Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.066476

ARTICLE

TGICP: A Text-Gated Interaction Network with Inter-Sample Commonality Perception for Multimodal Sentiment Analysis

Erlin Tian[1], Shuai Zhao[2,*], Min Huang[2], Yushan Pan[3,4], Yihong Wang[3,4] and Zuhe Li[1]

[1]School of Computer Science and Technology, Zhengzhou University of Light Industry, Zhengzhou, 450002, China
[2]School of Software, Zhengzhou University of Light Industry, Zhengzhou, 450002, China
[3]Department of Computing, Xi'an Jiaotong-Liverpool University, Suzhou, 215123, China
[4]Department of Computer Science, University of Liverpool, Liverpool, L69 7ZX, UK

*Corresponding Author: Shuai Zhao. Email: 332316040990@zzuli.edu.cn

Received: 09 April 2025; Accepted: 03 July 2025; Published: 29 August 2025

ABSTRACT: With the increasing importance of multimodal data in emotional expression on social media, mainstream methods for sentiment analysis have shifted from unimodal to multimodal approaches. However, the challenges of extracting high-quality emotional features and achieving effective interaction between different modalities remain two major obstacles in multimodal sentiment analysis. To address these challenges, this paper proposes a Text-Gated Interaction Network with Inter-Sample Commonality Perception (TGICP). Specifically, we utilize a Inter-sample Commonality Perception (ICP) module to extract common features from similar samples within the same modality, and use these common features to enhance the original features of each modality, thereby obtaining a richer and more complete multimodal sentiment representation. Subsequently, in the cross-modal interaction stage, we design a Text-Gated Interaction (TGI) module, which is text-driven. By calculating the mutual information difference between the text modality and nonverbal modalities, the TGI module dynamically adjusts the influence of emotional information from the text modality on nonverbal modalities. This helps to reduce modality information asymmetry while enabling full cross-modal interaction. Experimental results show that the proposed model achieves outstanding performance on both the CMU-MOSI and CMU-MOSEI baseline multimodal sentiment analysis datasets, validating its effectiveness in emotion recognition tasks.

KEYWORDS: Multi-modal sentiment analysis; multi-modal fusion; graph convolutional networks; inter-sample commonality perception; gated interaction

1 Introduction

With the rapid development of social media and multimedia content, users not only rely on text for emotional expression but also extensively use voice, images, and videos, among other modalities. This multimodal form of emotional expression presents significant challenges to traditional unimodal sentiment analysis methods, creating an urgent need for the development of new technologies to comprehensively capture and understand human emotions [1]. As a result, Multimodal Sentiment Analysis (MSA) has emerged. MSA integrates information from multiple perceptual channels, such as voice, facial expressions, gestures, and text, enabling AI systems to better understand and analyze human emotions. With MSA, AI systems can more accurately recognize, comprehend, express, and regulate emotions, providing a more intelligent and natural emotional experience in fields such as human-computer interaction, mental health, and social robotics. However, current multimodal sentiment analysis still faces several challenges, such

Copyright © 2025 The Authors. Published by Tech Science Press.
This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

as how to effectively extract high-quality emotional features and achieve efficient fusion of information from multiple modalities to fully leverage the emotional data from each modality, thereby enhancing model performance.

To obtain high-quality emotional features that are more comprehensive and rich in emotional information, a commonly used approach is to enhance the unimodal or fused multimodal representations by leveraging common features between modalities. MulT [2] uses a Transformer network for cross-sample translation to learn the commonalities between modalities, thereby refining unimodal features and effectively addressing the information disparity between modalities. MISA [3] constructs modality-invariant and modality-specific subspace representations separately, using the modality-invariant representation as a guiding signal to effectively learn the common features and individual differences between multimodal data. TCSP [4] and SWRM [5] design text-centered private-shared information extraction networks, utilizing information from the text modality to improve multimodal feature representation. Although these methods enhance modality features through invariant emotional feature commonality between modalities, they generally overlook the utilization of emotional commonalities within similar samples in each modality. This results in a certain degree of missing emotional information within modalities, thereby affecting the model's ability to capture common features within each modality.

In terms of multimodal fusion, existing multimodal sentiment analysis frameworks can generally be divided into two types based on the operational strategy: early fusion and late fusion [6]. In early fusion, methods such as TFN [7], MARN [8], and LMF [9] integrate the input features from various modalities through concatenation, summation, or similar techniques to obtain multimodal features. However, due to the differing parameter spaces of each modality's data, leading to modality information asymmetry, early fusion may limit the model's ability to learn modality-specific unique information. In late fusion, models like Self-MM [10], MKL [11], and MFN [12] process unimodal data independently and then concatenate the outputs at the final layer to form multimodal features for prediction. Although late fusion strategies avoid the issue of inconsistencies in the multimodal parameter space, their inherent drawback lies in the difficulty of capturing underlying interaction features between modalities, and they may weaken or overlook cross-modal semantic correlations. Therefore, there is an urgent need for a fusion approach that takes into account both the parameter space differences between modalities and the semantic relationships between them to some extent.

We propose a Text-Gated Interaction Network with Inter-Sample Commonality Perception (TGICP). The TGICP model consists of two main components: the Inter-sample Commonality Perception (ICP) module and the Text-Gated Interaction (TGI) module. In the ICP module, we enhance the original sample features by extracting common emotional features from similar samples within the same modality, thereby improving the expressive power of modality-specific features. The enhanced sample features are then input into the TGI module, which dynamically adjusts the cross-modal interaction process by calculating the information disparity between the text modality and nonverbal modalities. This mechanism dynamically perceives the emotional correlation levels between modalities and adaptively adjusts the multimodal interaction process, thereby effectively alleviating the issue of cross-modal information imbalance. Ultimately, the model fuses the multimodal representations and performs sentiment prediction, significantly improving the accuracy and robustness of sentiment analysis.

