



Hierarchical denoising representation disentanglement and dual-channel cross-modal-context interaction for multimodal sentiment analysis

Zuhe Li^a, Zhenwei Huang^a, Yushan Pan^{b,*}, Jun Yu^a, Weihua Liu^c, Haoran Chen^a, Yiming Luo^d, Di Wu^e, Hao Wang^d

^a School of Computer Science and Technology, Zhengzhou University of Light Industry, Zhengzhou 450002, China

^b Xi'an Jiaotong-Liverpool University, Suzhou, 215123, China

^c China Mobile Research Institute, Beijing, 100053, China

^d Xidian University, Xi'an, 710071, China

^e Norwegian University of Science and Technology, Aalesund, 6009, Norway

ARTICLE INFO

Keywords:

Multimodal sentiment analysis
Hierarchical disentanglement
Inter-modal enhancement
Cross-modal context interaction

ABSTRACT

Multimodal sentiment analysis aims to extract sentiment cues from various modalities, such as textual, acoustic, and visual data, and manipulate them to determine the inherent sentiment polarity in the data. Despite significant achievements in multimodal sentiment analysis, challenges persist in addressing noise features in modal representations, eliminating substantial gaps in sentiment information among modal representations, and exploring contextual information that expresses different sentiments between modalities. To tackle these challenges, our paper proposes a new Multimodal Sentiment Analysis (MSA) framework. Firstly, we introduce the Hierarchical Denoising Representation Disentanglement module (HDRD), which employs hierarchical disentanglement techniques. This ensures the extraction of both common and private sentiment information while eliminating interference noise from modal representations. Furthermore, to address the uneven distribution of sentiment information among modalities, our Inter-Modal Representation Enhancement module (IMRE) enhances non-textual representations by extracting sentiment information related to non-textual representations from textual representations. Next, we introduce a new interaction mechanism, the Dual-Channel Cross-Modal Context Interaction module (DCCMCI). This module not only mines correlated contextual sentiment information within modalities but also explores positive and negative correlation contextual sentiment information between modalities. We conducted extensive experiments on two benchmark datasets, MOSI and MOSEI, and the results indicate that our proposed method offers state-of-the-art approaches.

1. Introduction

Originally, sentiment analysis involved using Natural Language Processing (NLP) techniques to extract sentiment information, including opinions and feelings, from subjective text (Zhang, Xu, & Zhao, 2020). However, the rapid expansion of social media platforms like Twitter, TikTok, and YouTube has led to explosive growth in video data containing multimodal information—encompassing textual, acoustic, and visual elements (Shi, Fan, Wang, & Zhang, 2022). Traditional text-based sentiment analysis now struggles to handle the complexities of this data, prompting a growing interest in multimodal sentiment analysis, which extracts attitudes, opinions, and sentiment information from various modalities (Su & Kuo, 2022). Simultaneously, the widespread use of mobile devices not only facilitates the capture of

diverse modal sentiment cues from users (Michalis, Vassilis, Nicholas, & Petros, 2019) but also enables the application of multimodal sentiment analysis across various economic and social sectors (Wang et al., 2022). Consequently, an increasing number of researchers are delving into this promising and evolving field.

In recent years, deep learning methods have dominated multimodal sentiment analysis research, aiming to leverage complementary sentiment information between multimodal data to construct complex deep learning models (Abdu, Yousef, & Salem, 2021; Zhao, Jia, Yang, Ding, & Keutzer, 2021). While these methods have led to some improvements in accuracy, challenges persist (Zhu, Zhu, Zhang, Xu, & Kong, 2023). Effective representation disentanglement poses a key challenge, given

* Corresponding author.

E-mail addresses: zuheli@zzuli.edu.cn (Z. Li), 332207050671@email.zzuli.edu.cn (Z. Huang), yushanp@liverpool.ac.uk (Y. Pan), yujun@zzuli.edu.cn (J. Yu), liuweihuayjy@chinamobile.com (W. Liu), chenhaoran@zzuli.edu.cn (H. Chen), luoyiming@xidian.edu.cn (Y. Luo), di.wu@ntnu.no (D. Wu), wanghao@xidian.edu.cn (H. Wang).

<https://doi.org/10.1016/j.eswa.2024.124236>

Received 21 February 2024; Received in revised form 15 April 2024; Accepted 13 May 2024

Available online 16 May 2024

0957-4174/© 2024 Elsevier Ltd. All rights reserved, including those for text and data mining, AI training, and similar technologies.

the significant distribution gap between textual, acoustic, and visual representations. Existing methods address this by introducing domain separation into multimodal sentiment analysis, partitioning representations into modality-invariant and modality-specific subspaces to obtain appropriate multimodal representations (Bousmalis, Trigeorgis, Silberman, Krishnan, & Erhan, 2016; Hazarika, Zimmermann, & Poria, 2020). However, each modality representation not only contains information expressing common and private sentiment but also includes a considerable amount of irrelevant noise, which can significantly reduce the accuracy of sentiment analysis tasks. Therefore, the challenge lies in obtaining multimodal representations that balance common and private sentiments while mitigating the impact of noise.

Another challenge in multimodal sentiment analysis is the significant gap in modality information quality, leading to uneven distribution of sentiment information between modal representations. Enhancing representation information often becomes necessary to address this. Current approaches extract sentiment information related to the text modality from non-text modalities to strengthen text representation (Wang et al., 2019). However, this approach does not fully address sentiment information scarcity in non-textual representations. Leveraging sentiment information from textual representations to enhance non-textual representations remains a puzzle.

Modal interaction poses another challenge, as sentiment information in multimodal data is often complementary. Exploring these complementary pieces of information through modal interaction to generate effective multimodal representations is crucial. However, current methods tend to overlook negative correlation context information between modalities, despite its potential importance in certain contexts such as irony or delivering bad news (Vaswani et al., 2017). Effectively focusing on both positive and negative correlation context sentiment information between modalities remains a challenging task.

In response to the aforementioned challenges, we introduce a novel framework for Multimodal Sentiment Analysis (MSA). Firstly, we employ a hierarchical disentanglement technique to project the representation of each modality into modality-common, -private, and -noisy subspaces. These representations are then constrained to ensure their appropriateness. Secondly, recognizing the significant gap in sentiment information between modal representations, we aim to leverage the rich emotional information in the textual modality to enhance the acoustic and visual modalities. Additionally, we seek to mine contextual information within each modality and between modalities to enrich the sentiment semantics of multimodal representations.

Our study's contributions can be summarized as follows: introducing a novel Multimodal Sentiment Analysis (MSA) framework that tackles key challenges in multimodal sentiment analysis. By employing hierarchical disentanglement, leveraging textual emotional information, and mining contextual cues, our framework aims to enhance the effectiveness and accuracy of multimodal sentiment analysis.

- We introduce a hierarchical denoising representation disentanglement module, which decomposes modal representations through representation constraints. This allows modal representations to incorporate both commonality and individuality information while eliminating noise that may negatively impact sentiment analysis tasks.
- We have designed an inter-modal representation enhancement module to bridge the gap between modalities. This module extracts emotional information related to acoustic and visual content from textual representations, thereby bridging the substantial divide between textual and non-textual modalities.
- We introduce a dual-channel cross-modal context interaction module, which utilizes multiple attention mechanisms to simultaneously emphasize complementary contextual emotional information within and between modalities. This approach enables the extraction of contextual clues with rich semantics.

The subsequent sections of this paper are organized as follows: Section 2 conducts a review of pertinent literature in the realm of multimodal sentiment analysis. Section 3 presents the framework of the model and provides a detailed design of each module. Section 4 introduces the datasets used and outlines the experimental configurations. Section 5 demonstrates the efficacy of the proposed framework through various experiments. Finally, Section 6 provides a summary of the research outcomes, accompanied by directions for future exploration.

2. Related work

With the widespread adoption of social networks and the rapid development of deep learning technology (Biswas & Tešić, 2022), multimodal sentiment analysis has become a key focus of research within the multimodal domain. This approach harnesses diverse data sources, including textual, acoustic, and visual information, to comprehend sentiments (Wu, Lin, Zhao, Qin, & Zhu, 2021). Multimodal sentiment analysis based on deep learning aims to establish a reliable mapping between multimodal data and emotional polarity, a task reliant on the effective fusion of multimodal data. Existing work can be broadly categorized into attention-unrelated and attention-based methods based on their fusion approaches.

2.1. Attention-unrelated methods

These early approaches include the TFN proposed by Zadeh, Chen, Cambria, Poria, and Morency (2017), utilizing the Cartesian product to fuse modal representations. Responding to the complex computation of TFN, Liu et al. (2018) proposed LMF to simplify the computational complexity using low-rank tensors. With the development of feature fusion technology (Yu, Yu, Fan, & Tao, 2017), researchers try to decompose and re-fuse representations, aiming to learn more distinctive representations through factorization (Chen, Shen, Ding, Deng, & Li, 2024). For example, Wang, Yan, Lee, and Livescu (2016) reanalyzed the LVMS using deep variational CCA, obtaining modality variables that include private and shared variables. The MV-LSTM network proposed by Rajagopalan, Morency, Baltruaitis, and Goecke (2016) models consistent and complementary information among multiple modalities using multi-view LSTM blocks. The MFM model designed by Tsai, Liang, Zadeh, Morency, and Salakhutdinov (2018) decomposes the joint representation of multimodal data into intra-modal and inter-modal correlations. Hazarika et al. (2020) proposed the MISA framework, which uses different encoders to learn modal representations from the perspectives of modality-invariant and modality-specific. While these methods have indeed improved sentiment prediction accuracy to some extent, the presence of irrelevant noise significantly impacts model performance. This is because noise lacking emotional information often interferes with sentiment analysis. Furthermore, there has been insufficient attention paid to contextual interaction information within and between modalities.

2.2. Attention-based methods

These methods use various attention mechanisms (Li, Cai, Dong, Lai, & Xie, 2023) to achieve inter-modal and intra-modal information interaction for more effective multi-modal representations (Xiao et al., 2021). In the MARN model, Zadeh, Liang, Poria et al. (2018) employ multiple attention blocks to obtain diverse cross-modal emotional contexts, storing them in a mixed memory block. Ou, Chen, and Wu (2021) proposed a multimodal local-global attention network in the MMLGAN model to fuse representations from different modalities. The Transformer (Vaswani et al., 2017), initially developed for machine translation, has gained attention for its unique advantage in modeling context for sequential data. Researchers have explored its utilization in various domains. Tsai et al. (2019) employed the MulT for the interaction and fusion of multimodal sequences with

different time steps. Chen, Hong, Guo, and Song (2023) proposed the TCDN framework, utilizing a three-modal collaborative network to acquire intra- and inter-modal contextual sentiment information while eliminating irrelevant features between modalities. Wang, Guo et al. (2023) proposed the TETFN, obtaining consistent interaction information between modalities through text-guided cross-modal mapping. Tang, Liu et al. (2023) proposed the BAFN network, using dynamic enhancement blocks and bidirectional attention blocks to explore intra-modality emotional context and more advanced emotional context inter-modally. Wang, Tian et al. (2023) proposed the TEDT framework, which, through a Transformer-based modality-enhancing module, translates non-linguistic modalities into linguistic modalities while filtering out erroneous information between modalities. However, in the process of acquiring cross-modal interaction information, the aforementioned studies tend to focus on obtaining contextual information expressing similar sentiments between modalities while overlooking contextual information expressing differential sentiments. Additionally, the imbalance between modalities poses a significant challenge to the quality of cross-modal contextual interaction.

