sentiment tendency of the other person based on the presentation of the facial expressions, perceived intonation of the voice, and the understanding of the semantic information in the text. Thus, it is significant and applicable to use the information from text, audio, and vision to identify sentiment.

Multimodal sentiment analysis mainly uses two or more modalities to recognize the actual sentiment tendency of a person, involving text, audio, vision, electroencephalogram (EEG) signals and so on. Sentiment classification broadly categorizes sentiment tendency into three types: positive, neutral, and negative. Specifically, these three types can be divided into more fine-grained classifications based on sentiment intensity (Sun et al., 2023). Currently, for unimodality, there are many methods to obtain stable classification results (Lu et al., 2023). However, it is difficult to mine more comprehensive information using a single modality in

* Corresponding author at: School of Computer Engineering and Science, Shanghai University, Shanghai, China.
E-mail addresses: wanglan1997@shu.edu.cn (L. Wang), jjie.peng@shu.edu.cn (J. Peng), cangzhizheng@shu.edu.cn (C. Zheng), zhaotong@shu.edu.cn (T. Zhao), blossom@shu.edu.cn (L. Zhu).

https://doi.org/10.1016/j.ipm.2024.103675
Received 23 July 2023; Received in revised form 21 January 2024; Accepted 24 January 2024
Available online 29 January 2024
0306-4573/© 2024 Elsevier Ltd. All rights reserved.



Utterance: "and a lot of people have been saying that."

Text	Audio	Vision	Sentiment
Comment	Low	Frown	Negative

Fig. 1. An example of analysis of text, audio, and vision.

some cases. An example of the analysis of three modalities in video is illustrated in Fig. 1. The person in the video makes a comment 'and a lot of people have been saying that'. It is not easy to infer the speaker's sentiment state based on the text modality alone. However, considering the frowning and low voice of the person in the video, it can be judged that the speaker's sentiment polarity is negative. Thus, combining multiple complementary information can improve the accuracy of sentiment classification. Multimodal sentiment analysis is gradually coming into the limelight as a new research field, aiming to break through the limitations of unimodal sentiment analysis. With the rapid popularity of smartphones and the emergence of social media platforms such as TikTok, YouTube, and Little Red Book, users tend to upload videos online to record their daily lives and discuss hot topics. As a result, multimodal sentiment analysis has become a hot research area with the proliferation of video data.

Feature extraction and cross-modal fusion are the two challenges in multimodal sentiment analysis. For feature extraction, many methods have been proposed by researchers, such as Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Convolutional Neural Network (CNN), and attention mechanism (Brady et al., 2016; Cho et al., 2014; Lim et al., 2016; Vaswani et al., 2017). For cross-modal fusion, it can be broadly classified into feature-level fusion, decision-level fusion, and hybrid fusion. On the whole, the results of the multimodal sentiment analysis highly depend on whether the cross-modal fusion strategies are reasonable and efficient or not. Thus, cross-modal fusion has become one of the critical issues in multimodal sentiment analysis.

Video-based multimodal sentiment analysis often involves text, audio, and vision modalities. Combining the three modalities provides a thorough understanding the sentiment tendency of the person in the video. Due to the heterogeneity of the data, text modality provides higher-order semantic information compared with audio and vision modalities. Many works use cross-modal fusion strategies to capture correlation information between modalities to obtain a more efficient joint feature representation. Due to the consistency and difference between modalities, noisy information is inevitably generated in the fusion process. Multi-Task Learning (MTL) is a promising method to eliminate the negative effects of cross-modal interactions. MTL-based multimodal sentiment analysis method mainly utilizes unimodal prediction tasks to assist the mainline task and introduces multiple objective loss functions to improve the robustness of the model. The specific information provided by each modality is crucial to the sentiment analysis. However, there is a limitation to narrowing down the differences among modalities with the help of unimodal prediction tasks alone. The reason is that the information provided by the bimodal modality cannot be ignored for the final decision. Conflicting information can be introduced during cross-modal interactions since consistency and differences of the affective cues are provided by the unimodal and bimodal. To tackle the above-mentioned problems, we introduce multi-task learning and propose a cross-modal hierarchical fusion method for multimodal sentiment analysis. The method can mine the independence and correlation information among different levels of modalities. The main contributions of this paper are listed below:

- To eliminate the differences among modalities, a cross-modal hierarchical fusion model for multimodal sentiment analysis is proposed. It improves the performance of sentiment analysis through a multi-tasking assistance module.
- Considering that the cross-modal fusion inevitably brings conflicting information, we incorporate the relationship among the unimodal, bimodal, and trimodal to improve the robustness of the model.
- We conduct extensive experiments on both English and Chinese datasets to demonstrate the correctness and effectiveness of the proposed model.

2. Related work

Multimodal sentiment analysis primarily makes use of external and internal characteristics to identify one's true sentiment. Video-based multimodal sentiment analysis task involves multiple heterogeneous data such as text, audio, and vision. There are two major challenges that can be focused on for multimodal sentiment analysis; one is unimodal representation learning, and the other is multimodal fusion strategies (Gandhi et al., 2023).

Unimodal representation learning has incurred extensive work that focuses on modeling textual, acoustic, and visual data. The sentence vectors obtained by the large nature language pre-training BERT are highly desirable. Open-source toolkits such as LibROSA (McFee et al., 2015) and OpenSMILE (Eyben et al., 2013) are often used to extract acoustic features such as rhythmic features, spectral-based features, sound quality features, and so on. The OpenFace toolkit is proposed by Baltrušaitis et al. (2016) to execute vision modality and extract facial features. Many multimodal fusion algorithms have been developed by researchers based on feature concatenation (You et al., 2015), memory networks (Zadeh, Liang, Mazumder et al., 2018), etc. However, the methods mentioned above do not capture the deeper interaction information among modalities well.

Benefiting from the excellent performance of the Transformer, some researchers have introduced it to fusion phase of the multimodal sentiment analysis. Li et al. (2023) designed a streamlined method for multimodal sentiment classification. They utilized a single-layer Transformer encoder to extract acoustic, visual, and cross modal information, respectively. Chen, Su et al. (2023) proposed hybrid fusion model based on information relevance for multimodal sentiment analysis. In their model, the attended visual features, mid-level visual representations, high-level features, and attended semantic features are distilled as input of the information relevance classifier. He et al. (2021) designed a unimodal reinforced Transformer module to extract refined unimodal information from multimodal fusion features. They introduced time squeeze fusion to explore the interaction information of the three modalities in the temporal dimension. Motivated by Low-rank Multimodal Fusion (LMF), Sahay et al. (2020) proposed a Transformer-based multimodal framework. In that method, low-rank multimodal fusion and fusion-based Transformer are used to capture the interaction information between unimodal, bimodal, and trimodal modalities. Xu et al. (2020) mined different levels of sentiment information via multi-head attention mechanism. Jiang et al. (2020) proposed a Fusion-Extraction Network (FENet) for multimodal sentiment classification. The method introduces a fine-grained attention mechanism to learn bidirectional interactive information between text and image. Wu et al. (2022) considered that cues in bimodal are crucial for sentiment analysis. They used three multi-headed attention to explore the relative importance between two pair-wise modalities. Xue et al. (2020) designed a multi-tensor fusion method with hybrid attention architecture. In that method, multi-head attention, cross-modal attention, and multi-tensor fusion are introduced to calculate the joint feature representation. Lai and Yan (2022) proposed an asymmetric windows mechanism to focus on contextual information with different timestamps. The method uses text, audio, vision, and the concatenation of three modalities as input and introduces multiple attention to capture intra-modality and inter-modality sentiment information. Kumar and Vepa (2020) leveraged self-attention to obtain contextual information about each modality and passed through cross-attention module to learn the interaction information between different modalities. Tzirakis et al. (2021) introduced Transformer-based module to extract embeddings of the original text and then used residual attention to explore the correlation between modalities. Xie and Zhang (2020) used the Transformer combined with gated mechanism to close the feature space among modalities to improve the accuracy of multimodal sentiment analysis. Huddar et al. (2020) computed a bimodal attention matrix on the basis of unimodal attention weights and bimodal contribution values. Furthermore, the trimodal attention matrix is obtained from residual attention network.

The above-mentioned methods mainly focus on cross-modal fusion and pay less attention to considering the conflict information that can be brought by the inter-model and bimodal interactions. Zhang et al. (2020) noticed that noises exist in text modality can affect the accuracy of multimodal sentiment classification. Therefore, they designed a denoising autoencoder (DAE) to augment its representation capacity and introduced cross-feature-fusion module based on the attention mechanism to explore the complementary and comprehensive information. Multitask learning as an effective paradigm has been widely used in many field (Sener & Koltun, 2018). Multimodal sentiment analysis based on multi-task learning can introduce subtasks to mitigate the negative effects of conflicting information. Yang et al. (2022) presented Two Phase Multi-task method for sentiment analysis, which uses fine-tuned BERT to obtain textual contextual features. In their method, the information-enhanced audio and vision features are obtained from unidirectional cross modality Transformer of the text. Considering that the sentiment analysis task and emotion recognition task are highly related, Akhtar et al. (2019) proposed a multi-task framework based on contextual information. Yu et al. (2021) designed a unimodal label generation module based on self-supervised learning and introduced adaptive weights to balance the variability between modalities. Peng et al. (2023) manually labeled unimodal and bimodal labels on two public datasets to explore the impact of fine-grained labels for sentiment prediction. Followed by, they proposed a multi-stage model for multimodal sentiment analysis. In addition, the model introduces six unimodal tasks and three bimodal tasks to assist joint feature learning. Zhang et al. (2019) proposed an end-to-end speech emotion recognition network. It contains three emotion-related tasks of arousal, valence, and dominance prediction. Fang et al. (2022) designed Sense Block, Language Block, and Multimodal Block for multimodal sentiment analysis. They introduced high-level cross-modal attention layer to obtain multimodal feature representations in Multimodal Block. Moreover, text, audio, and vision predictions are used as auxiliary tasks.