To evaluate the performance advantages of the TGICP model, we conducted experimental validation using two multimodal sentiment analysis benchmark datasets: CMU-MOSI and CMU-MOSEI. Compared to existing state-of-the-art (SOTA) methods, TGICP demonstrates significant performance improvements on both datasets, validating its technical innovation and effectiveness.

The main contributions of this paper can be summarized as follows:

- We propose a Text-Gated Interaction Network with Inter-Sample Commonality Perception (TGICP) for multimodal sentiment analysis. This model enhances feature representation by extracting common information from intra-modality similar samples and utilizes a text-guided gating mechanism to dynamically regulate cross-modal feature fusion.
- We designed a Inter-sample Commonality Perception (ICP) module. This module extracts the emotional commonalities between similar samples within each modality and uses this common information to supplement and enhance the original samples, thereby enriching their unimodal emotional feature space and helping the model learn more comprehensive and richer unimodal emotional features.
- We introduce the Text Gated Interaction (TGI) module, which quantifies the information disparity between textual and non-textual modalities and dynamically adjusts the cross-modal interaction process. This mechanism enables the text modality to adaptively provide emotional context, effectively enhancing the emotional expression capability of non-textual modalities and significantly alleviating the issue of information asymmetry between modalities, thereby improving the model's predictive performance.
- We conducted comprehensive experiments on the CMU-MOSI and CMU-MOSEI datasets to evaluate the proposed model. The experimental results demonstrate that our model performs exceptionally well on the majority of evaluation metrics, thoroughly validating its effectiveness and superiority.

2 Related Work

2.1 Multimodal Sentiment Analysis

The core objective of multimodal sentiment analysis is to integrate multiple modalities of information (such as text, visual, and auditory data) from the same data sample and deeply explore and assess the emotional components it contains. The challenge in this process lies in the fact that different modalities perceive and interpret the same event from their unique perspectives. For example, text can provide emotional expression through words, the visual modality may convey emotion through facial expressions or body language, while the auditory modality reflects emotional changes through features like tone, volume, and speech rate. Therefore, the key challenge is how to effectively extract useful features from different modalities, enable information interaction and collaboration between modalities, and ultimately fuse these heterogeneous features effectively. This has been a long-standing research focus in the field of multimodal sentiment analysis.

In early research, Zadeh et al. [7] innovatively proposed the Tensor Fusion Network (TFN), which generates feature representations containing unimodal, bimodal, and trimodal combinations through Cartesian product operations. To further enhance the feature extraction capability, they also designed an architecture that combines the Long Short-Term Memory (LSTM) structure with a memory fusion mechanism, aimed at constructing comprehensive intra-modal and cross-modal feature views. Majumder et al. [13] used GRU to model two modalities layer by layer, enabling sufficient interaction between the two modalities and extracting information. With the development of the Transformer [14] model proposed by Google in various fields, methods using Transformers to process multimodal information have also emerged in the field of multimodal sentiment analysis. Lv et al. [15] proposed a Progressive Modality Reinforcement (PMR) method, which introduces an information center and progressive reinforcement strategy to address the issue of asynchronous data fusion in multimodal sentiment recognition. They also use cross-modal attention to reinforce features from different modalities and gradually enhance information flow. Rahman et al. [16] enhanced the encoding capabilities for different modalities by adding their designed Multimodal Adaptive Gate (MAG) to Bidirectional Encoder Representations from Transformers (BERT) and XLNET. Han et al. [17] effectively reduced information loss in key task processes by combining mutual information maximization theory with modality interaction theory, significantly improving model performance. Hou et al. [18] proposed a multimodal sentiment analysis method named TF-BERT, which gradually enhances information complementarity between different modalities by introducing Tensor-based Cross-modal Fusion (TCF) modules and Tensor-based Cross-modal Transformer (TCT) modules, enabling the simultaneous fusion of three modalities.

In previous multimodal sentiment analysis models, researchers typically strengthened the joint representation of each modality or multimodal features by extracting common characteristics between different modalities, which, to some extent, supplemented the emotional space of the samples and enhanced the sample features. However, these models generally overlook a key issue: the features of each modality are not always complete. Relying on these incomplete features for sentiment prediction may cause the model to overlook subtle emotional signals within modalities, leading to biased prediction results. Therefore, in our research, we enhance and supplement the original features by learning the commonality information between similar samples within each modality, thereby optimizing the feature quality of each modality and ultimately improving the overall performance of the model.

2.2 Inter-Sample Feature Enhancement

With the growing demand for multimodal data processing, cross-sample feature enhancement techniques have rapidly developed and have now become an important research direction in the field of feature engineering. These techniques have achieved significant success in various cross-modal applications. In multimodal sentiment analysis, this method can effectively enhance the model's learning capability in unimodal or multimodal sentiment analysis tasks by extracting emotional commonalities between similar samples. Traditional multimodal sentiment analysis methods typically focus on integrating information from different modalities (e.g., text, speech, vision) to improve the accuracy of sentiment prediction. However, these methods often overlook the emotional commonalities between similar samples within the same modality, which actually provide rich latent information that helps the model better understand and capture emotional features. As a result, researchers have begun to explore modeling the relationships between samples and leveraging the similarity between samples to strengthen feature representation.