2.3. Issues of acquiring cross-modal interaction information

Some researchers are attempting to address the issue of poor contextual interaction caused by emotional gaps between modalities. The RAVEN model proposed by Wang et al. (2019) utilizes cross-modal attention to integrate relevant non-verbal information with language representations. The MAG model by Rahman et al. (2020) uses acoustic and visual representations as auxiliary features, fine-tuning the position of the text representation in the sentiment space. However, the aforementioned studies often leverage non-textual modalities to enrich textual representations with emotional information. Yet, the influence of emotion-poor non-textual modalities on cross-modal context interaction has been overlooked.

To address these issues, we propose a novel approach to multimodal sentiment analysis. We utilize hierarchical disentanglement techniques to decompose modality representations into common, private, and noisy representations through two rounds of factorization. Different loss functions constrain these representations, enabling the learning of modality representations that encompass aspects of commonality, individuality, and noise. Subsequently, we enhance acoustic and visual representations by extracting related sentiment information from textual representations through an emotion correlation mining network. Furthermore, using the dual-channel concept, after completing the exploration of intra-modality contextual information, we simultaneously explore inter-modality positive and negative correlation contextual information within two channels.

3. Methodology

In this section, a comprehensive exploration of the various structures of the proposed model will be presented. The overall structure of the model is depicted in Fig. 1 and mainly comprises five parts: Feature Extraction module, HDRD module, IMRE module, DCCMCI module, and Sentiment Prediction module. The multimodal raw data is divided into textual, acoustic and visual modal data, which is then fed into the Feature Extraction module to obtain three modal representations containing both temporal and feature information.

In the HDRD module, the representation of each modality undergoes hierarchical representation disentanglement technology, enabling the learning of both common and private sentiments in the representation while eliminating sentiment-unrelated noise. Moving to the IMRE module, the denoised textual representation is utilized to enhance non-textual representations, thereby enriching the sentiment information embedded in these non-textual representations.

Within the DCCMCI module, the process begins by employing multi-head self-attention to extract contextual sentiment information for each

modality. Subsequently, a dual-channel mechanism is used to separately extract contextually positive and negative correlation sentiment information between modalities. Finally, the various mined sentiment information and the modal representations are fused. This involves concatenating the modality representations, each fused with various sentiment information, to obtain a complete multimodal representation with rich multimodal semantic interactions. This resulting representation is then passed to the Sentiment Prediction module, yielding the ultimate multimodal sentiment prediction outcome.

3.1. Task setup

In the benchmark dataset, each video segment containing a collection of video frames has been assigned an overall emotional label. Consequently, we construct a model that utilizes textual, acoustic, and visual signals within video segments to detect emotional information. Features from different modalities are extracted from each video segment to serve as model inputs.

Let the text representation be denoted as $X_t \in \mathbb{R}^{T \times d_t}$, the audio representation as $X_a \in \mathbb{R}^{T \times d_a}$, and the video representation as $X_v \in \mathbb{R}^{T \times d_v}$, where $T_{m \in \{t,a,v\}}$ represents the sequence length of the corresponding modality, and $d_{m \in \{t,a,v\}}$ represents the dimension. Correspondingly, the model produces a result $\hat{y} \in \mathbb{R}$ representing the sentiment intensity of a video clip. Specifically, $\hat{y} < 0$ signifies that the video expresses negative sentiments, $\hat{y} = 0$ signifies that the video expresses neutral sentiments, and $\hat{y} > 0$ signifies that the video expresses positive sentiments.

3.2. Feature extraction module

In this section, we will provide a detailed explanation of the feature extraction module, wherein the raw data of each modality undergoes processing to extract both featural and temporal information for the respective modality.

Featural information: Traditional methods for textual feature representation often struggle to leverage contextual information for distinguishing polysemy. To overcome this challenge, we adopt a recently successful pre-trained language model, specifically utilizing a BERT model comprising 12 Transformers to extract textual features from the transcripts of video segments. Each layer incorporates a multi-head attention mechanism with 12 heads and a feedforward neural network. This model is proficient in capturing bidirectional contextual information, generating sentence representations imbued with rich sentiment information. Drawing from past experiences, we select the initial word vector from the final layer as the textual representation.

For the acoustic modality, we sample and frame the audio data corresponding to each video segment and extract pitch and spectrum features, such as zero-crossing rate, Mel-Frequency Cepstral Coefficients (MFCCs), and Constant-Q Transform (CQT), from each audio frame. Extensive evidence suggests that this acoustic representation closely relates to the speakers' emotions. Regarding the visual modality, facial information, including expressions, head movements, and eye directions, contains rich emotional cues. Therefore, we utilize OpenFace 2.0 to identify facial cues in each video frame and reduce data volume through average pooling-based frame downsampling. This approach allows us to obtain an ordered set of audio and visual features corresponding to each video segment containing a collection of video frames. These features will then be fed into the LSTM network to capture temporal information.

Temporal information: The textual representation extracted by the BERT model inherently includes temporal information, obviating the need for additional operations. However, the extracted acoustic and visual representations lack temporal information. Consequently, we process these representations through the Bi-LSTM network to obtain temporal information. The operations outlined in Formula (1) result

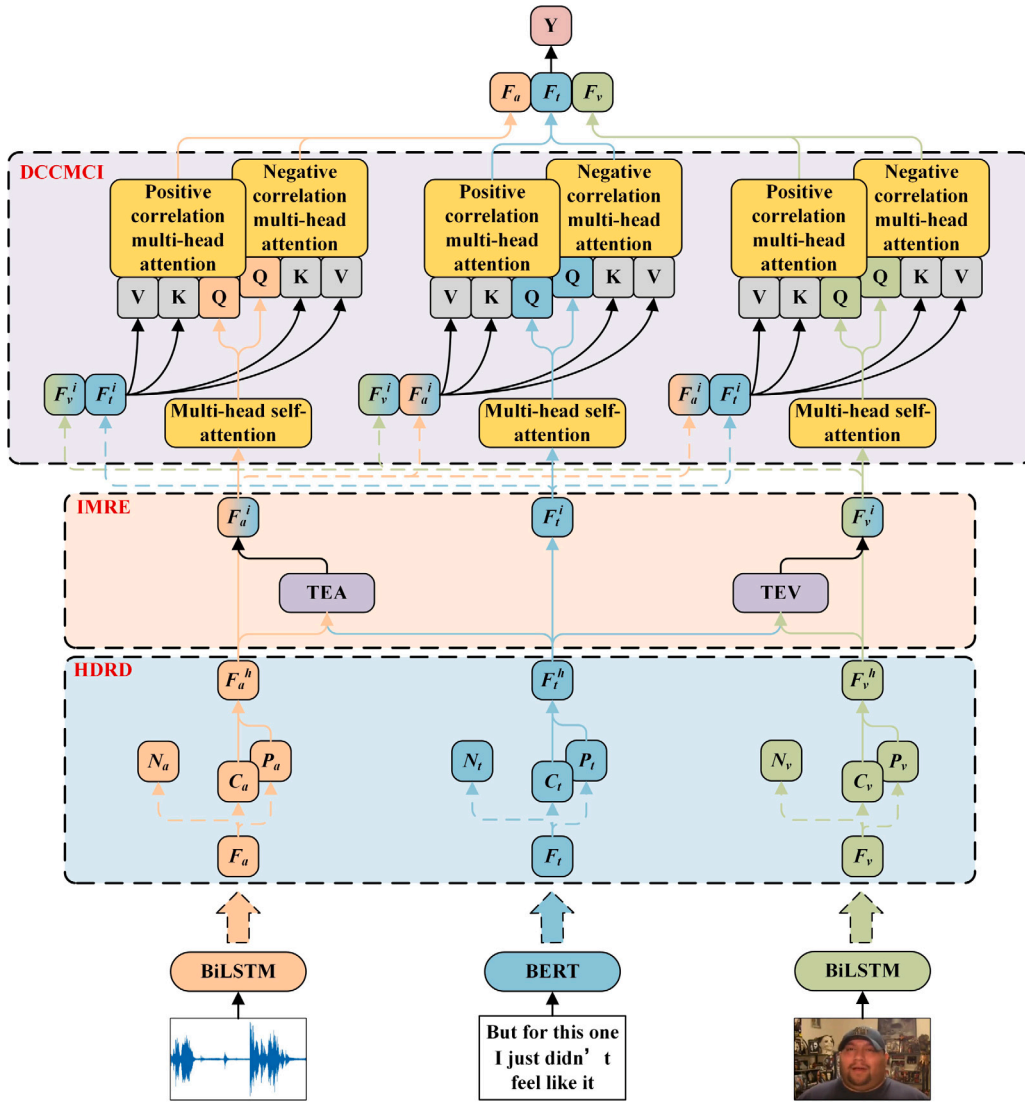


Fig. 1. Overall structure diagram of the proposed model.

in a 768-dimensional textual representation, a 16-dimensional acoustic representation, and a 32-dimensional visual representation.

$$\begin{aligned} F_t &= BERT(X_t; \theta_t^{BERT}) \\ F_a &= BiLSTM(X_a; \theta_a^{BiLSTM}) \\ F_v &= BiLSTM(X_v; \theta_v^{BiLSTM}) \end{aligned} \quad (1)$$

where X_n represents the original features of each modality, while θ_t^{BERT} , θ_a^{BiLSTM} and θ_v^{BiLSTM} represent the network parameters used for feature extraction in each model, $n \in \{t, a, v\}$.

3.3. Hierarchical denoising representation disentanglement module

Following feature extraction, we employed the HDRD module to obtain modal representations that encompass both common and private sentiment information while excluding noise. The HDRD module, as depicted in Fig. 2, primarily consists of two layers of structure. The first layer is the common sentiment learning layer, comprising a common encoder designed to learn a representation capable of expressing common sentiments among modalities. The second layer is the private sentiment learning layer, featuring a private encoder tasked with addressing the issue of noise in modalities by learning a representation capable of expressing private sentiments among modalities.

Our objective is to refine these representations, ensuring that the common representation is homogenized, the private representation is diversified, and the noisy representation is minimized. This process aims to retain diverse sentiment information while reducing noise. Finally, we merge the common representation and private representation into a new modal representation.