3. Methodology

An utterance-level video containing three modalities can be described by $u = \{I_t, I_a, I_v\}$. We represent text, audio, and vision modalities as I_t, I_a, I_v. The goal of the multimodal sentiment analysis issue is to select a suitable method to automatically predict one's sentiment polarity $\hat{y}_{tav} \in \{\text{Sentiment}_1, \text{Sentiment}_2, \ldots, \text{Sentiment}_n\}$. Sentiment polarity can be categorized into n classifications according to intensity.

Multimodal sentiment analysis focuses on unimodal feature learning and cross-modal fusion, where the latter is a more challenging subject. Designing efficient fusion strategies is a promising idea for solving this problem. Nowadays, most fusion strategies mainly explore cross-modal interactions. Using the obtained interaction information, multimodal feature representations are generated for sentiment prediction. Cross-modal interactions continuously reduce the differences between modalities, mapping each modality feature to a shared space. Due to the semantic variability of the three kinds of data, the fusion process brings noisy information and weakens the unique cues associated with sentiment, which affects the final sentiment prediction. To mitigate this problem, we introduce multi-task learning and propose a cross-modal hierarchical fusion model or CMHFM for short for multimodal sentiment analysis.

The overall framework of the model is shown in Fig. 2. The CMHFM contains four modules, i.e., Unimodal Feature Learning (UFL), Inter-Modal Interaction (IMI), Multi-Head Attention (MHA), and Multi-Tasking Assisted Learning (HTAL). We model the depth feature of the unimodal in Section 3.1. Then, we use TFN and multi-head attention to explore cross-modal interactions in Sections 3.2 and 3.3. Finally, we introduce multi-task learning and cross-modal hierarchical fusion to aid the final sentiment prediction in Section 3.4.

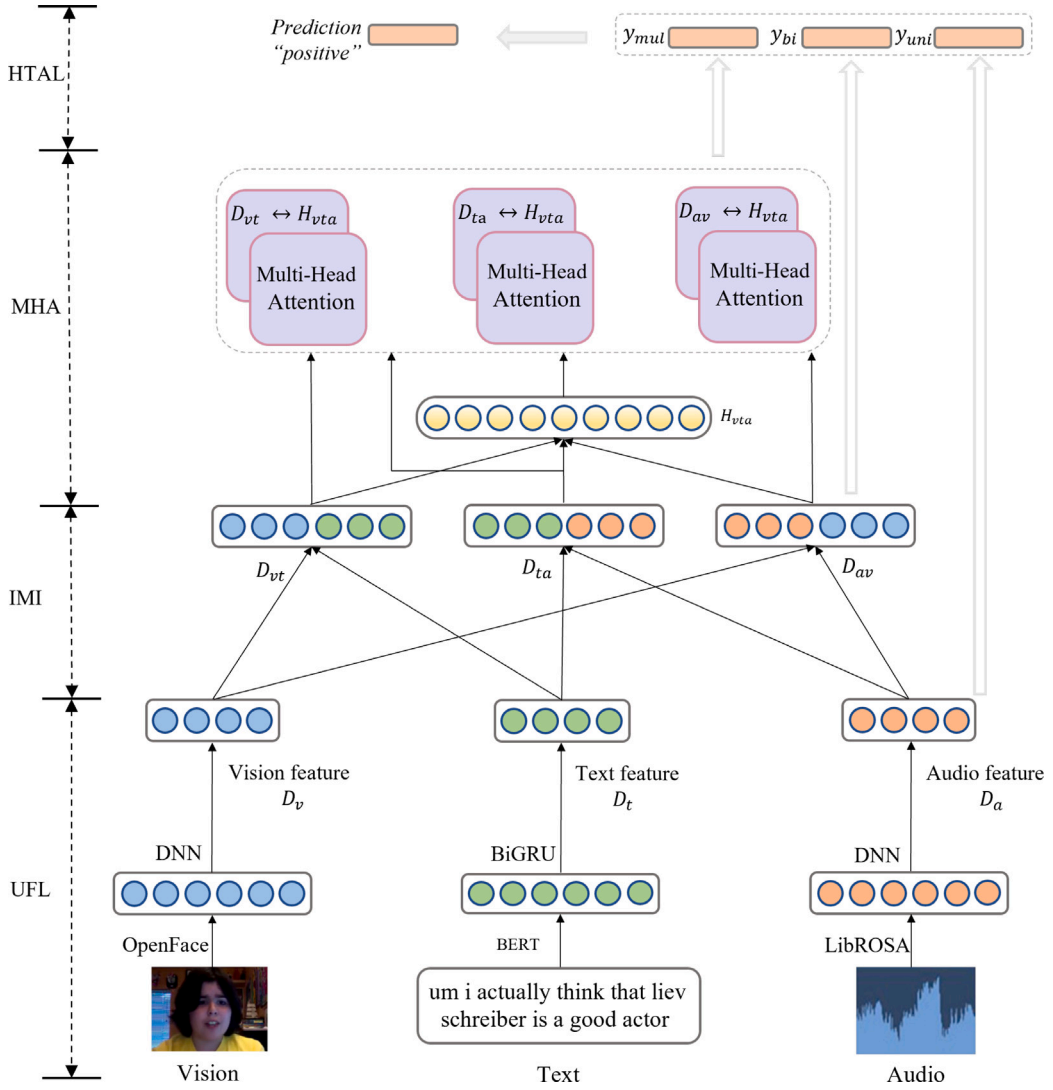


Fig. 2. The overall framework of CMHFM.

3.1. Unimodal feature learning

We assume that a dataset consists of N short video clips, which can be represented by $U = \{u_1, u_2, u_3, \ldots, u_N\}$. Each video clip contains three modalities: text, audio, and vision.

To make the best of the semantic information contained in text modality, we employ manual transcription to obtain text from the video. Since pre-trained BERT has a powerful feature representation, we use it to extract the initial features of the text $X_t \in R^{l_t \times d_t}$, where l_t and d_t denote the sequence length and embedding dimension of the text feature, respectively. Bidirectional Gate Recurrent Unit (BiGRU) can effectively capture contextual information relevance. Therefore, we introduce BiGRU to further mine the depth features of the text D_t, which is calculated as below:

$$D_t = \text{BiGRU}\left(X_t; \theta_{\text{BiGRU}}\right) \tag{1}$$

where θ_{BiGRU} is the parameter of the BiGRU model.

For audio modality, following the previous works (Wu et al., 2022), we use the LibROSA toolkit to mine features such as Mel Frequency Cepstral Coefficients (MFCCs), energy, and pitch. The initial audio features are described by $X_a \in R^{l_a \times d_a}$, where l_a and d_a are the sequence length and embedding dimension of the audio feature, respectively. For vision modality, FFmpeg toolkit is used to extract frames from the video clips, and MTCNN is applied for face recognition and alignment. The 2D keypoint coordinates (in pixels), 3D keypoint coordinates (in millimeters), head pose (position and rotation), facial action units, and eye gaze are obtained by the OpenFace toolkit. Let $X_v \in R^{l_v \times d_v}$ denote the original video features, where l_v and d_v are the sequence length and embedding

dimension of the video feature, respectively. The background noise in the audio modality and the presence of undetected faces in the aligned images are interference information that can adversely affect the accuracy of classification. Deep neural networks (DNN) can tolerate the presence of noisy information to a certain extent. Therefore, we utilize a three-layer DNN to extract the depth features of the audio and video modality, respectively. The equations are shown as follows:

$$\begin{aligned} D_a &= \mathrm{DNN}\left(X_a; \theta_a\right) \\ D_v &= \mathrm{DNN}\left(X_v; \theta_v\right) \end{aligned} \tag{2}$$

where θ_a and θ_v are parameters of the DNN network.