In the early stages of cross-sample feature enhancement, Chen et al. [19] proposed an innovative knowledge transfer method—cross-sample similarity, which is derived from deep metric learning models. By passing the similarity information between different samples as "dark knowledge" to smaller student models, this method enhances their performance while reducing computational complexity. With the widespread application of multimodal data, more and more researchers began to introduce this method into the processing of multimodal data. Huang and Lin [20] focused on modeling the relationships between samples, considering the interrelationships among multiple samples, and designed the AttnMixup strategy to perform weighted mixing of several samples, thereby enhancing data features and improving the model's generalization ability. Li et al. [21] proposed a strategy combining within-sample and between-sample contrastive learning, aiming to model the relationships between samples, enabling the model to effectively capture similarities between similar samples and differences between dissimilar samples. Recently, researchers have also started applying cross-sample feature enhancement methods to multimodal sentiment analysis. Zhang et al. [22] designed a dynamic sample selection strategy based on neighborhood density, selecting nearby samples with higher density to identify representative and distinguishable samples, thus effectively focusing on high-quality samples in the cross-sample feature enhancement process. Huang et al. [23] proposed a new cross-sample fusion strategy by merging the modality information of different samples and incorporating adversarial training, pairwise prediction tasks, and two-stage text-centered contrastive learning, which effectively enhances the retention and fusion of specific modality information.

In previous cross-sample feature enhancement methods, the extraction of common features between samples primarily relied on autoencoders to decompose the feature space. However, since autoencoders are optimized based on global loss functions to a certain extent, they focus on the global representation of features and overlook the potential complex and nonlinear local relationships between samples (such as the similarity between neighboring samples). Although autoencoders learn complex mappings by stacking network layers, they still rely on a fixed network structure for data mapping and cannot dynamically adjust weights and topological structures based on graph structures, as graph convolutional networks can, to adapt to different data distributions. Moreover, in previous works, feature enhancement using common features between samples typically employed concatenation or weighted summation, methods that often introduce redundant information, leading to a more complicated feature space. To better leverage the common information between similar samples for feature enhancement, we propose a method based on mutual information constraints using graph neural networks to learn the common information between similar samples. Additionally, we design a Commonality Embedding Optimization Module to achieve multidimensional enhancement of the original features.

2.3 Text-Center Based Methods

In Multimodal Sentiment Analysis (MSA), each modality provides a different emotional interpretation of the same event or behavior through its unique perceptual approach. This multi-perspective analysis enables a more comprehensive and in-depth capture and understanding of the diversity of emotions. Compared to visual and auditory modalities, the text modality typically responds most directly and instinctively to emotional information. This is because language, as a carrier of emotional expression, carries rich emotional content in everyday communication. Humans can precisely convey their emotional world through the combination of vocabulary, sentence structure, and context. Therefore, the text modality not only provides high accuracy in expressing emotions but also reflects subtle emotional changes in greater detail. This has led many researchers to view the text modality as the foundational modality for multimodal sentiment analysis, often relying first on textual data for emotion recognition, and then using data from other modalities for supplementary validation and enhancement.

Sun et al. [24] achieved efficient feature learning by constraining the outer product matrix of text and optimizing with canonical correlation analysis loss. Mai et al. [25] focused on improving the LSTM structure by designing a special gating mechanism to enhance text representation and perform information correction, while also applying auxiliary methods for text information calibration. Meanwhile, with the widespread application of attention-based methods in Natural Language Processing (NLP), Computer Vision (CV), and multimodal domains, new approaches and more advanced text encoders, such as BERT [26] and RoBERTa [27], have been introduced to the field of multimodal sentiment analysis. With the support of these stronger text encoders, researchers are increasingly using the text modality as the core modality for multimodal sentiment analysis. Wu et al. [4] studied the shared and private semantics in nonverbal modalities through a cross-modal prediction task. Wang et al. [28] injected information from the text modality into nonverbal modalities using a text-guided attention mechanism and a cross-modal Transformer, enhancing their representational power. Huang et al. [29] associated three-modal data using a text-centered unimodal and multimodal cross-modal Transformer. Li et al. [30] designed a semantic enhancement framework based on a cross-modal Transformer, which effectively strengthens the emotional representation capability of the text modality by coordinating the interaction mechanisms between the textual and nonverbal modalities.

Based on the previous work, it is clear that current text-centered approaches predominantly rely on cross-modal Transformers for cross-modal translation. However, in multimodal sentiment analysis, the emotional expressions in the text and nonverbal modalities may be asymmetric. Some emotional information may be more prominent in the text, while others are more prominent in nonverbal modalities. Direct cross-modal translation might miss out on the potential emotional information within each modality. Therefore, in order to achieve a more balanced interaction between modality information, we first quantify the information correlation between the text modality and nonverbal modality before performing cross-modal translation. Based on different levels of mutual information, we selectively incorporate text information into nonverbal information, compensating for the information asymmetry and improving the accuracy of sentiment analysis.

3 Methodology

In this section, we will provide a detailed introduction to the Text-Gated Interaction Network with Inter-Sample Commonality Perception (TGICP) proposed in this paper. The core objective of TGICP is to extract the common emotional features between each sample and its similar samples using a graph convolutional network, and enhance the original sample features with these common features, thereby enriching the emotional feature space of the sample across different modalities. At the same time, during the cross-modal interaction and fusion process, TGICP uses the strong explicit emotional information in the text modality as a gating mechanism to dynamically regulate the information interaction between modalities, thereby improving the model's efficiency in utilizing information from each modality.

3.1 Task Setup

Multimodal Sentiment Analysis (MSA) aims to predict the emotional tendency and intensity of the speaker in a video by simultaneously utilizing information from multiple modalities. In existing MSA research, this task can be modeled as either a classification problem or a regression problem. In our work, we choose to model and analyze it as a regression task. MSA typically involves three modalities: text (t), visual (v), and audio (a). We denote the input raw data as I_m, where $m \in \{t, v, a\}$, representing the three different modalities. Our goal is to design a robust model that can effectively integrate the information from each modality to accurately predict the emotional intensity score $y \in \mathbb{R}^n$ for the input sample.