Common sentiment learning layer: In the common sentiment learning layer, the central component is the common encoder, as illustrated in Formula (2). The common encoder allows us to derive the common representation for each modality.

$$C_n = L_n^c(F_n; \theta^c) \quad (2)$$

where L_n^c represents the common encoder composed of fully connected layers, which utilizes the same set of parameters θ^c to extract common representations for each modality, $n \in \{t, a, v\}$.

Next, as demonstrated in Formula (3), we will introduce the specific process of the common sentiment learning layer. We decompose the modal representation F_n into common and non-common representations, facilitating the separation of the non-common representation from the modal representation F_n .

$$\begin{aligned} F_n &= C_n + F_n^{unc} \\ F_n^{unc} &= F_n - C_n \end{aligned} \quad (3)$$

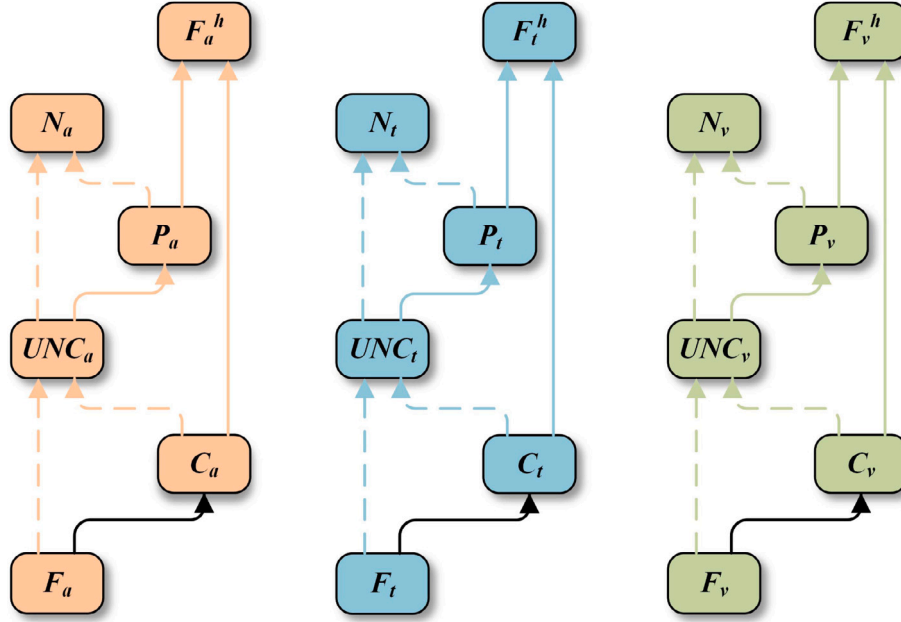


Fig. 2. Hierarchical denoising representation disentanglement framework.

where the common representation C_n contains information expressing common sentiments among various modalities. However, the non-common representation F_n^{unc} is not entirely useless for sentiment analysis; it includes information that reflects private sentiments among modalities. This information is particularly helpful in predicting ironic sentiments in contexts such as satire, $n \in \{t, a, v\}$.

private sentiment learning layer: In the private sentiment learning layer, the primary component is the private encoder, as illustrated in Formula (4). Leveraging the private encoder enables us to obtain the private representation for each modality.

$$P_n = L_n^p(F_n^{\text{unc}}; \theta_n^p) \quad (4)$$

where L_n^p represents the private encoder composed of fully connected layers, and a unique set of parameters θ_n^p is assigned when extracting the private representation for each modality, $n \in \{t, a, v\}$.

The detailed process of the private sentiment learning layer is presented in Formula (5), where the non-common representation F_n^{unc} includes information about private sentiments among modalities that positively contributes to sentiment analysis. Therefore, we decompose the non-common representation F_n^{unc} into private and noisy representations. Subsequently, we separate the noisy representation from the non-common representation F_n^{unc} .

$$\begin{aligned} F_n^{\text{unc}} &= P_n + N_n \\ N_n &= F_n^{\text{unc}} - P_n \end{aligned} \quad (5)$$

where the private representation P_n contains information expressing private sentiments among various modalities. On the other hand, the noisy representation N_n does not contain any sentiment information, such as chaotic backgrounds and changing noise. We believe that such information has no positive impact on sentiment analysis, $n \in \{t, a, v\}$.

Finally, as shown in Formula (6), merge the common representation C_n and private representation P_n into a new modal representation F_n^h .

$$F_n^h = C_n + P_n \quad (6)$$

Constraint condition: We have defined several constraints to ensure the effectiveness of the learned common, private, and noisy representations.

The common loss γ_{cl} represents the difference between common representations among various modalities. The smaller the value, the

more representative the learned common representations are. Therefore, in our work, we use Euclidean Distance (ED) to assess the difference between two representations. It measures the straight-line distance between two vectors in Euclidean space, representing the length of the line connecting these two vectors. As shown in Formula (7), we calculate the sum of the Euclidean distances between common representations of any two modalities as the common loss.

$$\gamma_{cl} = ED(C_t, C_a) + ED(C_t, C_v) + ED(C_a, C_v) \quad (7)$$

The private loss γ_{pl} is used to measure the redundancy among private representations of various modalities. This loss can assess whether the model has learned private representations that can capture the private emotions of modalities. We use an orthogonality constraint to compute this loss. Suppose A and B are two representation matrices whose rows are private representation vectors. The orthogonality constraint can be expressed as $OC(A, B) = \|A^T B\|_F^2$, where $\|*\|_F^2$ is the squared Frobenius norm. The smaller the value of OC, the more orthogonal the two representations of A and B are, indicating a greater difference between A and B. As shown in Formula (8), we calculate the sum of soft orthogonality constraints between private representations of any two modalities as the private loss.

$$\gamma_{pl} = OC(P_t, P_a) + OC(P_t, P_v) + OC(P_a, P_v) \quad (8)$$

The noise loss γ_{nl} is used to assess the magnitude of noisy representations, and the smaller the value, the less noisy is present in noisy representations. We use the squared L2 norm ($\|*\|_2^2$). As shown in Formula (9), we calculate the sum of the squared L2 norms for noisy representations of each modality as the noise loss. The specific hierarchical denoising representation disentanglement strategy is illustrated in Algorithm 1.

$$\gamma_{nl} = \|N_t\|_2^2 + \|N_a\|_2^2 + \|N_v\|_2^2 \quad (9)$$

3.4. Inter-modal representation enhancement module

In multimodal data, the textual modality contains richer sentiment information compared to acoustic and visual modalities. Therefore, in the IMRE module, we use the textual representation obtained from the HDRD module to enhance acoustic and visual representations,

Algorithm 1: Hierarchical Denoising Representation Disentanglement Module (HDRDM)

Input: Modality representations F_n
Output: Denoised modality representations F_n^h

```

1 for  $k \in [1, End]$  do
2   for batch in dataLoader do
3     for  $n \in [t, a, v]$  do
4       Compute common representation  $C_n$  using Equation (2)
5       Compute non-common representation  $F_n^{unc}$  using Equation (3)
6       Compute private representation  $P_n$  using Equation (4)
7       Compute noisy representation  $N_n$  using Equation (5)
8       Compute denoised modality representation  $F_n^h$  using Equation (6)
9     end
10    Compute common loss  $\gamma_{cl}$  using Equation (7)
11    Compute private loss  $\gamma_{pl}$  using Equation (8)
12    Compute noise loss  $\gamma_{nl}$  using Equation (9)
13  end
14 end

```

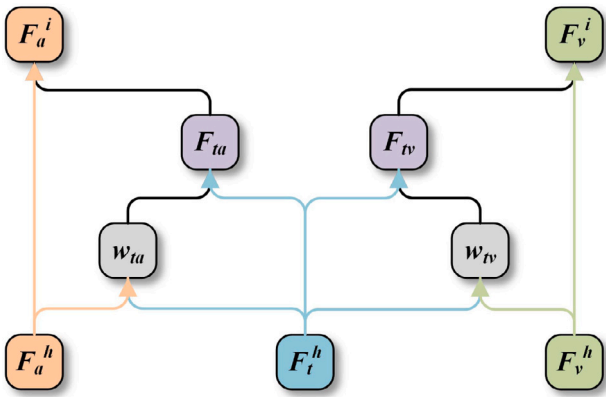


Fig. 3. Inter-modal representation enhancement framework.

addressing the lack of sentiment information in acoustic and visual representations. Fig. 3 illustrates the specific structure of the IMRE module, mainly comprising TEA and TEV components. TEA receives textual and acoustic representations as input, while TEV receives textual and visual representations as input. As outlined in Formula (10), we initiate the process by concatenating the textual representation with acoustic and visual representations to obtain $[F_t^h; F_a^h]$ and $[F_t^h; F_v^h]$. Then, we use them to generate two enhancement factors ω_{ta} and ω_{tv} .

$$\omega_{tm} = \text{RELU}(L([F_t^h; F_m^h]); \theta_{tm}) \quad (10)$$

where $m \in \{a, v\}$, θ_{tm} are network parameters, RELU is a non-linear activation function, and L is a fully connected layer network. These enhancement factors extract sentiment information from the textual representation to strengthen the acoustic and visual representations.

Then, as shown in Formula (11), we blend F_t^h with their respective enhancement factors to obtain emotional vectors F_{ta} and F_{tv} , which are designed to enhance the enhancing acoustic and visual features.

$$F_{tm} = \omega_{tm} * L(F_t^h; \theta_m) \quad (11)$$

where $m \in \{a, v\}$, θ_m are network parameters, and L is a fully connected layer network.

Finally, as outlined in Formula (12), we concatenate and fuse acoustic and visual representations with emotional vectors to obtain new acoustic and visual representations F_a^i and F_v^i . To ensure that the emotional vector remains within an ideal range, we use a scaling factor

φ for constraint.

$$\varphi_m = \min\left(\frac{\|F_m^h\|_2}{\|F_{tm}\|_2}, \mu, 1\right) \quad (12)$$

$$F_m^i = \text{Dropout}\left(\text{LN}([F_m^h; \varphi_m F_{tm}])\right)$$

where $m \in \{a, v\}$, μ is a hyperparameter selected through cross-validation, $\|\cdot\|_2$ is the L2 norm, and Dropout and LN are the dropout layer and normalization layer, respectively.