3.2. Inter-modal interaction

After the BERT model and DNN network, we obtain the depth feature representation of text, audio, and vision as D_t, D_a, and D_v, respectively. Cross-modal fusion learns interaction information between text, audio, and vision modalities. We introduce TFN to detect inter-modal interactions. TFN is not only effective in capturing useful information between modalities, but also simple to operate. We perform Cartesian product operation on the feature matrices of two modalities. The text-audio feature vectors, audio-vision feature vectors, and vision-text feature vectors are expressed as follows:

$$\begin{aligned} D_{vt} &= D_v \otimes D_t, D_{vt} \in R^{d_{vt}} \\ D_{ta} &= D_t \otimes D_a, D_{ta} \in R^{d_{ta}} \\ D_{av} &= D_a \otimes D_v, D_{av} \in R^{d_{av}} \end{aligned} \tag{3}$$

3.3. Multi-head attention

Attention mechanism can focus on information-intensive features. The attention mechanism tends to concern the current feature subspace and does not assign weights from the global perspective. The multi-head attention mechanism maps the input feature into different subspaces thereby mining more comprehensive knowledge. The attention mechanism uses the query (Q) and the key (K)-value (V) to calculate the weight. Therefore, the equation is shown as follows:

$$\mathrm{Attention}(Q, K, V) = \mathrm{Softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \cdot V \tag{4}$$

The feature matrix of each attention head is obtained by multi-head processing. The value of the ith head is calculated as follows:

$$\mathrm{head}_i = \mathrm{Attention}\left(Q_i, K_i, V_i\right) = \mathrm{Softmax}\left(\frac{Q_i \cdot K_i^T}{\sqrt{d_{k_i}}}\right) \cdot V_i \tag{5}$$

where d_{k_i} is the dimension of K_i. Q_i, K_i, and V_i denote the matrices of Q, K, and V mapping to the ith feature space, respectively. The calculation of Q_i, K_i, and V_i are defined as follows:

$$\begin{aligned} Q_i &= Q \cdot W_{Q_i} \\ K_i &= K \cdot W_{K_i} \\ V_i &= V \cdot W_{V_i} \end{aligned} \tag{6}$$

where W_{Q_i}, W_{K_i}, W_{V_i} denote the mapping matrix.

The value of multi-head attention is obtained by cascading the features of all head outputs, which is computed as follows:

$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}\left(\mathrm{head}_1, \mathrm{head}_2, \ldots, \mathrm{head}_h\right) \cdot W_o \tag{7}$$

where W_o is the training parameter. We represent the number of multi-head with h. Concat($\cdot$) indicates the value that conjoins all attention heads.

The multi-head attention mechanism captures key cues from different mapping subspaces to obtain a richer feature representation. Therefore, we introduce multi-head attention mechanism to further explore the importance information among pair-wise modalities. Grounded on the above principles, we take bimodal fusion features D_{ta}, D_{av}, and D_{vt} as Q, take the global feature H_{vta} as K and V. The equation for H_{vta} is determined as below:

$$H_{vta} = \mathrm{Concat}\left(D_{ta}, D_{av}, D_{vt}\right) \in R^{3d} \tag{8}$$

Take the text-audio feature vector D_{ta} as a demonstration, we map D_{ta} and H_{vta} to h feature spaces:

$$\begin{aligned} D_{H_1}^i &= H_{vta} \cdot W_K^i \\ D_{H_2}^i &= H_{vta} \cdot W_V^i \\ D_{ta}^i &= D_{ta} \cdot W_Q^i \end{aligned} \tag{9}$$

Table 1
The basic information of CH-SIMS, CMU-MOSI, and CMU-MOSEI.

Dataset	Reference	Language	Train	Valid	Test	Total
CMU-MOSI	Zadeh et al. (2016)	English	1 284	229	686	2 199
CMU-MOSEI	Zadeh, Liang, Poria et al. (2018)	English	16 326	1871	4659	22 856
CH-SIMS	Yu et al. (2020)	Chinese	1 368	456	457	2 281

where $D^i_{H_1}$ and $D^i_{H_2}$ are the inputs of K and V corresponding to the ith feature space. W^i_K, W^i_V, and W^i_Q mean the weight parameters. We denote the index of different feature subspaces with i.

We use the attention mechanism to mine the correlation between the bimodal fusion feature D_{ta} and global feature H_{vta}. The value of the ith head is formulated as:

$$\text{head}_i = \text{Attention}\left(D^i_{ta}, D^i_{H_1}, D^i_{H_2}\right) = \text{Softmax}\left(\frac{D^i_{ta} \cdot \left(D^i_{H_1}\right)^T}{\sqrt{d^i_{H_1}}}\right) \cdot D^i_{H_2} \tag{10}$$

where $d^i_{H_1}$ is the dissension of $D^i_{H_1}$.

The values of the h heads are cascaded together to obtain the final multi-head attention value. The equation is as follows:

$$M_{ta} = \text{MultiHead}\left(D_{ta}, D_{H_1}, D_{H_2}\right) = \text{Concat}\left(\text{head}_1, \text{head}_2, \ldots, \text{head}_h\right) \cdot W_o \tag{11}$$

The multi-head attention values M_{av} and M_{vt} of bimodal features D_{av} and D_{vt} are calculated according to the method mentioned above, respectively.

3.4. Multi-tasking assisted learning

TFN dynamically learns the important features between modalities, and multi-head attention captures critical information more comprehensively. Thus, we introduce TFN and multi-head attention to explore inter-modal interaction and bimodal interaction, respectively. Cross-modal interactions reduce the differences between modalities, but it brings noise and generates conflicting information that affects the accuracy of sentiment classification. To reduce the impact of negative information, we introduce multi-task learning to improve the robustness of the model. We execute cross-modal hierarchical fusion strategy to complement information from different layers while utilizing unimodal, bimodal, and global tasks to assist the final sentiment analysis. Unimodal fusion F_{uni}, bimodal fusion F_{bi}, and global fusion F_{mul} are calculated as follows:

$$\begin{aligned} F_{\text{uni}} &= \text{Concat}\left(D_t, D_a, D_v\right) \\ F_{\text{bi}} &= \text{Concat}\left(D_{vt}, D_{ta}, D_{av}\right) \\ F_{\text{mul}} &= \text{Concat}\left(M_{vt}, M_{ta}, M_{av}\right) \end{aligned} \tag{12}$$

Unimodal prediction y_{uni}, bimodal prediction y_{bi}, and global prediction y_{mul} are expressed as below:

$$\begin{aligned} y_{\text{uni}} &= W_1 \cdot F_{\text{uni}} + b_1 \\ y_{\text{bi}} &= W_2 \cdot F_{\text{bi}} + b_2 \\ y_{\text{mul}} &= W_3 \cdot F_{\text{mul}} + b_3 \end{aligned} \tag{13}$$

where W_1, W_2, and W_3 are the learnable parameters. b_1, b_2, and b_3 denote the biases.

The final classification is determined jointly by the y_{uni}, y_{bi}, and y_{mul}, which is named y_{fusion}. It is calculated as:

$$y_{\text{fusion}} = W_4 \cdot F_{\text{fusion}} + b_4 \tag{14}$$

where F_{fusion} is cascaded from y_{uni}, y_{bi}, and y_{mul}. W_4 is the learnable parameter, b_4 denotes the biases.

We introduce three subtasks to aid the final sentiment prediction, so the training loss consists of four parts, which is defined as follows:

$$\mathcal{L} = \text{Loss}\left(F_{\text{uni}}\right) + \text{Loss}\left(F_{\text{bi}}\right) + \text{Loss}\left(F_{\text{mul}}\right) + \text{Loss}\left(F_{\text{fusion}}\right) \tag{15}$$

4. Experiments

4.1. Experimental setup

4.1.1. Datasets

We employ three public multimodal sentiment analysis datasets, CMU-MOSI, CMU-MOSEI, and CH-SIMS, where the first two are in English and the last one is in Chinese. The basic information of the three datasets are shown in Table 1.

CMU-MOSI: CMU-MOSI (Zadeh et al., 2016) collects a total of 93 opinion videos from the YouTube website, which are professionally processed into 2199 individual utterance video clips. Each video clip needs to ensure that the voice and face of the

speaker appear simultaneously without any interference from other people. In addition, each video clip must remain for a period of time and is labeled with sentiment value between [−3,3]. Specifically, the labeled values of the seven types are not fixed values, but rather all values within the range belong to the category. We use 1284 utterances as training set, 229 utterances as validation set, and 686 utterances as test set.

CMU-MOSEI: CMU-MOSEI (Zadeh, Liang, Poria et al., 2018) has a significantly higher number of speakers, total video duration, and number of video clips than CMU-MOSI. The dataset has a total of 3228 videos, which are split into 22 856 individual utterances. The dataset has the same labeled values as the CMU-MOSI dataset. In addition, there are 16 326 video clips in the training set, 1871 video clips in the validation set, and 4659 video clips in the test set.

CH-SIMS: CH-SIMS (Yu et al., 2020) collects a total of 60 raw videos from movies, TV shows, and variety shows, which are split into 2281 video clips. Each video clip is labeled with sentiment intensity between [−1,1]. The dataset is divide into training set, validation set, and test set in a 6:2:2 ratio.

For CMU-MOSI and CMU-MOSEI, the aligned and unaligned versions are available. The CH-SIMS only provides the unaligned version. For aligned version, the three modalities in the datasets are aligned according to the time step. The aligned time step is obtained based on the word-level features of the text modality. For unaligned version, the three modalities are not processed based on word-level features, so there is no direct correlation among them. Moreover, the modalities in the aligned datasets have the same feature length. The modalities in the unaligned datasets have different feature lengths.