3.2 Overall Architecture

As shown in Fig. 1, TGICP mainly consists of four components: feature extraction, Inter-sample Commonality Perception, Text-Gated Interaction (TGI), and sentiment prediction. The feature extraction component transforms the raw multimodal data into vector representations. The Inter-sample Commonality Perception (ICP) part extracts the common emotional features between the sample and its emotionally similar samples, injecting these common features back into the original sample to enrich the emotional feature space of the sample across different modalities. The Text-Gated Interaction (TGI) component dynamically controls the cross-modal interaction of the sample's multimodal representation by calculating the information disparity between the text and nonverbal modalities. Finally, the sentiment prediction component analyzes and predicts the fused sample features, providing the final emotional intensity score. In this section, we will provide a detailed introduction to these components.

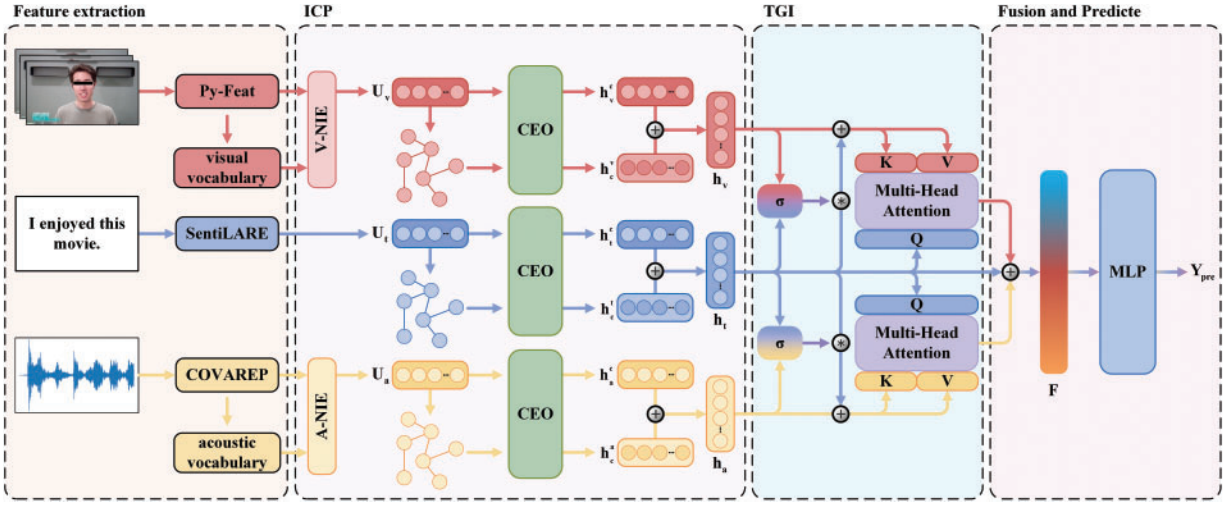

Figure 1: The overall structure of TGICP primarily consists of four components: feature extraction, Inter-sample Commonality Perception (ICP), Text-Gated Interaction (TGI), and fusion and sentiment prediction

3.3 Feature Extraction

For the text modality, we use the pre-trained language model SentiLARE [31] to obtain text features. SentiLARE is a novel pre-trained language model that, compared to the previous BERT [26] model, enhances text feature representation by incorporating word-level linguistic knowledge through the addition of part-of-speech embeddings and sentence-level word polarity embeddings. The part-of-speech embeddings are derived using the Stanford Log-Linear Part-of-Speech Tagger [32], while the sentiment polarity labels come from SentiWordNet [33]. After passing the text through SentiLARE, we obtain the word embeddings of all the words in the text, denoted as $\mathbf{H}_t \in \mathbb{R}^{n \times l_t \times d_t}$. Since the [CLS] token effectively captures the overall semantic features of a sentence and performs well in various tasks, we choose the embedding feature of the [CLS] token from each sentence, denoted as $\mathbf{F}_t \in \mathbb{R}^{n \times d_t}$, as the sentence-level high-level feature representation.

$$\mathbf{H}_t = SentiLARE\left(I_t, \theta_t^{SL}\right) \in \mathbb{R}^{n \times l_t \times d_t} \tag{1}$$

where n is the number of samples, l_t is the length of the text sequence, and d_t represents the dimensionality of the word embeddings, θ_t^{SL} represents the learnable parameters in SentiLARE.

For the visual and auditory modalities, we use the pre-trained open-source tools Py-Feat [34] and COVAREP [35] to extract features $\mathbf{X}_v \in \mathbb{R}^{n \times l_v \times d_v}$ and $\mathbf{X}_a \in \mathbb{R}^{n \times l_a \times d_a}$ from the raw data. In the work of Wang et al. [36], they constructed nonverbal dictionaries for the visual and auditory modalities using unsupervised clustering on the training set data, transforming real nonverbal modality data into indexes. Based on their work, we adopt the nonverbal dictionaries they constructed to obtain the clustering indexes for the visual and auditory modalities, denoted as $\mathbf{C}_v \in \mathbb{R}^{n \times l_v}$ and $\mathbf{C}_a \in \mathbb{R}^{n \times l_a}$. Here, n represents the number of samples, l_m and d_m represent the sequence length and feature dimension of modality m, where $m \in \{t, v, a\}$ denotes the three different modalities.