3.5. Dual channel cross-modal-context interaction module

The DCCMCI module is designed to explore contextual sentiment information from three perspectives: intra-modal, cross-modal positive correlation, and cross-modal negative correlation, with the goal of obtaining comprehensive multimodal representations. As depicted in Fig. 4, first, we use a multi-head self-attention mechanism to explore context information within each modality. Then, for cross-modal context interaction, it is divided into two channels. In the first channel, we apply a positive correlation multi-head attention mechanism to facilitate interaction between contexts that express similar sentiments across different modalities. In another channel, our specially designed negative correlation multi-head attention mechanism is employed to explore context information that expresses differing sentiments between modalities. Our aim is to extract context information expressing intra-modal, cross-modal similarities, and cross-modal differences through context interaction, thereby obtaining a multimodal representation with complete sentiment semantics. Taking the acoustic modality as an example, the operations of the DCCMCI module include the following steps.

Intra-modal context interaction: As shown in Formula (13), we transform the acoustic representation F_a^i into query Q_a , key K_a , and value V_a through a fully connected layer. We calculate the similarity weight matrix using vectors Q_a and K_a , then sum the weighted vector V_a to obtain a new representation vector. Each computation is treated as a separate head, and the outputs of multiple heads are concatenated to obtain intra-modal sentiment representation F_a^s .

$$\begin{aligned} Q_a, K_a, V_a &= F_a^i W_Q, F_a^i W_K, F_a^i W_V \\ \text{head}_j(Q_a, K_a, V_a) &= \text{softmax}\left(\frac{Q_a(K_a)^T}{\sqrt{d}}\right)V_a \end{aligned} \quad (13)$$

$$\begin{aligned} F_a^s &= \text{MHSA}(Q_a, K_a, V_a) \\ &= \text{LN}\left(\text{Dropout}\left(F_a^i + \text{Concat}(\text{head}_1, \dots, \text{head}_h)W_0\right)\right) \end{aligned}$$

where W_Q, W_K, W_V, W_0 represents the corresponding weights, d denotes the dimension, and h is the number of heads.

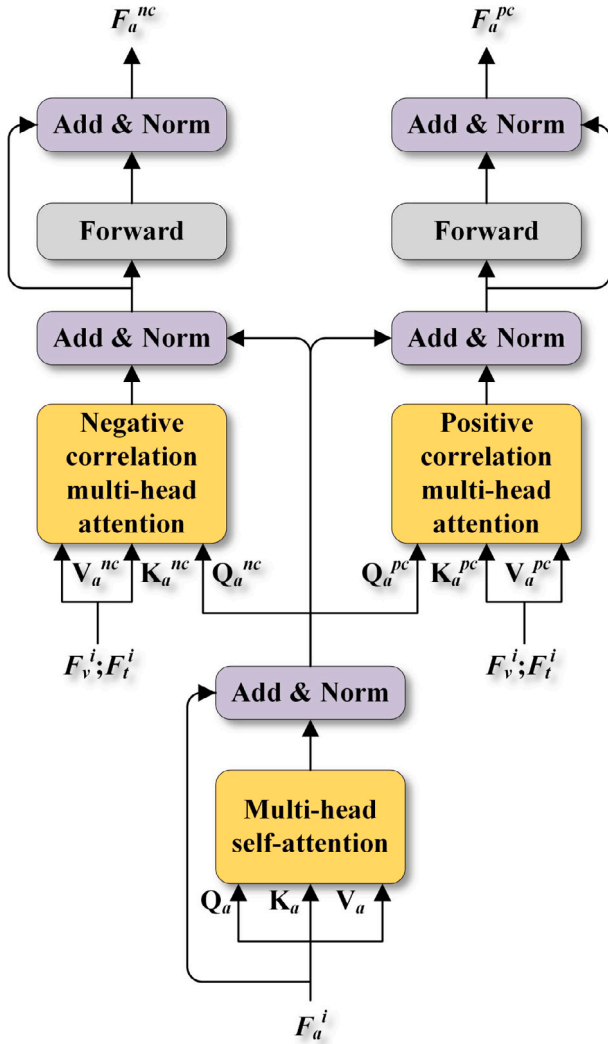


Fig. 4. Dual channel cross-modal-context interaction framework.

Inter-modal context interaction: In the first channel, as shown in Formula (14), we use the acoustic representation F_a^s and the concatenated visual and textual representations $[F_v^i; F_t^i]$ to obtain query Q_a^{pc} , key K_a^{pc} , and value V_a^{pc} . Then, we perform calculations to obtain the cross-modal similar sentiment representation, denoted as F_a^{pc} .

$$Q_a^{pc}, K_a^{pc}, V_a^{pc} = F_a^s W_Q^{pc}, [F_v^i; F_t^i] W_K^{pc}, [F_v^i; F_t^i] W_V^{pc}$$

$$head_j(Q_a^{pc}, K_a^{pc}, V_a^{pc}) = softmax\left(\frac{Q_a^{pc}(K_a^{pc})^T}{\sqrt{d}}\right) V_a^{pc} \quad (14)$$

$$F_a^{pc} = PC - MHA(Q_a^{pc}, K_a^{pc}, V_a^{pc})$$

$$= LN(Dropout(F_a^s + Concat(head_1, \dots, head_h)W_0^{pc}))$$

where $W_Q^{pc}, W_K^{pc}, W_V^{pc}, W_0^{pc}$ represents the corresponding weights.

In the other channel, as depicted in Formula (15), we use the acoustic representation F_a^s and the concatenated visual and textual representations $[F_v^i; F_t^i]$ to obtain query Q_a^{nc} , key K_a^{nc} , and value V_a^{nc} . However, after obtaining the similarity weight matrix, we perform the inverse operation to focus on cross-modal different sentiment information when calculating the weighted sum. Subsequently, we compute the cross-modal differential sentiment representation, which is represented

by F_a^{nc} .

$$Q_a^{nc}, K_a^{nc}, V_a^{nc} = F_a^s W_Q^{nc}, [F_v^i; F_t^i] W_K^{nc}, [F_v^i; F_t^i] W_V^{nc}$$

$$head_j(Q_a^{nc}, K_a^{nc}, V_a^{nc}) = softmax\left(\frac{negate(Q_a^{nc}(K_a^{nc})^T)}{\sqrt{d}}\right) V_a^{nc} \quad (15)$$

$$F_a^{nc} = NC - MHA(Q_a^{nc}, K_a^{nc}, V_a^{nc})$$

$$= LN(Dropout(F_a^s + Concat(head_1, \dots, head_h)W_0^{nc}))$$

where $W_Q^{nc}, W_K^{nc}, W_V^{nc}, W_0^{nc}$ represents the corresponding weights, and $negate()$ denotes the inverse operation.

Finally, after fusing the obtained cross-modal similar and different sentiment representations F_a^{pc} and F_a^{nc} , they are passed through the FFN layer. Additionally, as shown in Formula (16), the output of each layer undergoes residual transformation and normalization.

$$F_a^d = LN(Dropout(F_a^i + FFN(F_a^{pc} + F_a^{nc}))) \quad (16)$$

where FFN represents the feedforward neural network layer.

3.6. Sentiment prediction module

As illustrated in Formula (17), we concatenate the representations of the three modalities and input them into a Multilayer Perceptron (MLP) module for sentiment classification. The MLP module consists of a three-layer network. The first two layers are feed-forward layers utilizing the Rectified Linear Unit (ReLU) activation function. The last layer of the MLP serves as the output layer, directly providing a continuous value representing emotional intensity without employing an activation function.

$$\hat{y} = MLP(Concat(F_t^d, F_a^d, F_v^d)) \quad (17)$$

where \hat{y} is a continuous value representing the sentiment intensity.

The entire model finally finds the best fitting parameters during training by minimizing the overall loss shown in Formula (18).

$$L_{all} = x_s \gamma_{sl} + x_c \gamma_{cl} + x_p \gamma_{pl} + x_n \gamma_{nl} \quad (18)$$

where γ_{sl} represents the sentiment prediction task loss, γ_{cl} represents the common loss, γ_{pl} represents the private loss, γ_{nl} represents the noise loss, and x_s, x_c, x_p and x_n represent the weight of each loss. We employ the mean square error as the loss function for the sentiment prediction task. Additionally, the common loss is calculated as the sum of Euclidean distances between common representations, while the private loss is computed as the sum of soft orthogonal constraints between private representations. Furthermore, the noise loss is determined as the sum of L2 norms of each noise representation.

4. Experiment settings

In this section, we will provide a detailed introduction to the datasets, baselines, and basic settings used in our work.

4.1. Datasets

In this study, we assess the performance of the proposed model through experiments conducted on two widely-used benchmark datasets, CMU-MOSI and CMU-MOSEI, both within the domain of multimodal sentiment analysis. The CMU-MOSI dataset stands as the pioneering corpus for online video sentiment analysis, encompassing 2199 short video clips extracted from 93 movie review videos (Zadeh, Zellers, Pincus, & Morency, 2016). On the other hand, the CMU-MOSEI dataset consists of a larger collection of video samples, comprising over 20,000 video clips extracted from speeches delivered by 1000 different speakers (Zadeh, Liang, Vanbriessen et al., 2018). Additionally, sentiment annotations for the video samples in both CMU-MOSI and CMU-MOSEI datasets are provided within the range of $[-3, 3]$. For further details regarding the statistical breakdown of the training, validation, and testing sets for these datasets, refer to Table 1.

Table 1
Statistics for CMU-MOSI and CMU-MOSEI dataset.

Dataset	Train	Valid	Test	All
CMU-MOSI	1284	299	686	2199
CMU-MOSEI	16326	1871	4659	22856

4.2. Baselines

To validate the efficacy of the proposed model, we conduct comparisons with baselines and state-of-the-art models within the realm of multimodal sentiment analysis.

MuT (Tsai et al., 2019) The Multimodal Transformer leverages paired cross-modal Transformer mechanisms to explore context information between modalities without the need for data alignment.

ICCN (Sun, Sarma, Sethares, & Liang, 2020) The Interaction Canonical Correlation Network leveraged Deep Canonical Correlation Analysis to explore correlations among three modalities.

MISA (Hazarika et al., 2020) The Modality-Invariant and Specific Representations projects modal representations through spatial mapping into modality-invariant and modality-specific subspaces, thereby learning modal representations containing shared and private characteristics.

Self-MM (Yu, Xu, Yuan, & Wu, 2021) The Self-Supervised Multi-Task Learning jointly trains multimodal and unimodal tasks using multimodal labels and generated unimodal labels, enabling the learning of similarities and differences between modalities.

MMIM (Han, Chen, & Poria, 2021) The Multimodal InfoMax framework maximizes mutual information between unimodal representations and multimodal fusion representations in a layered manner, aiming to include as much task-related information as possible in the multimodal fusion representation.

TMPSA (Yang et al., 2022) The Two-Phase Multi-Task Sentiment Analysis framework leverages staged training and multi-task learning strategies to fully exploit the powerful capabilities of pre-trained models while preserving the sentiment information of each representation.

MMCL (Lin & Hu, 2022) The Multimodal Contrastive Learning framework uses unimodal contrastive coding and a pseudo-siamese network to filter out noise from non-text modality and capture the interaction information among modalities.