4.1.2. Baselines

To verify the performance and effectiveness of the proposed model, we compare the performance of our model with that of twelve methods. For the mentioned models, we take three text (I_t), audio (I_a), and vision (I_v) modalities from given utterance-level video as input. The multimodal representation R_{tav} obtained from fusion strategy to predict the finally sentiment polarity $\hat{y}_{tav} \in \{\text{Sentiment}_1, \text{Sentiment}_2, \ldots, \text{Sentiment}_n\}$. Sentiment polarity can be categorized into n classifications based on intensity. Specifically, we broadly categorize the comparison models into three types, which are described as follows:

Neural Network based methods. This type methods can be formalize as: Input: $I_t, I_a, I_v \rightarrow$Neural Network$(I_t, I_a, I_v) \rightarrow$ Prediction$(R_{tav}) \rightarrow$ Output: $\hat{y}_{tav}$.

- EF-LSTM: The EF-LSTM (Williams et al., 2018) concatenates different feature vectors and feeds them into an LSTM to predict the sentiment polarity.
- MFN: The MFN (Zadeh, Liang, Mazumder et al., 2018) combines LSTM and memory mechanism to capture the temporal dependence of modalities.
- DFG: The DFG (Zadeh et al., 2016) introduces graph neural networks to multimodal sentiment analysis task. It explores the interactions between unimodal, bimodal, and trimodal through dynamic fusion.
- TFN: The TFN (Zadeh et al., 2017) explores the interaction between unimodal, bimodal, and trimodal by calculating the Cartesian product.
- LMF: The LMF (Liu et al., 2018) is a low-rank multimodal fusion method to capture inter-modal information with high efficiency.
- MulT: The MulT (Tsai et al., 2019) performs cross-modal attention module based on Transformer to mine the interaction between sequences.
- BIMHA: The BIMHA (Wu et al., 2022) applies multi-head attention mechanism to distill relative relationships and importance between two pair-wise modalities.
- RAVEN: The RAVEN (Wang et al., 2019) designs a multimodal transformation module to integrate non-verbal features into word-level features, learning higher-order semantic information.

Translation based methods. This type methods can be formalize as: Input: $I_t, I_a, I_v \rightarrow$Seq2Seq$(I_{\text{source}} \rightarrow I'_{\text{target}}) \rightarrow$Prediction$(R_{tav}) \rightarrow$Output: $\hat{y}_{tav}$, where $I_{\text{source}} \in \{I_t, I_a, I_v\}$ and $I'_{\text{target}} \in \{I'_t, I'_a, I'_v\}$.

- MCTN: The MCTN (Pham et al., 2019) translates the source modality into the target modality to learn the joint feature representation, using pair-wise modalities during the training process to prevent the loss of key information.

MTL-based methods. This type methods can be formalize as: Input: $I_t, I_a, I_v \rightarrow \{\text{task}_1, \text{task}_2, \ldots, \text{task}_n\} \rightarrow$ Prediction$\{R_1, R_2, \ldots, R_n\} \rightarrow$Fusion$\rightarrow$Output: $\hat{y}_{tav}$, where $\{1, 2, \ldots, n\} \in \{a, t, v, at, av, vt, atv\}$. Although there are prediction results from multiple tasks, we only take the output of the multimodal prediction task as the final result.

- Self-MM: The Self-MM (Yu et al., 2021) is a multi-task model based on self-supervised learning. The unimodal sentiment labels are automatically generated by the model.
- MTFN: The MTFN (Yu et al., 2020) manually annotates unimodal labels on the CH-SIMS dataset to explore the impact of fine-grained labels. It uses text, audio, and vision subtasks to improve the accuracy of the model.
- FmlMSN: The FmlMSN (Peng et al., 2023) is a fine-grained modal label-based multi-stage learning. It aids the final sentiment prediction with six unimodal tasks and three bimodal tasks to improve the robustness of the model.

Table 2
Hyperparameters of CMHFM for different datasets.

Dataset	Parameters										
	lr	bs	tdrp	adrp	vdrp	thid	ahid	vhid	tout	heads	wgd
MOSI-aligned	0.002	128	0	0	0	128	16	128	64	6	0
MOSI-unaligned	0.002	16	0	0	0	256	32	256	128	6	0
MOSEI-aligned	0.002	128	0	0	0	128	16	128	64	6	0
MOSEI-unaligned	0.002	128	0	0	0	256	16	128	64	8	0
CH-SIMS	0.002	64	0	0	0	128	16	128	128	6	0

Table 3
Experimental results of the different methods on aligned versions of CMU-MOSI and CMU-MOSEI, which are labeled in the last column of the table. There are 2199 video clips in CMU-MOSI and 22 856 video clips in CMU-MOSEI.

Model	Acc-2 (%)↑	Acc-5 (%)↑	Acc-7 (%)↑	F1 (%)↑	MAE↓	Corr↑	Dataset
EF-LSTM[a]	77.38/78.48	40.15	35.90	77.35/78.51	0.949	0.669	MOSI
MFN[a]	77.67/78.87	40.47	35.83	77.63/78.90	0.927	0.670	MOSI
DFG[a]	77.14/78.35	38.63	34.64	77.08/78.35	0.956	0.649	MOSI
MCTN[a]	79.30	–	35.60	79.10	0.909	0.677	MOSI
RAVEN[a]	78.00	–	33.20	76.60	0.915	**0.691**	MOSI
MulT[a]	78.70	–	33.60	78.40	0.964	0.662	MOSI
CMHFM	**80.17/81.71**	**42.42**	**37.17**	**80.12/81.72**	**0.907**	0.683	MOSI
EF-LSTM[a]	77.84/80.79	51.16	50.01	78.34/80.67	0.601	0.683	MOSEI
MFN[a]	78.94/82.86	52.76	51.34	79.55/82.85	0.573	0.718	MOSEI
DFG[a]	81.28/83.48	52.69	51.37	81.48/83.23	0.575	0.713	MOSEI
MCTN[a]	79.80	–	49.60	80.60	0.609	0.670	MOSEI
RAVEN[a]	79.10	–	50.00	79.50	0.614	0.662	MOSEI
MulT[a]	80.20	–	46.60	79.80	0.657	0.661	MOSEI
CMHFM	**84.07/84.45**	**54.39**	**52.83**	**83.85/83.97**	**0.548**	**0.737**	MOSEI

[a] The results of models are obtained from Wu et al. (2022).

4.1.3. Experimental parameters

We use Adam (Kingma & Ba, 2015) as the optimizer. All experiments are implemented on the PyTorch. We conduct experiments on aligned and unaligned versions of the CMU-MOSI and CMU-MOSEI, respectively. Meanwhile, we report the results on the CH-SIMS. The relevant hyper-parameters used by the CMHFM in different datasets include learning rate (lr), batch size (bs), the dropout (tdrp, adrp, vdrp) and hidden units (thid, ahid, thid) of text, audio, and vision feature networks, number of multi-head attention heads (heads), output dimensions of text subnetwork (tout), and weight decay (wgd). The hyper-parameters are shown in Table 2.

4.1.4. Evaluation metrics

To be consistent with prior works (Wasifur et al., 2019; Yu et al., 2020), for CMU-MOSI and CMU-MOSEI datasets, sentiment can be categorized into seven types according to intensity, corresponding to [−3, −2): highly negative, [−2, −1): negative, [−1, 0): weakly negative, [0]: neutral, (0, 1]: weakly positive, (1, 2]: positive, and (2, 3]: highly positive, respectively. Here, we take 7-class accuracy (Acc-7), 5-class accuracy (Acc-5), 2-class accuracy (Acc-2), F1 score (F1), Mean Absolute Error (MAE), and Pearson Correlation Coefficient (Corr) as the evaluation metrics. More specifically, 5-class contains negative, weakly negative, neutral, weakly positive, and positive. Acc-2 and F1 are classified into two cases: one is negative/non-negative (non-include neutral), and the other is negative/positive (include neutral). Corr is the quotient of the covariance and standard deviation between two variables. MAE represents the mean absolute error between actual values and predicted values. The MAE is calculated as follows:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{y}^i - y^i\right| \tag{16}$$

where $\hat{y}$ and y are predicted value and actual label value, respectively. We denote the total number of sample with N.

For CH-SIMS dataset, the sentiment labels are categorized into five classifications, where {−1.0,−0.8} represents negative, {−0.6, −0.4, −0.2} represents weakly negative, {0} represents neutral, {0.2, 0.4, 0.6} represents weakly positive, and {0.8, 1.0} represents positive. Thus, we report 5-class accuracy (Acc-5), 3-class accuracy (Acc-3), 2-class accuracy (Acc-2), F1 score (F1), MAE, and Corr as the evaluation metrics. Specifically, 3-class contains weakly negative, neutral, and weakly positive.

4.2. Results and analysis

In this part, we conduct comprehensive experiments on three multimodal sentiment datasets, including CMU-MOSI and CMU-MOSEI in English, the CH-SIMS in Chinese. The corresponding results are shown in Tables 3, 4, 5, 6, 7, and 8, respectively. In these tables, the results with black bolded font indicate the best results, the results with blue font represent the second-best results, ↑ means the higher the better, ↓ indicates that smaller results are better. In Tables 3–6, the dataset is labeled in the last column of the tables, abbreviated as MOSI and MOSEI. Specifically, for Acc-2 and F1, the left of the slash denotes the negative/non-negative (Non0-Acc, Non0-F1), and the right of the slash denotes the negative/positive (Has0-Acc, Has0-F1).

Table 4
The average results and standard deviations of the different methods on aligned versions of CMU-MOSI and CMU-MOSEI. There are 2199 video clips in CMU-MOSI and 22 856 video clips in CMU-MOSEI.