3.4 Inter-Sample Commonality Perception

To more fully utilize the emotional commonalities between samples and enrich their unimodal emotional feature space, we designed a Inter-sample Commonality Perception (ICP) module. This module consists of three components: Nonverbal Modality Information Enhancement (NIE), Inter-sample Commonality Mining (ICM), and Commonality Embedding Optimization (CEO). The following section provides a detailed description of each of these components.

3.4.1 Nonverbal Modality Information Enhancement

In the feature extraction section, we obtained the initial feature representations $\mathbf{X}_{\tilde{m}}$ for the two nonverbal modalities, visual and auditory, as well as their clustering indices $\mathbf{C}_{\tilde{m}}$, where $\tilde{m} \in \{v, a\}$. The clustering index $\mathbf{C}_{\tilde{m}}$ reflects the similarity between video frames to some extent. Therefore, in order to better extract the shared information between the nonverbal modalities, we designed a Nonverbal Modality Information Enhancement (NIE) module, as shown in Fig. 2. First, we embed the clustering index $\mathbf{C}_{\tilde{m}}$, resulting in the enhanced clustering index features $\mathbf{C}^{e}_{\tilde{m}}$ for the nonverbal modalities.

$$\mathbf{C}^{e}_{\tilde{m}} = Embedding\left(\mathbf{C}_{\tilde{m}}, \theta^{embedding}_{\tilde{m}}\right) \in \mathbb{R}^{n \times l_{\tilde{m}} \times d_{\tilde{m}}} \tag{2}$$

After obtaining the index features $\mathbf{C}^{e}_{\tilde{m}}$, we input them along with the initial features $\mathbf{X}_{\tilde{m}}$ of the nonverbal modalities into the NIE for interaction between the initial features and the index features. We first apply a multi-head attention mechanism to each feature type, allowing them to mutually guide the learning of the feature information between the two representations. Specifically, as shown in Eq. (3), we compute the query, key, and value for both feature types. Then, as described in Eq. (4), we set the index feature $Q^{\tilde{m}}_{C^{e}_{\tilde{m}}}$ as the query, while the modal initial features $K^{\tilde{m}}_{X_{\tilde{m}}}$ and $V^{\tilde{m}}_{X_{\tilde{m}}}$ serve as the key and value for the multi-head attention mechanism.

$$\begin{aligned} Q^{\tilde{m}}_{E} &= Linear(E) = W^{\tilde{m}}_{q} E \\ K^{\tilde{m}}_{E} &= Linear(E) = W^{\tilde{m}}_{k} E \\ V^{\tilde{m}}_{E} &= Linear(E) = W^{\tilde{m}}_{v} E \end{aligned} \tag{3}$$

$$C^{e}_{\tilde{m}}{}' = Softmax\left(\frac{Q^{\tilde{m}}_{C^{e}_{\tilde{m}}} \cdot K^{\tilde{m}}_{X_{\tilde{m}}}{}^{\mathrm{T}}}{\sqrt{d_{\tilde{m}}}}\right) V^{\tilde{m}}_{X_{\tilde{m}}} \tag{4}$$

where $E \in \{\mathbf{C}^{e}_{\tilde{m}}, \mathbf{X}_{\tilde{m}}\}$ represents the two feature types, $Q^{\tilde{m}}_{E}$, $K^{\tilde{m}}_{E}$, and $V^{\tilde{m}}_{E}$ denote the query, key, and value derived from the feature E, respectively. $W^{\tilde{m}}_{q}$, $W^{\tilde{m}}_{k}$ and $W^{\tilde{m}}_{v}$ are learnable parameters. $\mathbf{C}^{e}_{\tilde{m}}{}'$ represents the new index features obtained after the interaction guided by the index features. For the case where the modal initial features serve as the guide in the multi-head attention, a similar approach is employed for the computation. After obtaining the features resulting from the mutual guidance and interaction between the two feature types, these features are concatenated and passed through self-attention, which generates the further refined features $\mathbf{H}_{\tilde{m}}$ for the two nonverbal modalities.

$$\mathbf{H}_{\tilde{m}} = Self - Attention(\mathbf{X}_{\tilde{m}} \parallel \mathbf{C}^{e}_{\tilde{m}}\prime) \tag{5}$$

Additionally, to capture the contextual temporal features of these two modalities, we use two independent Bidirectional Long Short-Term Memory networks (BiLSTM) to further process the extracted nonverbal modality features $\mathbf{H}_{\tilde{m}}$. As shown in Eqs. (6) and (7), the entire sequence is passed through the BiLSTM, resulting in the high-level feature representations for the visual and auditory modalities, $\mathbf{F}_{v}$ and $\mathbf{F}_{a}$, respectively.

$$\mathbf{F}_{v} = BiLSTM\left(\mathbf{H}_{v}, \theta^{BiLSTM}_{v}\right) \in \mathbb{R}^{n \times d_{v}} \tag{6}$$

$$\mathbf{F}_a = BiLSTM\left(\mathbf{H}_a, \theta_a^{BiLSTM}\right) \in \mathbb{R}^{n \times d_a} \tag{7}$$

where θ_v^{BiLSTM} and θ_a^{BiLSTM} represent the learnable parameters in the BiLSTM for the visual and auditory modalities, respectively.

Finally, as shown in Eq. (8), the high-level feature representation of each modality is mapped to a shared space through independent encoders $Encoder_m$ and transformed into the input $\mathbf{U}_m$ for downstream tasks.

$$\mathbf{U}_m = Encoder_m\left(\mathbf{F}_m, \theta_m^{Encoder_m}\right) \in \mathbb{R}^{n \times d_{hidden}} \tag{8}$$

where $m \in \{t, v, a\}$ represents the three modalities, and $\theta_m^{Encoder_m}$ denotes the learnable parameters in the encoder corresponding to each modality.