SenBERT (Fang, Liu, & Zhang, 2022) This Sense-aware BERT network utilizes cross-modal multi-head attention to explore interactions between multimodal data and uses multimodal representations to fine-tune the BERT model.

SIMR (Wang, Wang, Lin, Xu and Guo, 2023) The Speaker-Independent Multimodal Representation Framework decomposes non-verbal representations into personal style coding and sentiment representation to mitigate the influence of individual styles, and uses an enhanced cross-modal Transformer to explore context interaction between modalities.

MUTA-Net (Tang, Xiao et al., 2023) The Modal-Utterance-Temporal Attention Network applies utterance-level representations to interactions between different modalities, minimizing intra-class distance and maximizing inter-class distance to enhance the discriminative power of the representations.

TETFN (Tang, Liu et al., 2023) The Text-Enhanced Transformer Fusion Network obtains effective unified multimodal representations through learning text-based pairwise cross-modal mappings and focuses on inter-modal differences through unimodal labels.

AOBERT (Kim & Park, 2023) The All-modalities-inOne BERT network leverages a single-stream Transformer to integrate three types of modal representations and utilizes multimodal representations for training MMLM and AP tasks.

4.3. Basic settings

Experiment details: In the audio feature extraction module, the sampling rate of audio frames and the hop length between frames are 22.05 kHz and 512 respectively. In the visual feature extraction module, the pooling size of video frames is 5. Our model is trained using the Adam optimizer. Taking into account computing resources and training efficiency, we set the fine-tune range of the batch size to {16, 32}. Considering the convergence speed and stability of the model, we set the initial learning rate of the pre-trained model BERT to $1e-5$, and the fine-tune range of other parameters' learning rate to { $1e-5$, $5e-6$, $1e-6$ }. The weights for commonality, individuality, and noise constraints are selected from the range {0.1, 0.3, 0.5, 0.8}. In order to strike a balance between the generalization ability of the model and the risk of overfitting, we set the Dropout value range in the TEA and TEV components to {0.0, 0.1, 0.3, 0.5}. Experiments are conducted on a TESLA-V100 GPU, utilizing grid search to identify the optimal hyperparameters within the specified ranges.

The total number of parameters in our model is 163.483 million, with approximately 110 million parameters attributed to the pre-trained BERT model used for textual representation extraction. The remaining parameters, excluding BERT, amount to about 53 million. Furthermore, the computational complexity of the model is measured at 6.074 GFLOPs.

Assessment Metrics: We evaluate the model's performance in both regression and classification tasks. In regression tasks, Mean Absolute Error (MAE) and Pearson correlation coefficient (Corr) are employed as assessment metrics. For classification tasks, we derive discrete labels (positive and negative) representing the sentiment polarity of samples from the continuous label values within the $[-3, 3]$ interval. We calculate binary classification accuracy (Acc-2) and F1 score in both negative/non-negative and negative/positive manners as assessment metrics. A smaller MAE and larger values for the other metrics indicate better model performance.

5. Results and analysis

5.1. Quantitative analysis

We comprehensively compare our model with previous works using the respective evaluation metrics of classification and regression tasks. For the assessment metrics of classification tasks, the right side of “/” represents “negative/positive” and the left side represents “negative/non-negative”. The optimal metric values are emphasized in bold.

Table 2 presents the experimental result on the CMU-MOSI dataset. This suggests a notable enhancement in the performance of our model in classification tasks. When assessing the method as ‘negative/positive,’ our model exhibits improvements of 1.89% and 1.87% in Acc-2 and F1 scores, respectively, compared to the suboptimal model (TMPSA). In comparison to the worst-performing model (MuT), these scores have increased by 7.79% and 7.87%, respectively. When assessing the method as ‘negative/non-negative,’ our model shows enhancements of 1.26% and 1.12% in Acc-2 and F1 scores, respectively, compared to the suboptimal model (AOBERT). These improvements rise to 4.66% and 4.82%, respectively, when compared to the worst-performing model (MISA). In the regression task, our model reduces the Mean Absolute Error (MAE) by 0.017 and 0.205, respectively, compared to the suboptimal model (MMIM) and the worst model (MuT). Additionally, compared to the suboptimal model (Sen-BERT) and the worst model (MuT), our model improves the Pearson correlation coefficient (Corr) by 0.01 and 0.129, respectively.

In addition, we also conducted relevant experiments on the CMU-MOSEI dataset. Table 3 presents the experimental results on the CMU-MOSEI dataset. In the ‘negative/positive’ assessment method, our model demonstrates improvements of 0.31% in both Acc-2 and F1

Table 2
Result on the CMU-MOSI dataset.

Model	MAE↓	Corr↑	Acc-2↑	F1-Score↑
MuT (Tsai et al., 2019)	0.889	0.686	-/81.10	-/81.00
ICCN (Sun et al., 2020)	0.862	0.714	-/83.07	-/83.02
MISA (Hazarika et al., 2020)	0.783	0.761	81.80/83.40	81.70/83.60
Self-MM (Yu et al., 2021)	0.713	0.798	84.00/85.98	84.42/85.95
MMIM (Han et al., 2021)	0.701	0.801	84.14/86.06	84.00/85.98
TMPSA (Yang et al., 2022)	0.704	0.799	-/87.00	-/87.00
MMCL (Lin & Hu, 2022)	0.705	0.797	84.00/86.30	83.80/86.20
SenBERT (Fang et al., 2022)	0.702	0.805	83.67/85.37	83.66/85.40
SIMR (Wang, Wang et al., 2023)	0.706	0.798	84.20/86.10	84.00/86.10
MUTA-Net (Tang, Xiao et al., 2023)	0.708	0.798	83.00/84.90	82.90/84.90
TETFN (Tang, Liu et al., 2023)	0.717	0.800	84.05/86.10	83.83/86.07
AOBERT (Kim & Park, 2023)	0.856	0.700	85.20/85.60	85.40/86.40
Our model	0.684	0.815	86.46/88.89	86.52/88.87

Table 3
Result on the CMU-MOSEI dataset.

Model	MAE↓	Corr↑	Acc-2↑	F1-Score↑
MuT (Tsai et al., 2019)	0.591	0.694	-/81.60	-/81.60
ICCN (Sun et al., 2020)	0.565	0.713	-/84.18	-/84.15
MISA (Hazarika et al., 2020)	0.555	0.756	83.60/85.50	83.80/85.30
Self-MM (Yu et al., 2021)	0.530	0.765	82.81/85.17	82.53/85.30
MMIM (Han et al., 2021)	0.526	0.772	82.24/85.97	82.66/85.94
TMPSA (Yang et al., 2022)	0.542	0.770	-/85.60	-/85.60
MMCL (Lin & Hu, 2022)	0.537	0.765	84.80/85.90	84.80/85.70
SenBERT (Fang et al., 2022)	0.534	0.768	84.57/85.39	84.59/85.15
SIMR (Wang, Wang et al., 2023)	0.580	0.696	82.50/82.90	81.90/82.90
MUTA-Net (Tang, Xiao et al., 2023)	0.537	0.764	81.90/85.20	82.30/85.20
TETFN (Tang, Liu et al., 2023)	0.551	0.748	84.25/85.18	84.18/85.27
AOBERT (Kim & Park, 2023)	0.515	0.763	84.90/86.20	85.00/85.90
Our model	0.514	0.761	85.46/86.51	85.47/86.23

scores compared to the suboptimal model (AOBERT). When compared to the worst-performing model (SIMR), these scores increase by 3.61% and 3.33%, respectively. When evaluating the ‘negative/non-negative’ method, our model exhibits enhancements of 0.56% and 0.47% in Acc-2 and F1 scores, respectively, compared to the suboptimal model (AOBERT). These improvements rise to 3.35% and 3.17%, respectively, when compared to the worst-performing model (MUTA-Net). In regression tasks, the Mean Absolute Error (MAE) of our model decreases by 0.001 and 0.066, respectively, compared to the suboptimal model (AOBERT) and the worst model (SIMR). Additionally, compared to the worst model (MuT), the Pearson correlation coefficient (Corr) of our model improves by 0.065.

We attribute the enhanced performance of our model to several key factors: The hierarchical denoising representation disentanglement network effectively reduces irrelevant noise in modal representations, preserving detailed sentiment features such as individuality and commonality. Additionally, the cross-modal representation enhancement network significantly augments sentiment information in acoustic and visual representations. Moreover, the dual-channel cross-modal contextual interaction network can simultaneously explore inter-modal positive and negative correlation sentiment information, enhancing the sentiment semantics of multimodal representations.

To validate the predictive performance of the model, we selected five video segments from the CMU-MOSI dataset for case studies. Fig. 5 illustrates the sentiment prediction results for each example. On the left side are the textual, acoustic, and visual data from video segments, while on the right side are the label values and model prediction values. Red indicates negative sentiment, green indicates positive sentiment, and white indicates neutral sentiment. This figure demonstrates that the predicted values are generally consistent with the corresponding label values, providing an intuitive showcase of the effectiveness of the proposed model.

5.2. Ablation study

The proposed model comprises three main components: HDRD, IMRE, and DCCMCI. To delve into the internal mechanisms of these

components and understand their contributions to the model, comprehensive ablation experiments were conducted using the CMU-MOSI dataset.

In these experiments, we first verified the efficacy of each internal component by selectively excluding them while maintaining the overall model structure. In the absence of HDRD, we mapped the extracted modal representations to the same feature space and eliminated the three representation constraint losses in the overall loss function. In the absence of IMRE, we used non-textual representations to enhance themselves instead of utilizing textual representation. In the absence of DCCMCI, we no longer performed inter-modal and intra-modal contextual interactions on modal representations, but simply concatenated them and fed them into the sentiment prediction network.

The experimental results in Table 4 indicate that when any internal component is eliminated, the model’s performance decreases to varying degrees. This confirms that each component is essential for enhancing the model’s performance. Specifically, when IMRE is removed, the Acc-2 and F1-Score of this model decrease by 1.74%/2.32% and 1.79%/2.35%, respectively. Removing HDRD results in a more significant drop, with the Acc-2 and F1-Score decreasing by 5.67%/5.09% and 5.6%/5.06%, respectively. Similarly, removing DCCMCI leads to a considerable decrease, with the Acc-2 and F1-Score dropping by 6.11%/5.56% and 6.03%/5.51%, respectively. These findings indicate a significant degradation in model performance when these components are removed. This suggests that there is a considerable amount of interference noise in multimodal data, reducing the predictive accuracy of the model. It also highlights that both cross-modal positive and negative correlation contextual interactions between multimodal data contribute to improving the accuracy of emotion prediction.

In the HDRD component, effective modal representations are obtained by applying different representation constraints to the separated representations. To further evaluate the efficacy of the HDRD component, we designed an ablation study to verify the effects of various representation constraints. In each round of model training, one of the three representation constraint losses was removed sequentially.