Model	Acc-2 (%)↑	Acc-5 (%)↑	Acc-7 (%)↑	F1 (%)↑	MAE↓	Corr↑	Dataset
EF-LSTM[a]	78.14 ± 0.65/79.91 ± 0.49	37.29 ± 2.74	34.31 ± 1.85	78.00 ± 0.68/79.86 ± 0.50	0.970 ± 0.034	0.647 ± 0.009	MOSI
MFN[a]	77.38 ± 1.16/78.57 ± 1.48	40.82 ± 2.04	35.74 ± 1.42	77.29 ± 1.02/78.55 ± 1.34	0.937 ± 0.031	0.663 ± 0.014	MOSI
DFG[a]	77.84 ± 0.72/78.99 ± 0.93	39.62 ± 3.08	35.16 ± 2.39	77.90 ± 0.98/79.30 ± 1.16	0.947 ± 0.033	0.665 ± 0.010	MOSI
MCTN[a]	79.21 ± 1.19	–	32.68 ± 0.70	79.30 ± 1.16	0.978 ± 0.030	0.672 ± 0.012	MOSI
RAVEN[a]	79.82 ± 0.52	–	36.21 ± 1.32	79.77 ± 0.56	0.918 ± 0.008	0.667 ± 0.010	MOSI
MulT[a]	80.49 ± 1.02	–	34.99 ± 0.76	80.52 ± 0.97	0.923 ± 0.014	**0.691 ± 0.008**	MOSI
CMHFM	**79.88 ± 0.26/80.95 ± 0.71**	**42.10 ± 1.29**	**37.03 ± 0.45**	**79.82 ± 0.23/81.29 ± 0.45**	**0.912 ± 0.009**	0.677 ± 0.007	MOSI
EF-LSTM[a]	78.20 ± 3.97/80.69 ± 1.60	50.75 ± 0.79	49.69 ± 0.60	78.55 ± 3.31/80.69 ± 1.60	0.603 ± 0.012	0.678 ± 0.013	MOSEI
MFN[a]	81.35 ± 1.13/83.94 ± 0.38	52.97 ± 0.46	51.46 ± 0.49	81.69 ± 0.95/83.81 ± 0.35	0.571 ± 0.003	0.722 ± 0.210	MOSEI
DFG[a]	81.81 ± 1.97/**84.11 ± 0.29**	53.30 ± 0.52	51.92 ± 0.53	82.10 ± 1.71/**83.96 ± 0.52**	0.565 ± 0.003	0.727 ± 0.004	MOSEI
MCTN[a]	79.19 ± 1.23/82.58 ± 0.71	49.29 ± 0.56	48.40 ± 0.63	80.24 ± 1.10/82.50 ± 0.67	0.614 ± 0.009	0.686 ± 0.012	MOSEI
RAVEN[a]	82.73 ± 0.23	–	49.93 ± 1.16	82.09 ± 0.35	0.606 ± 0.010	0.697 ± 0.007	MOSEI
MulT[a]	79.75 ± 1.42	–	52.03 ± 0.03	80.32 ± 1.21	0.563 ± 0.021	**0.731 ± 0.003**	MOSEI
CMHFM	**83.94 ± 0.18**/83.49 ± 0.55	**54.27 ± 0.11**	**52.61 ± 0.29**	**83.57 ± 0.22**/82.90 ± 0.62	**0.558 ± 0.007**	0.731 ± 0.004	MOSEI

[a] The results of models are reproduced using public code under the same platform.

Table 5
Experimental results of the different methods on unaligned versions of CMU-MOSI and CMU-MOSEI. There are 2199 video clips in CMU-MOSI and 22 856 video clips in CMU-MOSEI.

Model	Acc-2 (%)↑	Acc-5 (%)↑	Acc-7 (%)↑	F1 (%)↑	MAE↓	Corr↑	Dataset
TFN[a]	77.99/79.08	39.39	34.46	77.95/79.11	0.947	0.673	MOSI
LMF[a]	77.90/79.08	38.13	33.82	77.80/79.15	0.947	0.651	MOSI
MulT[a]	79.71/80.98	42.68	36.91	79.63/80.95	0.880	0.702	MOSI
MTFN[a]	79.01/80.26	39.36	35.32	79.63/80.95	0.936	0.658	MOSI
Self-MM[a]	**84.00/85.98**	–	–	**84.42/85.95**	**0.713**	**0.798**	MOSI
BIMHA[a]	78.57/80.30	43.29	36.44	78.50/80.03	0.925	0.671	MOSI
FmlMSN[a]	80.09/–	40.61	31.44	79.78/–	0.977	0.669	MOSI
CMHFM	79.45/81.10	**47.08**	**42.13**	79.30/81.02	0.822	0.723	MOSI
TFN[a]	78.50/81.89	53.10	51.60	78.96/81.74	0.573	0.714	MOSEI
LMF[a]	80.54/83.48	52.99	51.59	80.94/83.36	0.576	0.717	MOSEI
MulT[a]	–/81.60	–	50.70	–/81.60	0.591	0.694	MOSEI
MTFN[a]	82.37/82.59	50.95	50.29	82.14/82.03	0.597	0.699	MOSEI
Self-MM[a]	82.81/**85.17**	–	–	82.53/**85.30**	**0.530**	**0.765**	MOSEI
BIMHA[a]	84.07/83.96	53.36	52.11	83.35/83.50	0.559	0.731	MOSEI
FmlMSN[a]	83.45/84.29	54.17	**52.69**	83.56/84.10	0.569	0.719	MOSEI
CMHFM	**84.31**/84.37	**54.37**	52.61	**84.18**/84.01	0.548	0.747	MOSEI

[a] The results of models are obtained from Peng et al. (2023), Wu et al. (2022) and Yu et al. (2021).

4.2.1. Experiments on English datasets

The results of the proposed model compared with some classical methods on aligned versions of CMU-MOSI and CMU-MOSEI are listed in Table 3. With the CMU-MOSI dataset, we can find that the proposed model achieves competitive performance on all metrics except Corr. Compared with the best baseline MCTN, CMHFM improves by 0.002 on MAE. With the CMU-MOSEI dataset, it has been observed that our model substantially outperforms other comparison methods in all indicators. Compared with DFG, CMHFM exhibits highly competitive performance, with a 0.027 improvement on MAE. In addition, the performance of MCTN is second to that of CMHFM on the CMU-MOSI dataset. However, with the CMU-MOSEI dataset, DFG achieves relatively good results. It is worth noting that CMHFM has superior results on both aligned datasets. The results demonstrate the generalizability of our model. We further run five times with different random seeds and report average results and standard deviations of the compared models on aligned versions of CMU-MOSI and CMU-MOSEI. The results are presented in Table 4. With the CMU-MOSI dataset, CMHFM yields better performance in most metrics, achieving Acc-2 as high as 80.95%, and MAE as low as 0.912. With the CMU-MOSEI dataset, the average results and standard deviations of our model obtain the best performance across most metrics. The Acc-2 (negative/positive) of CMHFM is marginally lower than that of DFG, but MAE of CMHFM outperforms those of the other models. In addition, from Table 4, standard deviations of the proposed model on both aligned datasets suggest that the stability of CMHFM.

The results of the comparison models on unaligned versions of CMU-MOSI and CMU-MOSEI are shown in Table 5. With the CMU-MOSI dataset, our model has exhibited great advantages over TFN, LMF, MulT, and MTFN. The reason is that TFN and LMF do not extract deep-level features in cross-modal interactions, resulting in poor results. MulT introduces the Transformer module to learn the interaction information between modalities, but the results are unsatisfactory. MTFN uses only multimodal labels for training on the English datasets because unimodal labels are unavailable. This is the reason why MTFN does not perform well on the CMU-MOSI dataset. Our model obtains the best results on Acc-5 and Acc-7 and is slightly inferior to that of Self-MM on Acc-2, F1, MAE, and Corr. Self-MM learns fine-grained information via sentiment labels automatically generated by the model. The

Table 6
The average results and standard deviations of the different methods on unaligned versions of CMU-MOSI and CMU-MOSEI. There are 2199 video clips in CMU-MOSI and 22 856 video clips in CMU-MOSEI.