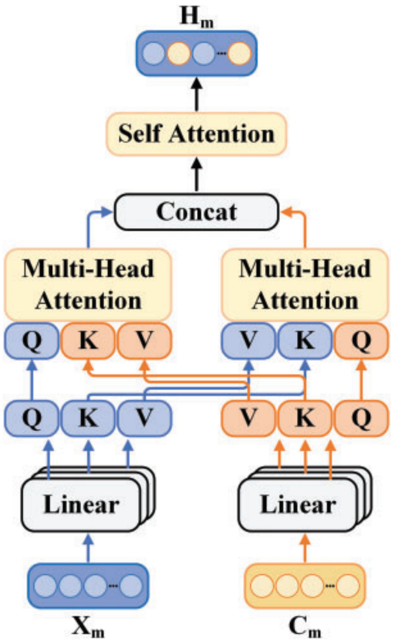

Figure 2: Schematic of nonverbal modality information enhancement (NIE)

3.4.2 Inter-Sample Commonality Mining

In the work of Mai et al. [37], the problem of multimodal sequence analysis was transformed into a graph learning problem, where Graph Convolutional Networks (GCN) were used to process the feature data of each modality. In this approach, an adjacency matrix is constructed within the graph structure by calculating the similarity between samples, and node features are propagated through graph convolution, thereby capturing both intra-modal and cross-modal feature dependencies. Specifically, Mai et al. proposed several methods to define the adjacency matrix, including non-parametric approaches (such as the generalized diagonal matrix) and learnable methods (where the adjacency matrix is learned through gradient descent). These methods, through effective feature propagation, capture the commonalities between intra-modal samples as well as the complex relationships between modalities. Inspired by this, to adaptively extract the emotional commonalities between similar samples based on their structure, we employ a similar graph construction method to design a weighted undirected graph $\mathcal{G} = (V, \mathcal{E})$ for each modality. This graph is based on sample-level attention and feature similarity to extract common features between similar samples, as shown in Fig. 3.

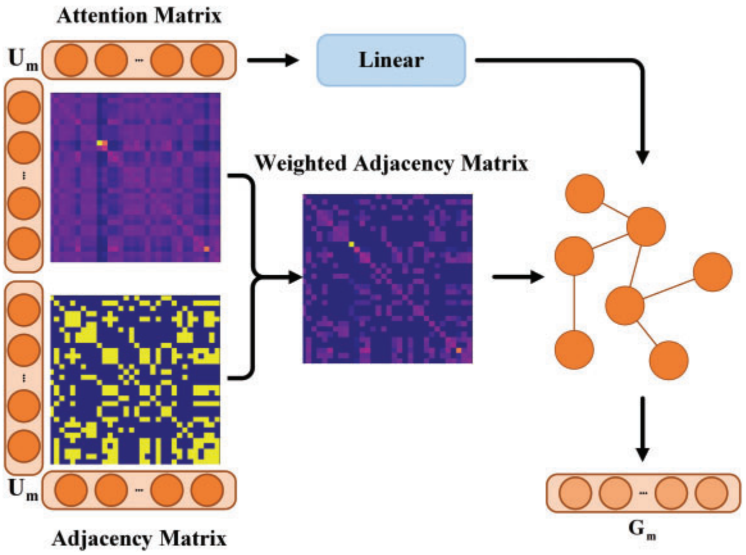

Figure 3: Schematic of inter-sample commonality mining (ICM)

Graph Construction: As shown in Eq. (9), we first calculate the cosine similarity between pairs of samples for each modality to obtain the feature similarity matrix $\mathbf{M}_{FCS}^{m}$. Then, as shown in Eq. (10), we set the values greater than or equal to the threshold τ_s to 1, and those smaller than the threshold to 0, resulting in the adjacency matrix $\mathbf{M}_{FC}^{m}$.

$$ {\mathbf{M}_{FCS}^{m}}^{ij} = \frac{\mathbf{U}_m^i \mathbf{U}_m^{j^{\mathrm{T}}}}{\|\mathbf{U}_m^i\|_2 \|\mathbf{U}_m^j\|_2} \tag{9} $$

$$ {\mathbf{M}_{FC}^{m}}^{ij} = \begin{cases} 0 & \mathbf{M}_{FCS}^{ij} < \tau_s \\ 1 & \mathbf{M}_{FCS}^{ij} \geq \tau_s \end{cases} \tag{10} $$

where $i, j \in \{1, 2, 3, \ldots, n\}$ represents two distinct samples.

In order to allow the model to dynamically adjust the weights for extracting commonalities between different samples, we compute the attention weight matrix $\mathbf{M}_{Att}^{m}$ for each modality, as shown in Eq. (11), to assign different weights to the extraction of common features between similar samples.

$$ \mathbf{M}_{Att}^{m} = \frac{\mathbf{U}_m^i \mathbf{U}_m^{j^{\mathrm{T}}}}{\sqrt{d_{\text{hidden}}}} \quad m \in \{t, v, a\} \tag{11} $$

where $\mathbf{U}_m^i$, $\mathbf{U}_m^i$ represent the features of any two samples under modality m, d_{hidden} denotes the dimensionality of the sample features, and $\mathbf{M}_{Att}^{m} \in \mathbb{R}^{n \times n}$.