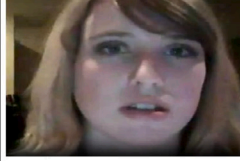

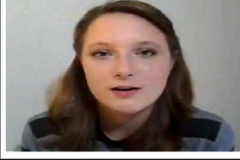
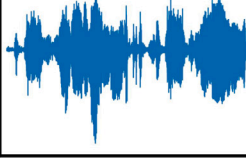
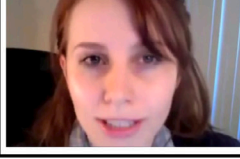
#	Text data	Audio data	Video data	Label	Output
Case1	Release the cracken on this movie dont go see it go see...			-2.8 (Negative)	-2.81904 (Negative)
Case2	And um while the trailer made the film look real nice and cute			-1.6 (Negative)	-1.64318 (Negative)
Case3	And there is like someone there was a lot of action			0.0 (Neutral)	0.044636 (Neutral)
Case4	And um michael sheen as aro as well he was rather enjoyable to watch			1.8 (Positive)	1.762993 (Positive)
Case5	Its extremely well made			2.8 (Positive)	2.829079 (Positive)

Fig. 5. Case study.

Table 4
Ablation experiment results for each component.

Model	MAE↓	Corr↑	Acc-2↑	F1-Score↑
w/o HDRD	0.709	0.803	80.79/83.80	80.92/83.81
w/o IMRE	0.699	0.809	84.72/86.57	84.73/86.52
w/o DCCMCI	0.709	0.799	80.35/83.33	80.49/83.36
Our model	0.684	0.815	86.46/88.89	86.52/88.87

The experimental results in Table 5 show that when we discard any one of the representation constraint losses, the model’s performance declines to varying degrees. When the private loss is removed, the Acc-2 and F1-Score of the model decrease by 1.74%/1.39% and 1.71%/1.36%, respectively. Removing the common loss results in a more significant drop, with the Acc-2 and F1-Score decreasing by 3.49%/3.24% and 3.46%/3.23%, respectively. Similarly, removing the noise loss leads to a noticeable decrease, with the Acc-2 and F1-Score dropping by 2.62%/2.78% and 2.57%/2.74%, respectively. These findings highlight the importance of each loss component, as the performance degradation is noticeable when any of them is removed. This indicates that the model relies more on the constraints of common and noise loss than on the private loss. When all representation constraints are in effect, the model’s predictive performance is optimal. This proves that representation learning results can be more effective through constraints of common, private, and noisy, thereby improving the model’s performance.

Additionally, the weight values of representation constraints play a crucial role in the model. To investigate the impact of different weight values of representation constraints on model performance, this

Table 5
Ablation experiment results for each constraint.

Model	MAE↓	Corr↑	Acc-2↑	F1-Score↑
w/o γ_{cl}	0.717	0.800	82.97/85.65	83.06/85.64
w/o γ_{pl}	0.694	0.809	84.72/87.50	84.81/87.51
w/o γ_{nl}	0.695	0.809	83.84/86.11	83.95/86.13
Our model	0.684	0.815	86.46/88.89	86.52/88.87

study conducted ablation experiments under various weight values for each representation constraint. The experimental results are shown in Fig. 6. Regarding the weight x_c of the common constraint, when x_c is set to 0, it signifies that the model does not utilize common constraints for representation learning, resulting in a significant amount of redundant information in the modal representation. In this case, the Acc-2 of the model dropped by 3.24%, while the Mean Absolute Error (MAE) increased by 0.033. Overall, the model’s performance was not satisfactory. However, when x_c is not equal to 0, the model’s performance improves. Specifically, when $x_c = 0.1$, the model achieves optimal performance.

Regarding the weight x_p of the private constraint, when x_p is set to 0, it signifies that the private constraint in the model is inactive, leading to ineffective discrimination of unique attributes in the modal representation. In this situation, the Acc-2 of this model has dropped to 87.50%, while the Mean Absolute Error (MAE) has increased to 0.694. Overall, the model’s performance is average. However, when x_p is not equal to 0, the model’s performance increases slightly, reaching its highest accuracy when $x_p = 0.8$.

For the weight x_n of the noisy constraint, when x_n is 0, it indicates that the model does not address interference noise in the

Table 6
Experimental results of ablation studies for each channel.

Model	MAE↓	Corr↑	Acc-2↑	F1-Score↑
w/o PC-MHA	0.691	0.810	83.41/85.65	83.43/85.59
w/o NC-MHA	0.694	0.814	84.15/87.50	85.24/87.51
Our model	0.684	0.815	86.46/88.89	86.52/88.87

representation, resulting in a significant presence of noise features in the modal representation. In this circumstance, the Acc-2 of the model has dropped to 86.11%, the Mean Absolute Error (MAE) has increased to 0.695. Thus, the overall performance of the model degrades significantly. However, when x_n is not 0, the noisy constraint reduces noise in the representation, improving the model's performance. The model achieves optimal performance when $x_n = 0.1$. For different representation constraints, the performance of models with non-zero weight values are consistently superior to that of models with weight values equal to zero. This demonstrates the positive impact of common, private, and noisy constraints on representation learning. Moreover, appropriate constraint weight values further enhance the model's performance.

In the DCCMCI component, a dual-channel approach is employed, utilizing the PC-MHA and NC-MHA mechanisms to explore contextual interactions, both positive and negative correlation, between modalities. To further validate the effectiveness of the DCCMCI component, we conducted ablation experiments by exploring modal contextual interactions solely through either the PC-MHA or NC-MHA mechanism, employing them as single-channel exploration modes.

The experimental results in Table 6 show that when using a single channel to explore contextual interactions between modalities, models using only the NC-MHA mechanism for information mining exhibit a more pronounced performance decline compared to models using only the PC-MHA mechanism. When only PC-MHA is used, the Acc-2 and F1-Score of the model have decreased by 2.31%/1.39% and 1.28%/1.36% respectively. When only using NC-MHA, the Acc-2 and F1-Score of the model have dropped by 3.05%/3.24% and 3.09%/3.28% respectively. This indicates that the contribution of contextually positive correlation interactions between modalities in sentiment analysis remains unique. Additionally, negative correlation contextual interactions between modalities also have a positive impact on model performance improvement. The model achieves optimal performance when both NC-MHA and PC-MHA mechanisms are simultaneously employed. This finding demonstrates that exploring both positive and negative correlation contextual interactions between modalities enriches the modal representation with more effective sentiment information, thereby enhancing model performance.

The exploration of contextual interactions between modalities through the PC-MHA and NC-MHA mechanisms in this study is an adaptive iterative process. To further explore the impact of different numbers of iterations on model performance, we set the number of iterations to range from 1 to 5 and retrained the model on the MOSI dataset. According to the data presented in Fig. 7, the model achieves its optimal performance when the iteration number is configured to 4. However, for other values of the number of iterations, the model's performance is adversely affected to varying degrees. When the number of iterations is 3, the Acc-2 and F1-Score of the model have dropped by 4.63% and 4.63% respectively, the Mean Absolute Error (MAE) has increased to 0.720, and the F1-Score has dropped to 0.798.

In fact, our goal is to obtain more distinctive modal representations by appropriately stacking iteratively and dynamically updating the mined cross-modal contextual interaction information. Therefore, the above experimental results indicate that too few iterations may not effectively unearth more distinctive emotional context. Conversely, excessive iterations may lead to the extraction of incorrect cross-modal emotional context due to sentiment bias between modalities.

Table 7
Experimental results of ablation studies for modality contribution.

Model	MAE↓	Corr↑	Acc-2↑	F1-Score↑
w/o text	0.731	0.795	81.66/84.72	81.78/84.73
w/o audio	0.717	0.794	81.66/84.26	81.78/84.28
w/o video	0.708	0.802	85.15/87.50	85.22/87.49
Our model	0.684	0.815	86.46/88.89	86.52/88.87

To explore the contribution of different modalities to the proposed model, we conducted ablation experiments by sequentially removing one of the three modalities and adjusting the network structure accordingly. Specifically, when the textual modality is retained, we maintain the IMRE module in the model. However, if the textual modality is removed, we exclude the IMRE module. Regardless of which modality is removed, we only consider the remaining two modalities when calculating the representation constraints.

The experimental results in Table 7 show that deleting any modality will lead to varying degrees of degradation in model performance. When the visual modality is deleted, the Acc-2 and F1-Score of the model have decreased by 1.31%/1.39% and 1.3%/1.38% respectively. When the textual modality is deleted, the Acc-2 and F1-Score of the model have dropped by 4.8%/4.17% and 4.74%/4.14% respectively. When the audio modality is deleted, the Acc-2 and F1-Score of the model have dropped by 4.8%/4.63% and 4.74%/4.59% respectively. These findings indicate a significant loss in model performance in all cases. It suggests that each modality contributes to the model's performance, but the textual and audio modalities contribute more significantly than the visual modality.

To demonstrate the excellent performance of our model in sentiment prediction tasks, we utilize the t-SNE method to visualize the fused representations learned by the model. The visualization results are shown in Fig. 8. We map predicted sentiment labels within the range $[-3, 3]$ to $[0, 1]$, where $[0, 0.5]$ represents positive sentiment representation, and $[0.5, 1]$ represents negative sentiment representation.

The four subplots in Fig. 8 represent the distribution of fused representations during the training process when the epoch is 1, 15, 30, and final, respectively. Through observation, it is evident that at epoch = 1, the fused representations representing different sentiment polarities exhibit a mixed distribution, indicating poor distinctiveness in the representations learned by the untrained model. When epoch = 15, fused representations representing the same sentiment polarity begin to cluster, indicating that the trained model can learn associative information among similar samples. When epoch = 30, the distribution of fused representations becomes more compact and separable, suggesting that the model, after further training, can accurately recognize more detailed information among different samples. When epoch = final, positive and negative fused representations form two independent distribution clusters, indicating that the model, after training completion, possesses good emotion classification capability. This also indirectly reflects the effectiveness of the proposed model in sentiment prediction.