Model	Acc-2 (%)↑	Acc-5 (%)↑	Acc-7 (%)↑	F1 (%)↑	MAE↓	Corr↑	Dataset
TFN[a]	78.18 ± 0.80/79.45 ± 0.97	39.85 ± 3.80	35.02 ± 2.36	77.00 ± 0.77/79.37 ± 0.91	0.935 ± 0.032	0.658 ± 0.012	MOSI
LMF[a]	76.76 ± 1.34/77.80 ± 1.79	38.51 ± 2.19	34.17 ± 1.55	76.71 ± 1.32/77.83 ± 1.76	0.963 ± 0.041	0.663 ± 0.070	MOSI
MulT[a]	78.78 ± 0.48/80.21 ± 0.45	39.21 ± 1.57	34.93 ± 0.80	78.64 ± 0.54/80.14 ± 0.51	0.916 ± 0.007	0.685 ± 0.003	MOSI
MTFN[a]	78.51 ± 0.73/79.60 ± 0.59	39.87 ± 1.41	34.90 ± 0.83	78.49 ± 0.73/79.64 ± 0.58	0.955 ± 0.009	0.659 ± 0.010	MOSI
Self-MM[a]	**82.59 ± 0.54/84.42 ± 0.58**	–	–	**82.52 ± 0.53/84.41 ± 0.57**	**0.719 ± 0.014**	0.719 ± 0.014	MOSI
BIMHA[a]	77.58 ± 0.30/78.72 ± 0.48	41.78 ± 1.42	35.35 ± 0.76	77.53 ± 0.25/78.74 ± 0.40	0.967 ± 0.020	0.646 ± 0.008	MOSI
FmlMSN[a]	78.11 ± 0.48/–	39.80 ± 2.05	34.37 ± 2.05	77.95 ± 0.61/–	0.939 ± 0.020	0.658 ± 0.006	MOSI
CMHFM	80.18 ± 0.83/81.55 ± 0.87	**45.97 ± 1.48**	**40.82 ± 1.37**	80.08 ± 0.78/81.52 ± 0.79	0.840 ± 0.013	**0.722 ± 0.004**	MOSI
TFN[a]	80.95 ± 1.90/83.25 ± 0.25	53.22 ± 0.37	51.87 ± 0.43	81.18 ± 1.58/83.00 ± 0.25	0.570 ± 0.008	0.721 ± 0.006	MOSEI
LMF[a]	78.65 ± 1.49/82.85 ± 0.40	53.67 ± 0.49	52.20 ± 0.48	79.32 ± 1.24/82.72 ± 0.47	0.569 ± 0.005	0.728 ± 0.006	MOSEI
MulT[a]	–/**84.76 ± 0.45**	–	52.03 ± 0.03	–/83.74 ± 0.40	0.565 ± 0.002	0.731 ± 0.003	MOSEI
MTFN[a]	82.55 ± 0.52/82.48 ± 0.80	50.85 ± 2.85	49.60 ± 1.77	82.26 ± 0.59/81.90 ± 0.92	0.603 ± 0.002	0.699 ± 0.002	MOSEI
Self-MM[a]	80.63 ± 3.66/83.87 ± 1.43	–	–	81.12 ± 3.23/**83.85 ± 1.28**	**0.530 ± 0.005**	**0.754 ± 0.004**	MOSEI
BIMHA[a]	82.43 ± 0.86/82.47 ± 0.79	51.17 ± 0.82	50.30 ± 0.66	82.20 ± 0.82/81.93 ± 0.72	0.593 ± 0.007	0.700 ± 0.009	MOSEI
FmlMSN[a]	82.87 ± 0.99/82.69 ± 0.66	52.14 ± 0.42	50.91 ± 0.47	82.50 ± 0.95/82.04 ± 0.75	0.578 ± 0.009	0.720 ± 0.006	MOSEI
CMHFM	**84.09 ± 0.12**/83.90 ± 0.43	**53.98 ± 0.33**	**52.39 ± 0.28**	**83.83 ± 0.22**/83.40 ± 0.55	0.560 ± 0.009	0.733 ± 0.008	MOSEI

[a] The results of models are reproduced using public code under the same platform.

Table 7
Experimental results of the different methods on the CH-SIMS. There are 2281 video clips in dataset.

Model	Acc-2 (%)↑	Acc-3 (%)↑	Acc-5 (%)↑	F1 (%)↑	MAE↓	Corr↑
TFN[a]	78.38	65.12	39.30	78.62	0.432	0.591
LMF[a]	79.43	66.52	42.89	79.75	0.411	0.630
MulT[a]	77.46	64.33	34.57	78.27	0.448	0.563
MTFN[a]	81.09	69.37	40.31	81.01	0.395	0.666
Self-MM[a]	80.04	65.47	41.53	80.44	0.424	0.569
BIMHA[a]	82.71	69.23	45.21	82.72	0.385	0.660
FmlMSN[a]	83.59	**70.90**	41.79	83.71	0.385	0.680
CMHFM	**84.46**	68.27	**45.29**	**84.80**	**0.380**	**0.685**

[a] The results of models are obtained from Peng et al. (2023), Wu et al. (2022) and Yu et al. (2020, 2021).

Table 8
The average results and standard deviations of the different methods on the CH-SIMS. There are 2281 video clips in dataset.

Model	Acc-2 (%)↑	Acc-3 (%)↑	Acc-5 (%)↑	F1 (%)↑	MAE↓	Corr↑
TFN[a]	80.66 ± 1.40	64.46 ± 1.70	38.38 ± 3.60	81.62 ± 1.10	0.425 ± 0.011	0.612 ± 0.012
LMF[a]	79.34 ± 0.40	64.38 ± 2.10	35.14 ± 4.60	79.96 ± 0.60	0.440 ± 0.016	0.600 ± 0.013
MulT[a]	77.94 ± 0.90	65.03 ± 0.90	35.34 ± 2.90	79.10 ± 0.90	0.485 ± 0.026	0.560 ± 0.006
MTFN[a]	82.45 ± 1.30	**69.02 ± 0.03**	37.20 ± 1.80	82.56 ± 1.20	0.407 ± 0.011	**0.670 ± 0.013**
Self-MM[b]	78.08 ± 0.45	65.91 ± 0.81	42.71 ± 0.64	78.11 ± 0.40	0.414 ± 0.008	0.595 ± 0.008
BIMHA[b]	80.79 ± 0.59	65.95 ± 1.56	42.71 ± 1.23	81.19 ± 0.65	0.403 ± 0.005	0.629 ± 0.001
FmlMSN[b]	81.27 ± 0.33	67.40 ± 1.35	40.48 ± 2.53	81.88 ± 0.30	0.411 ± 0.008	0.647 ± 0.001
CMHFM	**82.97 ± 0.80**	67.61 ± 0.84	**43.05 ± 1.25**	**83.21 ± 0.89**	**0.398 ± 0.010**	0.663 ± 0.014

[a] The results of models are obtained from Yu et al. (2020).
[b] The results of models are reproduced using public code under the same platform.

proposed model introduces multi-task learning to fuse cross-modal features from different layers. Therefore, CMHFM achieves top two results in all metrics. With the CMU-MOSEI dataset, the Acc-2, Acc-5, and F1 of our model have improvement compared with other methods do. The underlying reason could be that the cross-modal hierarchical fusion module can learn relevant information from different levels. Furthermore, CMHFM introduces multi-task learning to filter out the noisy information from cross-modal fusion. Compared with the classical model Self-MM, our method performs slightly worse on Has0-Acc, Has0-F1, MAE, and Corr. Compared with BIMHA, our model gets overwhelming results in all metrics, achieving as low as 0.548 on MAE. The Acc-2, Acc-5, Acc-7, and F1 of our model and FmlMSN are closer, but the MAE of our model is lower than that of FmlMSN. Therefore, our model outperforms FmlMSN. FmlMSN uses multimodal labels instead of fine-grained labels on the CMU-MOSEI dataset, resulting in suboptimal performance. Overall, CMHFM outperforms other comparison models in most metrics. Our model poses cross-modal hierarchical fusion strategy, which obtains information from different levels more effectively than other methods do. Meanwhile, the proposed model achieves the best performance on the aligned dataset. Overall, the excellent performances of CMHFM on aligned and unaligned datasets suggest the efficiency of the proposed model.

To demonstrate the stability and effectiveness of CMHFM, we perform experiments on unaligned versions of CMU-MOSI and CMU-MOSEI. Specifically, we run 5 times of the compared models under different random seeds and calculate their average

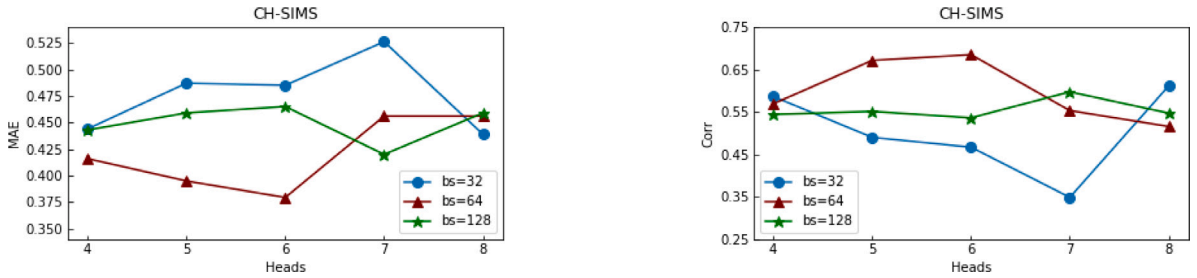


Fig. 3. Results of effect of heads and batch size (bs) on the CH-SIMS.

results and standard deviations. The detailed results are presented in Table 6. With the CMU-MOSI dataset, Self-MM still achieves competitive performance across most metrics. Our model obtains the best results on Acc-5, Acc-7, and Corr. Moreover, the Acc-2, F1, MAE, and Corr of Our model are slightly inferior to those of Self-MM. With the CMU-MOSEI dataset, compared with Self-MM, CMHFM only performs slightly not as good as it does in MAE and Corr. However, CMHFM outperforms Self-MM on Acc-2, Acc-5, Acc-7, and F1. A potential reason is that Self-MM automatically generates unimodal labels and utilizes them for the final sentiment prediction. Additionally, standard deviations of CMHFM obtain relatively small values, especially on the CMU-MOSEI dataset. Overall, from Table 6, the results suggest that the stability and effectiveness of the proposed model on both datasets.