After obtaining the adjacency matrix $\mathbf{M}_{FC}^{m}$ and the attention weight matrix $\mathbf{M}_{Att}^{m}$ for each modality, we calculate the weighted adjacency matrix $\mathbf{A}_m$ of the commonality extraction graph for each modality using element-wise multiplication, as shown in Eq. (12), to represent the similarity relationships between different samples.

$$\mathbf{A}_m = \mathbf{M}_{Att}^m * \mathbf{M}_{FC} \quad \mathbf{A}_m \in \mathbb{R}^{n\times n} \tag{12}$$

Graph Convolution: After the graph construction described above, we obtain the weighted adjacency matrix $\mathbf{A}_m$ for the commonality extraction graph of each modality. We then use $\mathbf{U}_m$ as input and build a graph convolutional neural network to extract the emotional commonalities $\mathbf{C}_m$ between emotionally similar samples in modality m. Specifically, given a weighted undirected graph $\mathcal{G} = (V, \mathcal{E})$, we compute the symmetrically normalized Laplacian matrix $\mathbf{P}$ of the graph $\mathcal{G}$ according to the method in Eq. (13).

$$\mathbf{P} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}_m\mathbf{D}^{-\frac{1}{2}} \tag{13}$$

where $\mathbf{D}$ represents the degree matrix of the graph $\mathcal{G}$. Since the self-loops of each node are already included when calculating the unweighted adjacency matrix $\mathbf{M}_{FC}$ of the graph $\mathcal{G}$, no additional self-loops are added to the nodes when computing the Laplacian matrix $\mathbf{P}$.

Subsequently, we use $\mathbf{U}_m$ as input and perform graph convolution according to Eqs. (14) and (15) to extract the emotional commonalities $\mathbf{G}_m$ between similar samples in modality m.

$$\begin{aligned}\mathbf{U}'_m &= Linear(\mathbf{U}_m, \theta_m^{\mathcal{G}})\\ &= \left(\mathbf{W}_m^{\mathcal{G}}\mathbf{U}_m + b_m^{\mathcal{G}}\right)\end{aligned} \tag{14}$$

$$\begin{aligned}\mathbf{G}_m &= \mathbf{P}\mathbf{U}'_m\\ &= \left(\mathbf{D}^{-\frac{1}{2}}\mathbf{A}_m\mathbf{D}^{-\frac{1}{2}}\right)\mathbf{U}'_m\end{aligned} \tag{15}$$

where $\mathbf{W}_m^{\mathcal{G}} \in \mathbb{R}^{n\times n}$ and $b_m^{\mathcal{G}} \in \mathbb{R}^n$ are the learnable parameters of the linear layer $Linear$, and $\mathbf{G}_m \in \mathbb{R}^{n\times d_{hidden}}$.

3.4.3 Commonality Embedding Optimization

Through the above steps, we have obtained the inter-sample commonality features $\mathbf{G}_m$ and high-level features $\mathbf{U}_m$ for each modality. To enable the high-level features of each modality to better represent emotional features, we design a Dual-View Commonality Embedding Optimization Layer (CEO), which embeds the inter-sample commonality features $\mathbf{G}_m$ into the high-level features $\mathbf{U}_m$ and performs optimization.

As shown in Fig. 4, the designed CEO consists of two components: the Local Mixer and the Global Mixer. Each component is composed of two linear layers and a Gaussian Error Linear Unit (GELU) activation function, with residual connections incorporated. As described in Eq. (16), we first stack the high-level feature $\mathbf{U}_m$ and the cross-sample commonality feature $\mathbf{G}_m$ along the first dimension to obtain $\mathbf{S}_m$.

$$\mathbf{S}_m = Stack(\mathbf{U}_m, \mathbf{G}_m) \in \mathbb{R}^{n\times 2\times d_{hidden}} \tag{16}$$

Subsequently, as shown in Eq. (17), the stacked $\mathbf{S}_m$ is fed into the Local Mixer, allowing the interaction and communication between the two features, $\mathbf{U}_m$ and $\mathbf{G}_m$, at the local level.

$$\begin{aligned}\mathbf{S}'_m &= LocalMixer(\mathbf{S}_m)\\ &= \mathbf{S}_m^{\mathrm{T}} + \mathbf{W}_2\Phi\left(\mathbf{W}_1 Norm(\mathbf{S}_m^{\mathrm{T}})\right) \in \mathbb{R}^{2\times n\times d_{hidden}}\end{aligned} \tag{17}$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are learnable parameters, Φ denotes the GELU activation function, and *Norm* represents layer normalization.

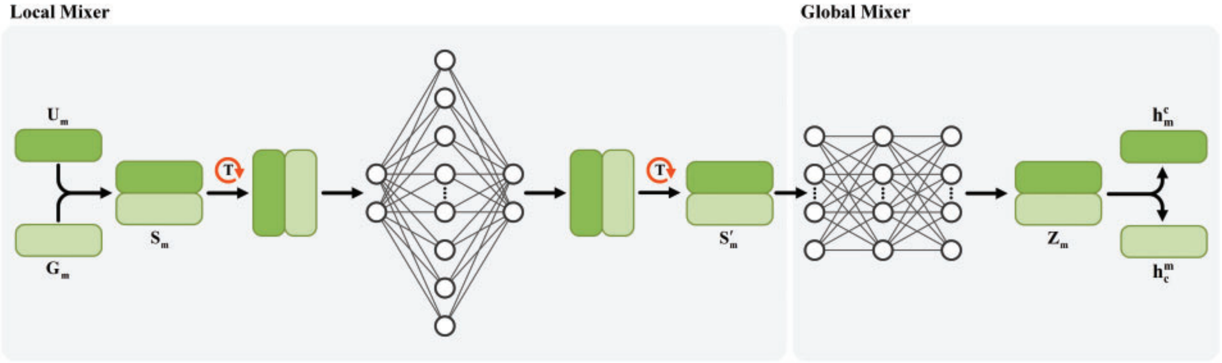

Figure 4: Schematic of commonality embedding optimization (CEO)