6. Conclusion and future work

In this paper, we present a multi-modal sentiment analysis model. Utilizing hierarchical disentanglement techniques and employing commonality and individuality encoders, we learn modal representations that encompass common sentiments, private sentiments, and interference noise simultaneously. Moreover, to tackle the challenge of disparate sentiment information distribution between modalities arising from varying data quality, we enrich the sentiment semantics of acoustic and visual representations by extracting relevant sentiment information from textual representations. Additionally, we incorporate a dual-channel mechanism and utilize distinct multi-head attentions to explore contextual sentiment information between modalities, both positive and negative correlation. This addresses the issue of inadequate

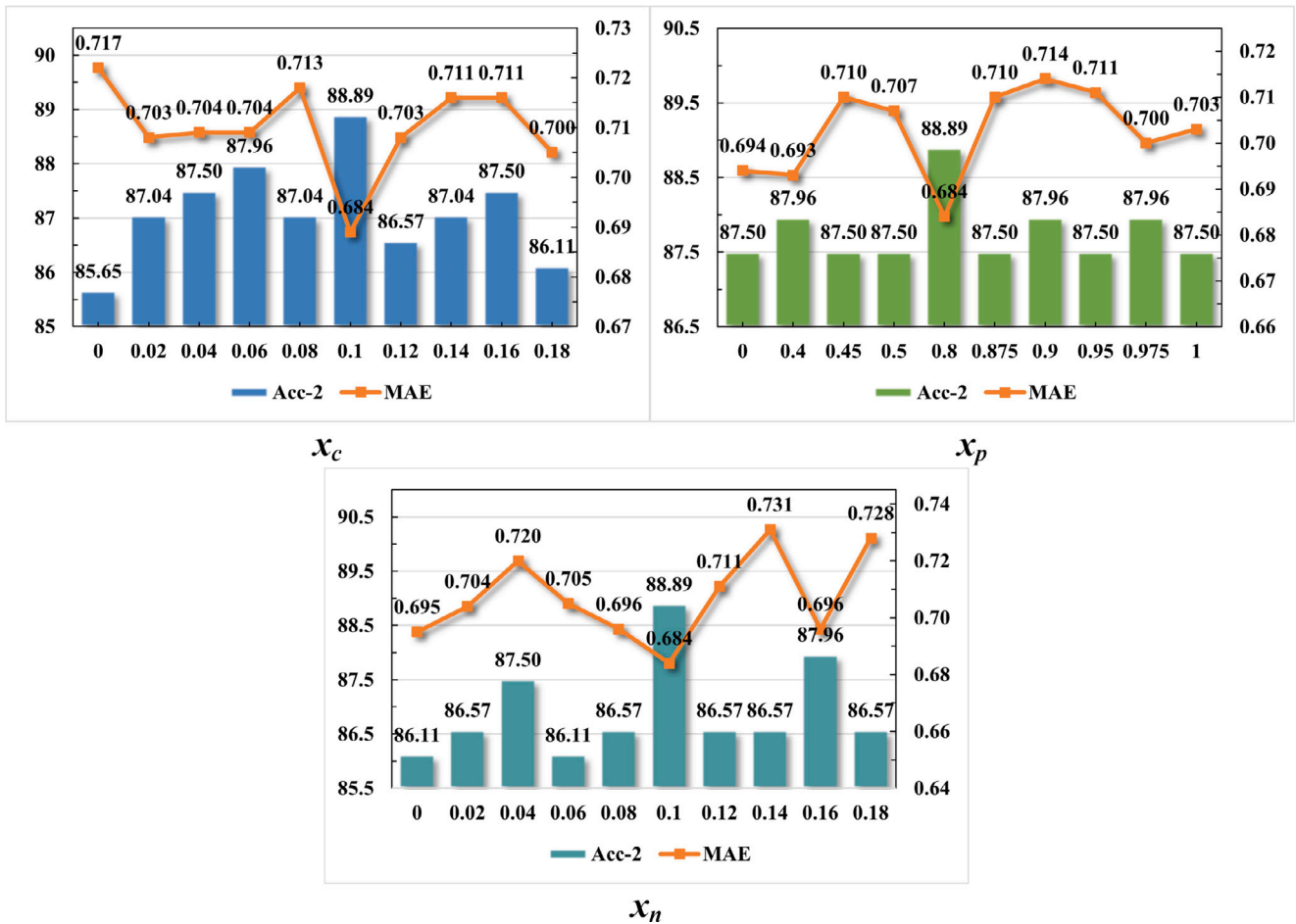


Fig. 6. Experimental results of ablation studies for different values of x_c , x_p and x_n .

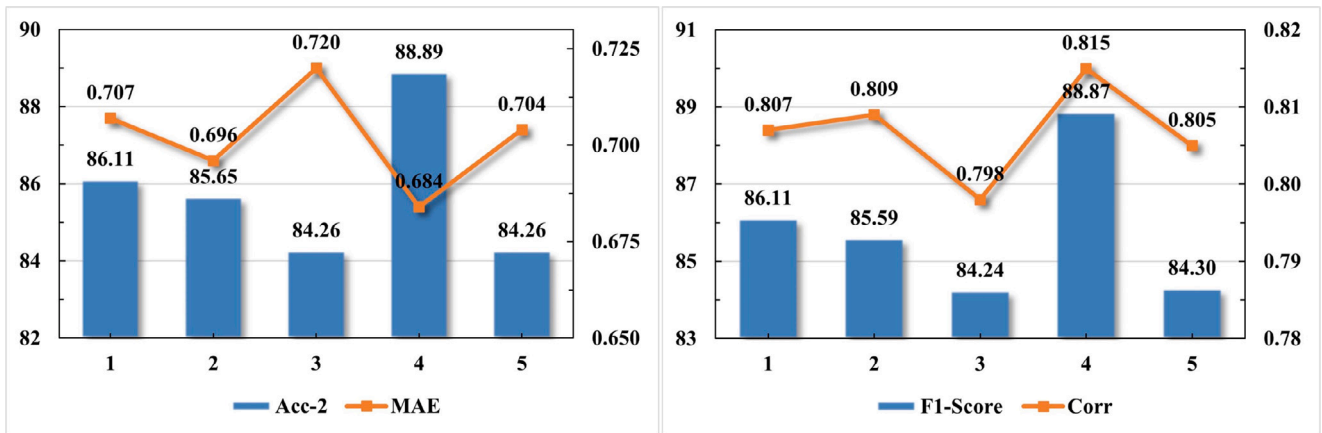


Fig. 7. Experimental results of ablation studies for different numbers of iterations.

attention to diverse contextual information between modalities. The combination of experimental results suggests that our proposed model achieves competitive performance on both benchmark datasets. For future work, our main objective is to enhance the model’s accuracy by integrating a multi-task learning framework. Furthermore, we aim to optimize the model for a more lightweight design while preserving its accuracy.

CRedit authorship contribution statement

Zuhe Li: Development or design of methodology, Creation of models, Conducting a research and investigation process, Specifically performing the experiments, or data/evidence collection, Preparation, Creation and/or presentation of the published work, Specifically writing

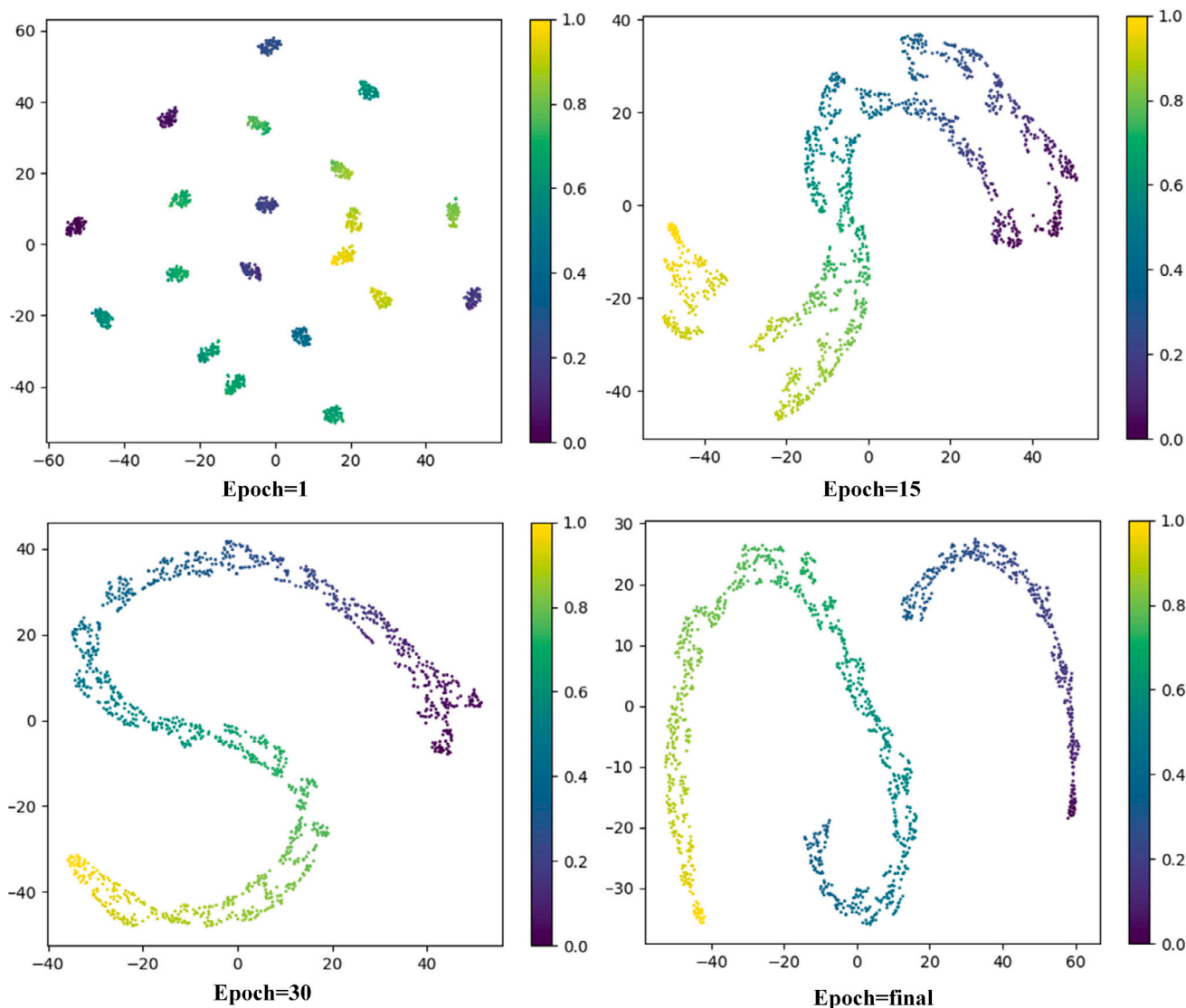


Fig. 8. t-SNE visualization of the fused representations learned by the model.

the initial draft. **Zhenwei Huang:** Programming, Software development, Designing computer programs, Implementation of the computer code and supporting algorithms, Testing of existing code components. **Yushan Pan:** Ideas formulation or evolution of overarching research goals and aims, Development or design of methodology creation of models, Oversight and leadership responsibility for the research activity planning and execution. **Jun Yu:** Preparation, creation and/or presentation of the published work, Specifically writing the initial draft. **Weihua Liu:** Management activities to annotate, Scrub data and maintain research data for initial use and later reuse. **Haoran Chen:** Provision of study materials, Computing resources. **Yiming Luo:** Preparation, creation and/or presentation of the published work, Specifically critical review, Commentary or revision. **Di Wu:** Preparation, creation and/or presentation of the published work, Specifically critical review, Commentary or revision. **Hao Wang:** Verification.