From the above results, our method gets promising results on aligned versions of CMU-MOSI and CMU-MOSEI. As listed in Table 1, we can observe that the number of video clips in the CMU-MOSEI is much bigger than that of the CMU-MOSI. The results of our model on the unaligned version of CMU-MOSI dataset are slightly worse than those of Self-MM. However, CMHFM achieves competitive performance in all metrics on the unaligned version of the CMU-MOSEI dataset as the amount of data increases. We can infer that the proposed model is dependent on the size of the dataset and performs better on larger dataset. In particular, the average results and standard deviations demonstrate the stability of the proposed method on both CMU-MOSI and CMU-MOSEI datasets. In general, benefiting from the cross-modal hierarchical learning module, the proposed model demonstrates competitiveness and robustness.

4.2.2. Experiments on Chinese dataset

To further evaluate the generality and robustness of the proposed model, we perform experiments on the CH-SIMS, a commonly used and unaligned Chinese dataset for multimodal sentiment analysis. The comparative results are presented in Tables 7 and 8. From Table 7, we can observe that the proposed method achieves the best performance across all metrics except for the Acc-3 which is only second to that of FmlMSN. These encouraging results can be attributed to multi-tasking assisted learning. EF-LSTM fails to capture interaction information among modalities, which leads to poor performance. MTFN and Self-MM introduce unimodal tasks to improve the accuracy of sentiment classification. The results indicate that MTFN outperforms the Self-MM. The underlying reason could be that Self-MM applies the automatically generated unimodal labels while MTFN leverages the manually annotated unimodal labels provided by the dataset. The proposed method achieves superior results compared to the second performer, FmlMSN, with an improvement of 0.005 on MAE and 3.50% on Acc-5. However, the second performer FmlMSN does not obtain good results on the CMU-MOSI and CMU-MOSEI datasets. Therefore, the generalization ability of the method is poor. Through the analysis of experimental results, we find our model is superior to the compared models. This is because cross-modal hierarchical fusion can capture the differences among the levels. We further give the average results and standard deviations of the compared models, which are listed in Table 8. The proposed method still gains superior results. Our method outperforms second performer MTFN on metric MAE by 0.009 and F1 by 0.65%. More importantly, the standard deviations of all indicators are less than 1.40. The results suggest that CMHFM is more stable compared to other models.

From the results of the above experiments, our model achieves promising performance on both Chinese and English datasets, which suggests the effectiveness and generalization of the proposed model. In addition, the repeated experimental results consistently demonstrate the stability of CMHFM. We can attribute these encouraging results to cross-modal hierarchical fusion, which effectively learns comprehensive information from different levels.

4.2.3. Sensitivity analysis

In this part, we conduct sensitivity analysis on the heads and batch size (bs) used in the proposed method. The experimental results on the CH-SIMS are illustrated in Fig. 3. We set the bs to 32, 64, and 128. The number of heads range from 4 to 8. MAE and Corr are selected as evaluation metrics. From Fig. 3, it is clear that the proposed method obtains best results in terms of both MAE and Corr with fix bs = 64 and heads = 6. Compared with heads = 8, it can find that our method achieves better performance when selecting heads = 7 and bs = 128. The underlying reason could be summarized as two factors. First, since there are 2281 video clips in the CH-SIMS, it is easy to cause model overfitting when setting bs = 128. Second, too many heads may introduce redundant information, which further causes suboptimal performance.

Table 9
The results on different combinations of modules.

Module	Acc-2 (%)↑	Acc-5 (%)↑	Acc-7 (%)↑	F1 (%)↑	MAE↓	Corr↑
With U	**84.01**/82.61	52.39	51.34	**83.50**/81.91	0.571	0.722
With U+I	83.28/82.31	52.67	51.21	82.78/81.57	0.575	0.715
W/o M	83.56/**83.93**	**53.10**	**52.07**	83.37/**83.47**	**0.569**	**0.725**
W/o H	83.30/83.32	51.73	50.20	83.11/82.86	0.577	0.718
H W/o I	83.77/83.39	52.26	50.70	83.39/82.55	0.584	0.713
H W/o U	83.54/82.72	53.21	51.56	83.03/81.98	0.576	0.721
CMHFM	**84.31**/**84.37**	**54.37**	**52.61**	**84.18**/**84.01**	**0.548**	**0.746**

Table 10
The results on different combinations of modalities.

Modal	Acc-2 (%)↑	Acc-5 (%)↑	Acc-7 (%)↑	F1 (%)↑	MAE↓	Corr↑
I_a	70.49/54.56	40.59	40.59	64.34/56.80	0.813	0.241
I_v	70.92/63.59	42.43	42.41	62.06/52.88	0.796	0.284
I_t	70.92/63.59	51.21	50.23	81.91/81.86	0.584	0.704
$I_a + I_v$	70.29/65.52	40.67	40.65	66.08/59.93	0.821	0.292
$I_a + I_t$	82.83/82.11	51.86	50.74	82.31/81.32	0.577	0.718
$I_v + I_t$	83.60/82.50	50.85	50.18	82.93/81.59	0.590	0.701
$I_a + I_v + I_t$	**84.31**/**84.37**	**54.37**	**52.61**	**84.18**/**84.01**	**0.548**	**0.746**

4.2.4. Ablation study on CMU-MOSEI

To investigate the necessity and contribution of Unimodal Feature Learning (U), Inter-Modal Interaction (I), Multi-Head Attention (M), and Multi-Tasking Assisted Learning (H), we perform experiments of each module on unaligned version of the CMU-MOSEI dataset. The results are shown in Table 9. We set up seven groups of experiments, divided into two parts. The descriptions of each module are shown below:

- **With U** indicates that only unimodal feature learning is used to predict sentiment.
- **With U+I** denotes the combination of unimodal feature learning and inter-modal interaction.
- **W/o M** indicates that the multi-head attention is discarded.
- **W/o H** represents cross-modal hierarchical fusion module not applied in the CMHFM.
- **H w/o I** indicates that bimodal prediction results are not used in multi-tasking assisted learning.
- **H w/o U** discards unimodal prediction results in multi-tasking assisted learning.

As we can infer from Table 9, 'W/o M' shows significant improvement over 'With U' and 'With U+I'. The underlying reason could be that 'W/o M' applies cross-modal hierarchical fusion module and considers the cues contained in the different layers. It can further supplement the rich unimodal information. 'H w/o I' utilizes unimodal and global prediction tasks in the HTAL module to assist multimodal sentiment analysis. 'H w/o U' performs bimodal and global fusion in the HTAL module to improve the accuracy of recognition. From the results of 'W/o H', 'H w/o I', and 'H w/o U', we can notice that HTAL retains either unimodal or bimodal to achieve cross-modal fusion and improves performance of the models to a great extent. From Table 9, we can infer that the best performance is obtained by the complete HTAL component. The experimental results demonstrate that the cross-modal hierarchical fusion module is helpful for the final sentiment prediction. HTAL module based on multi-task learning complements information from unimodal, bimodal, and global aspects and diminishes the negative effects of noise.

To further test the effect and importance of different combinations of modalities for the final sentiment prediction, we use audio (I_a), vision (I_v), text (I_t), audio-vision ($I_a + I_v$), audio-text ($I_a + I_t$), vision-text ($I_v + I_t$), and audio-vision-text ($I_a + I_v + I_t$) as the inputs of the model. The results of the ablation experiments are shown in Table 10. In the unimodal sentiment prediction tasks, we find that the text modality performs best, followed by the vision modality, and then the audio modality. The obtained results are reasonable as text contains higher-level semantic information, and BERT and BiGRU are effective in capturing the contextual relevance features. The presence of background noise in the audio leads to the extracted features to be less refined, which is one of the reasons for the suboptimal performance of the audio task. In bimodal sentiment prediction tasks, it is obvious that $I_a + I_t$ and $I_v + I_t$ perform better than $I_a + I_v$ does. Text plays a crucial role in multimodal sentiment analysis, so the pair-wise modalities with text have improvements in all indicators to some extent. Overall, the performance of the model with $I_a + I_v + I_t$ as the input substantially outperforms than that with unimodal and bimodal. In addition, text modality provides more useful information to the sentiment analysis model than audio and vision do.

5. Conclusion

Due to the correlation and differences between modalities, conflicting information is brought during cross-modal interactions, resulting in poor quality of the joint feature representation. To solve this problem, we introduce multi-task learning and propose a cross-modal hierarchical fusion method for multimodal sentiment analysis. The model uses a tensor fusion network to mine the

interaction between modalities and then utilizes multi-head attention mechanism to explore the importance of bimodal information. Afterward, multi-task learning is adopted to assist the final sentiment analysis and reduce misinformation due to cross-modal fusion. To verify the feasibility and validity of the proposed model in this paper, we conduct extensive experiments on the CH-SIMS, CMU-MOSI, and CMU-MOSEI datasets, and the experimental results show the effectiveness and generalization ability of CMHFM. In the future, we will collect some data with partial availability of three modalities and explore the way of cross-modal interactions for sentiment analysis.