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

This paper was supported by the National Natural Science Foundation of China under Grant 61702462, the Natural Science Foundation of Henan, China under Grant 242300421220, the XJTLU RDF-21-02-008, the Henan Provincial Science and Technology Research Project under Grants 242102211007, 242102211020, 232102211006, 232102210044 and 232102211017, the Science and Technology Innovation Project of Zhengzhou University of Light Industry, China under Grant 23XNKJTD0205, the Undergraduate Universities Smart Teaching Special Research Project of Henan Province under Grant Jiao Gao [2021] No. 489–29, the Doctor Natural Science Foundation of Zhengzhou University of Light Industry under Grants 2021BSJJ025 and 2022BSJJZK13.

References

- Abdu, Sarah A., Yousef, Ahmed H., & Salem, Ashraf. (2021). Multimodal video sentiment analysis using deep learning approaches, a survey. *Information Fusion*, 76, 204–226.

- Biswas, Debojyoti, & Tešić, Jelena (2022). Small object difficulty (sod) modeling for objects detection in satellite images. In *Proceedings of the 2022 14th international conference on computational intelligence and communication networks* (pp. 125–130).
- Bousmalis, Konstantinos, Trigeorgis, George, Silberman, Nathan, Krishnan, Dilip, & Erhan, Dumitru (2016). Domain separation networks. *Vol. 29*, In *Proceedings of the 30th advances in neural information processing systems*.
- Chen, Chen, Hong, Hansheng, Guo, Jie, & Song, Bin (2023). Inter-intra modal representation augmentation with trimodal collaborative disentanglement network for multimodal sentiment analysis. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 31, 1476–1488.
- Chen, Hao, Shen, Feihong, Ding, Ding, Deng, Yongjian, & Li, Chao (2024). Disentangled cross-modal transformer for RGB-d salient object detection and beyond. *IEEE Transactions on Image Processing*, 33, 1699–1709.
- Fang, Lingyong, Liu, Gongshen, & Zhang, Ru (2022). Sense-aware BERT and multi-task fine-tuning for multimodal sentiment analysis. In *Proceedings of the 2022 international joint conference on neural networks* (pp. 1–8).
- Han, Wei, Chen, Hui, & Poria, Soujanya (2021). Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 9180–9192).
- Hazarika, Devamanyu, Zimmermann, Roger, & Poria, Soujanya (2020). Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 1122–1131).
- Kim, Kyeonghun, & Park, Sanghyun (2023). AOBERT: All-modalities-in-one BERT for multimodal sentiment analysis. *Information Fusion*, 92, 37–45.
- Li, Zheng, Cai, Weibo, Dong, Junhao, Lai, Jianhuang, & Xie, Xiaohua (2023). Feature disentanglement and adaptive fusion for improving multi-modal tracking. In *Proceedings of the Chinese conference on pattern recognition and computer vision* (pp. 68–80).
- Lin, Ronghao, & Hu, Haifeng (2022). Multimodal contrastive learning via unimodal coding and cross-modal prediction for multimodal sentiment analysis. arXiv preprint arXiv:2210.14556.
- Liu, Zhun, Shen, Ying, Lakshminarasimhan, Varun Bharadhwaj, Liang, Paul Pu, Zadeh, Amir, & Morency, Louis-Philippe (2018). Efficient low-rank multimodal fusion with modality-specific factors. *Vol. 1*, In *Proceedings of the 56th annual meeting of the association-for-computational-linguistics* (pp. 2247–2256).
- Michalis, Angelou, Vassilis, Solachidis, Nicholas, Vretos, & Petros, Daras (2019). Graph-based multimodal fusion with metric learning for multimodal classification. *Pattern Recognition*, 95, 296–307.
- Ou, Yangjun, Chen, Zhenzhong, & Wu, Feng (2021). Multimodal local-global attention network for affective video content analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(5), 1901–1914.
- Rahman, Wasifur, Hasan, Md Kamrul, Lee, Sangwu, Zadeh, Amir, Mao, Chengfeng, Morency, Louis-Philippe, et al. (2020). Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 2359–2369).
- Rajagopalan, Shyam Sundar, Morency, Louis-Philippe, Baltruaitis, Tadas, & Goecke, Roland (2016). Extending long short-term memory for multi-view structured learning. *Vol. 9911*, In *Proceedings of the 14th European conference on computer vision* (pp. 338–353).
- Shi, Qianqian, Fan, Junsong, Wang, Zuoren, & Zhang, Zhaoxiang (2022). Multi-modal channel-wise attention transformer inspired by multisensory integration mechanisms of the brain. *Pattern Recognition*, 130, Article 108837.
- Su, Yuanhang, & Kuo, C. C. Jay (2022). Recurrent neural networks and their memory behavior: A survey. *APSIPA Transactions on Signal and Information Processing*, 11(1), Article e26.
- Sun, Zhongkai, Sarma, Prathusha K., Sethares, William A., & Liang, Yingyu (2020). Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. *Vol. 34*, In *Proceedings of the 34th AAAI conference on artificial intelligence* (pp. 8992–8999).
- Tang, Jiajia, Liu, Dongjun, Jin, Xuanyu, Peng, Yong, Zhao, Qibin, Ding, Yu, et al. (2023). BAFN: Bi-direction attention based fusion network for multimodal sentiment analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(4), 1966–1978.
- Tang, Zemin, Xiao, Qi, Zhou, Xu, Li, Yangfan, Chen, Cen, & Li, Kenli (2023). Learning discriminative multi-relation representations for multimodal sentiment analysis. *Information Sciences*, 641, Article 119125.
- Tsai, Yao-Hung Hubert, Bai, Shaojie, Liang, Paul Pu, Kolter, J. Zico, Morency, Louis-Philippe, & Salakhutdinov, Ruslan (2019). Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 6558–6569).
- Tsai, Yao-Hung Hubert, Liang, Paul Pu, Zadeh, Amir, Morency, Louis-Philippe, & Salakhutdinov, Ruslan (2018). Learning factorized multimodal representations. arXiv preprint arXiv:1806.06176.
- Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., et al. (2017). Attention is all you need. *Vol. 30*, In *Proceedings of the 31st annual conference on neural information processing systems*.
- Wang, Di, Guo, Xutong, Tian, Yumin, Liu, Jinhui, He, LiHuo, & Luo, Xuemei (2023). TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis. *Pattern Recognition*, 136, Article 109259.
- Wang, Yixin, Qiu, Shuang, Li, Dan, Du, Changde, Lu, Bao-Liang, & He, Huiguang (2022). Multi-modal domain adaptation variational autoencoder for EEG-based emotion recognition. *IEEE/CAA Journal of Automatica Sinica*, 9(9), 1612–1626.
- Wang, Yansen, Shen, Ying, Liu, Zhun, Liang, Paul Pu, Zadeh, Amir, & Morency, Louis-Philippe (2019). Words can shift: Dynamically adjusting word representations using nonverbal behaviors. *Vol. 33*, In *Proceedings of the 33rd AAAI conference on artificial intelligence* (pp. 7216–7223).
- Wang, Fan, Tian, Shengwei, Yu, Long, Liu, Jing, Wang, Junwen, Li, Kun, et al. (2023). TEDT: Transformer-based EncodingDecoding translation network for multimodal sentiment analysis. *Cognitive Computation*, 15(1), 289–303.
- Wang, Jianwen, Wang, Shiping, Lin, Mingwei, Xu, Zeshui, & Guo, Wenzhong (2023). Learning speaker-independent multimodal representation for sentiment analysis. *Information Sciences*, 628, 208–225.
- Wang, Weiran, Yan, Xinchun, Lee, Honglak, & Livescu, Karen (2016). Deep variational canonical correlation analysis. arXiv preprint arXiv:1610.03454.
- Wu, Yang, Lin, Zijie, Zhao, Yanyan, Qin, Bing, & Zhu, Li-Nan (2021). A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing ACL-IJCNLP*, (pp. 4730–4738).
- Xiao, Guorong, Tu, Geng, Zheng, Lin, Zhou, Teng, Li, Xin, Ahmed, Syed Hassan, et al. (2021). Multimodality sentiment analysis in social internet of things based on hierarchical attentions and CSAT-TCN with MBM network. *IEEE Internet of Things Journal*, 8(16), 12748–12757.
- Yang, Bo, Wu, Lijun, Zhu, Jinhua, Shao, Bo, Lin, Xiaola, & Liu, Tie-Yan (2022). Multimodal sentiment analysis with two-phase multi-task learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30, 2015–2024.
- Yu, Wenmeng, Xu, Hua, Yuan, Ziqi, & Wu, Jiele (2021). Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. *Vol. 35*, In *Proceedings of the 35th AAAI conference on artificial intelligence* (pp. 10790–10797).
- Yu, Zhou, Yu, Jun, Fan, Jianping, & Tao, Dacheng (2017). Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision* (pp. 1821–1830).
- Zadeh, Amir, Chen, Minghai, Cambria, Erik, Poria, Soujanya, & Morency, Louis-Philippe (2017). Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing, proceedings* (pp. 1103–1114).
- Zadeh, Amir, Liang, Paul Pu, Poria, Soujanya, Vij, Prateek, Cambria, Erik, & Morency, Louis-Philippe (2018). Multi-attention recurrent network for human communication comprehension. *Vol. 32*, In *Proceedings of the 32th AAAI conference on artificial intelligence* (pp. 5642–5649).
- Zadeh, Amir, Liang, Paul Pu, Vanbriesen, Jonathan, Poria, Soujanya, Tong, Edmund, Cambria, Erik, et al. (2018). Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. *Vol. 1*, In *Proceedings of the 56th annual meeting of the association for computational linguistics* (pp. 2236–2246).
- Zadeh, Amir, Zellers, Rowan, Pincus, Eli, & Morency, Louis-Philippe (2016). Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv preprint arXiv:1606.06259.
- Zhang, Yaojie, Xu, Bing, & Zhao, Tiejun (2020). Convolutional multi-head self-attention on memory for aspect sentiment classification. *IEEE/CAA Journal of Automatica Sinica*, 7(4), 1038–1044.
- Zhao, Sicheng, Jia, Guoli, Yang, Jufeng, Ding, Guiguang, & Keutzer, Kurt (2021). Emotion recognition from multiple modalities: Fundamentals and methodologies. *IEEE Signal Processing Magazine*, 38(6), 59–73.
- Zhu, Linan, Zhu, Zhechao, Zhang, Chenwei, Xu, Yifei, & Kong, Xiangjie (2023). Multimodal sentiment analysis based on fusion methods: A survey. *Information Fusion*, 95, 306–325.