CRediT authorship contribution statement

Lan Wang: Writing – review & editing, Writing – original draft, Software, Methodology, Conceptualization. **Junjie Peng:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **Cangzhi Zheng:** Writing – review & editing, Conceptualization. **Tong Zhao:** Formal analysis. **Li'an Zhu:** Visualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors would like to thank the funding from the Shanghai Service Industry Development Fund, China and the resources and technical support from the High Performance Computing Center of Shanghai University, China, and Shanghai Engineering Research Center of Intelligent Computing System, China (No. 19DZ2252600).

Ethics approval

This article has never been submitted to more than one journal for simultaneous consideration. This article is original.

Consent to participate

The authors have approved this article before submission, including the names and order of authors.

References

Akhtar, M. S., Chauhan, D. S., Ghosal, D., Poria, S., Ekbal, A., & Bhattacharyya, P. (2019). Multi-task learning for multi-modal emotion recognition and sentiment analysis. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: human language technologies* (pp. 370–379).

Baltrušaitis, T., Robinson, P., & Morency, L. P. (2016). OpenFace: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision* (pp. 1–10).

Brady, K., Gwon, Y., Khorrami, P., Godoy, E., Campbell, W., Dagli, C., & Huang, T. S. (2016). Multi-modal audio, video and physiological sensor learning for continuous emotion prediction. In *Proceedings of the 6th international workshop on audio/visual emotion challenge* (pp. 97–104).

Chen, G., Peng, J., Zhang, W., Huang, K., Cheng, F., Yuan, H., & Huang, Y. (2023). A region group adaptive attention model for subtle expression recognition. *IEEE Transactions on Affective Computing*, *14*(2), 1613–1626.

Chen, D., Su, W., Wu, P., & Hua, B. (2023). Joint multimodal sentiment analysis based on information relevance. *Information Processing & Management*, *60*(2), Article 103193.

Cho, K., Van, M. B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1723–1734).

Eyben, F., Weninger, F., Gross, F., & Schuller, B. (2013). Recent developments in openSMILE, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on multimedia* (pp. 835–838).

Fang, L., Liu, G., & Zhang, R. (2022). Sense-aware BERT and multi-task fine-tuning for multimodal sentiment analysis. In *2022 international joint conference on neural networks* (pp. 1–8).

Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., & Hussain, A. (2023). Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, *91*, 424–444.

He, J., Mai, S., & Hu, H. (2021). A unimodal reinforced transformer with time squeeze fusion for multimodal sentiment analysis. *IEEE Signal Processing Letters*, *28*, 992–996.

Huddar, M. G., Sannakki, S. S., & Rajpurohit, V. S. (2020). Multi-level context extraction and attention-based contextual inter-modal fusion for multimodal sentiment analysis and emotion classification. *International Journal of Multimedia Information Retrieval*, *9*, 103–112.

Jiang, T., Wang, J., Liu, Z., & Ling, Y. (2020). Fusion-extraction network for multimodal sentiment analysis. In *Advances in knowledge discovery and data mining: 24th Pacific-Asia conference* (pp. 785–797).

Kingma, D. P., & Ba, J. L. (2015). Adam: A method for stochastic optimization. In *3rd international conference on learning representations.* (pp. 1–15).

Kumar, A., & Vepa, J. (2020). Gated mechanism for attention based multi modal sentiment analysis. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing* (pp. 4477–4481).

Lai, H., & Yan, X. (2022). Multimodal sentiment analysis with asymmetric window multi-attentions. *Multimedia Tools and Applications*, *81*(14), 19415–19428.

Li, J., Chen, Y., Zhang, X., Nie, J., Li, Z., Yu, Y., Zhang, Y., Hong, R., & Wang, M. (2023). Multimodal feature extraction and fusion for emotional reaction intensity estimation and expression classification in videos with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5837–5843).

Lim, W., Jang, D., & Lee, T. (2016). Speech emotion recognition using convolutional and recurrent neural networks. In *2016 Asia-Pacific signal and information processing association annual summit and conference* (pp. 1–4).

Lin, H., Zhang, P., Ling, J., Yang, Z., Lee, L. K., & Liu, W. (2023). PS-Mixer: A polar-vector and strength-vector mixer model for multimodal sentiment analysis. *Information Processing & Management*, *60*(2), Article 103229.

Liu, Z., Shen, Y., Lakshminarasimhan, V. B., Liang, P. P., Zadeh, A., & Morency, L. (2018). Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th annual meeting of the association for computational linguistics,* (pp. 2247–2256).

Lu, Q., Sun, X., Long, Y., Gao, Z., Feng, J., & Sun, T. (2023). Sentiment analysis: Comprehensive reviews, recent advances, and open challenges. *IEEE Transactions on Neural Networks and Learning Systems*, 1–21.

McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., & Nieto, O. (2015). LibROSA: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference* (pp. 18–24).

Peng, J., Wu, T., Zhang, W., Cheng, F., Tan, S., Yi, F., & Huang, Y. (2023). A fine-grained modal label-based multi-stage network for multimodal sentiment analysis. *Expert Systems with Applications*, *221*, Article 119721.

Pham, H., Liang, P. P., Manzini, T., Morency, L.-P., & Póczos, B. (2019). Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 6892–6899).

Sahay, S., Okur, E., Kumar, S. H., & Nachman, L. (2020). Low rank fusion based transformers for multimodal sequences. (pp. 1–6). arXiv preprint arXiv:2007.02038.

Sener, O., & Koltun, V. (2018). Multi-task learning as multi-objective optimization. In *Advances in neural information processing systems* (pp. 525–536).

Sun, H., Liu, J., Chen, Y., & Lin, L. (2023). Modality-invariant temporal representation learning for multimodal sentiment classification. *Information Fusion*, *91*, 504–524.

Tsai, Y. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. association for computational linguistics.* (pp. 6558–6569).

Tzirakis, P., Chen, J., Zafeiriou, S., & Schuller, B. (2021). End-to-end multimodal affect recognition in real-world environments. *Information Fusion*, *68*, 46–53.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 1–11.

Wang, Y., Shen, Y., Liu, Z., Liang, P. P., Zadeh, A., & Morency, L.-P. (2019). Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 7216–7223).

Wasifur, R., Md., K. H., Sangwu, L., Amir, Z., Mao, C., Louis-Philippe, M., & Ehsan, H. (2019). Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 2359–2369).

Williams, J., Kleinegesse, S., Comanescu, R., & Radu, O. (2018). Recognizing emotions in video using multimodal DNN feature fusion. In *Proceedings of grand challenge and workshop on human multimodal language* (pp. 11–19).

Wu, T., Peng, J., Zhang, W., Zhang, H., Tan, S., Yi, F., Ma, C., & Huang, Y. (2022). Video sentiment analysis with bimodal information-augmented multi-head attention. *Knowledge-Based Systems*, *235*, Article 107676.

Xie, L., & Zhang, X. (2020). Gate-fusion transformer for multimodal sentiment analysis. In *Pattern recognition and artificial intelligence: international conference* (pp. 28–40).

Xu, M., Zhang, F., & Khan, S. U. (2020). Improve accuracy of speech emotion recognition with attention head fusion. In *2020 10th annual computing and communication workshop and conference* (pp. 1058–1064).

Xue, H., Yan, X., Jiang, S., & Lai, H. (2020). Multi-tensor fusion network with hybrid attention for multimodal sentiment analysis. In *2020 international conference on machine learning and cybernetics* (pp. 169–174).

Yang, B., Wu, L., Zhu, J., Shao, B., Lin, X., & Liu, T.-Y. (2022). Multimodal sentiment analysis with two-phase multi-task learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 2015–2024.

You, Q., Luo, J., Jin, H., & Yang, J. (2015). Joint visual-textual sentiment analysis with deep neural networks. In *Proceedings of the 23rd annual ACM conference on multimediaConference.* (pp. 1071–1074).

Yu, W., Xu, H., Meng, F., Zhu, Y., Ma, Y., Wu, J., Zou, J., & Yang, K. (2020). CH-SIMS: a Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 3718–3727).

Yu, W., Xu, H., Yuan, Z., & Wu, J. (2021). Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 10790–10797).

Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. (2017). Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing.* (pp. 1103–1114).

Zadeh, A., Liang, P. P., Mazumder, N., Poria, S., Cambria, E., & Morency, L. (2018). Memory fusion network for multi-view sequential learning. In *Proceedings of the thirty-second AAAI conference on artificial intelligence* (pp. 5634–5641).

Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E., & Morency, L.-P. (2018). Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th annual meeting of the association for computational linguistics* (pp. 2236–2246).

Zadeh, A., Zellers, R., Pincus, E., & Morency, L.-P. (2016). MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. (pp. 1–10). arXiv preprint arXiv:1606.06259.

Zhang, K., Geng, Y., Zhao, J., Liu, J., & Li, W. (2020). Sentiment analysis of social media via multimodal feature fusion. *Symmetry*, *12*(12), 2010.

Zhang, Z., Wu, B., & Schuller, B. (2019). Attention-augmented end-to-end multi-task learning for emotion prediction from speech. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing* (pp. 6705–6709).

Zhao, T., Peng, J., Huang, Y., Wang, L., Zhang, H., & Cai, Z. (2023). A graph convolution-based heterogeneous fusion network for multimodal sentiment analysis. *Applied Intelligence*, *53*(14), 30455–30468.