After the local-level interaction and communication, as shown in Eq. (18), the locally interacted feature vector $\mathbf{S}'_m$ obtained from the Local Mixer is fed into the Global Mixer. This enables the interaction of the two locally interacted features at the global level, resulting in the feature vector $\mathbf{Z}_m$, which is the optimized common embedding of the dual perspectives.

$$\begin{aligned}\mathbf{Z}_m &= GlobalMixer(\mathbf{S}'_m)\\ &= \mathbf{S}'^{\mathrm{T}}_m + \mathbf{W}_4\Phi\left(\mathbf{W}_3 Norm(\mathbf{S}'^{\mathrm{T}}_m)\right)\end{aligned} \tag{18}$$

where $\mathbf{W}_3$ and $\mathbf{W}_4$ are learnable parameters. Finally, as shown in Eq. (19), the dual-perspective interaction vector $\mathbf{Z}_m$ is unstacked along the first dimension to obtain the feature vectors of its two channels. These feature vectors are then added together to obtain the feature h_m, which represents the optimized common embedding for each modality. After two rounds of interaction, both locally and globally, h_m not only retains the unique characteristics of the sample to some extent but also incorporates the emotional commonality with similar samples. Compared to the original sample features, h_m offers a richer representation in the emotional space.

$$h_m = h^m_c + h^c_m \in \mathbb{R}^{n\times d_{hidden}} \tag{19}$$

where $h^m_c, h^c_m \in \mathbb{R}^{n\times d_{hidden}}$ are the two channels of $\mathbf{Z}_m$. According to the Information Bottleneck theory proposed by Tishby and Zaslavsky [38], in supervised learning tasks, it is possible to compress the input data in a way that retains as much information as possible about the output (label) variable. This compression-prediction trade-off is achieved by minimizing mutual information. The goal is to preserve as much information about the output as possible while removing irrelevant information for prediction, thus finding the most concise feature representation while maintaining the correlation between the input and output. Building on this theoretical foundation and inspired by Lambert et al. [39], we compute the mutual information loss $Loss_{mi}$ between the input features $\mathbf{U}_m$ and the intermediate features h_m as shown in Eq. (20) and minimize it to reduce their dependency.

$$Loss_{mi} = exp\left(\frac{h_m}{\|h_m\|^2}\left(\frac{Linear(\mathbf{U}_m)}{\|Linear(\mathbf{U}_m)\|^2}\right)^{\mathrm{T}}\right) \tag{20}$$

3.5 *Text Gated Interaction*

Compared to nonverbal modalities, the text modality is better suited to directly express rich emotional information, with its semantic content being more precise and detailed. In contrast, the emotional information conveyed by nonverbal modalities tends to be more abstract and implicit (e.g., micro-expressions, tone variations), and although the relationship between these signals and text is more indirect, they remain important signals in emotional analysis. However, in existing cross-modal interactions based on relevance, these potentially crucial emotional signals are often overlooked or lost. To address this issue and compensate for the emotional information loss caused by modality information asymmetry during cross-modal interactions, we have designed a text-gated interaction module, as shown in Fig. 5. This module aims to leverage the accurate emotional information in text to dynamically compensate for the information asymmetry between modalities, thereby improving the accuracy and comprehensiveness of emotional analysis.

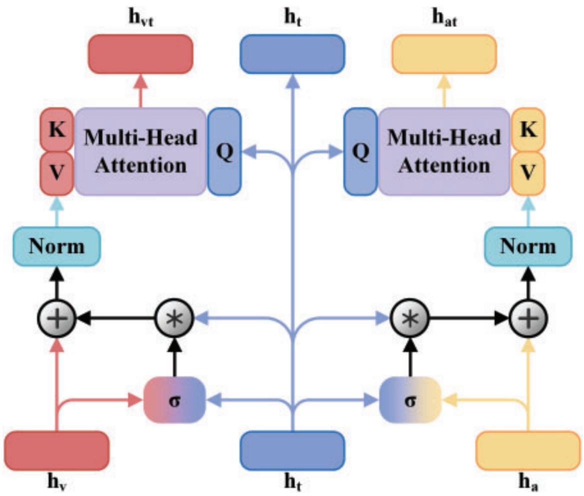

Figure 5: Schematic of text-gated interaction (TGI)

To provide a concise and intuitive introduction to this part of the method, we use the text-visual modality interaction as an example. To effectively integrate the emotional information from both text and video modalities, it is essential to assess the correlation and complementarity between the two modalities, thereby guiding the feature fusion process. In this context, mutual information is employed to quantify and optimize the emotional information sharing between the modalities, helping to avoid redundancy and address the informational discrepancies between them. First, we calculate the level of mutual information between the two modalities using the CPC method [40], as shown in Eq. (21):

$$\mathbf{M}_{tv} = exp\left(\frac{h_v^{\mathrm{T}}}{\|h_v^{\mathrm{T}}\|^2}\left(\frac{Linear(h_t^{\mathrm{T}})}{\|Linear(h_t^{\mathrm{T}})\|^2}\right)^{\mathrm{T}}\right) \tag{21}$$

When the information correlation level between two modalities is low, it indicates that the degree of information dependence between the features of the two modalities is also low, with a smaller intersection of emotional information. In this case, we assign a higher weight to the textual modality, enabling it to provide more accurate emotional information to the visual modality, thus helping the visual modality better understand and express emotions. Conversely, if the information correlation level is high, the weight of the textual modality is reduced to avoid the convergence of multimodal data due to the excessive involvement of the textual modality. Therefore, after calculating the mutual information level between the two modalities, we first compute the information complementarity weight g_{tv} between the textual and visual modalities according to Eq. (22). Subsequently, we implement dynamic and controllable gated interaction between the textual and visual modalities by incorporating the information complementarity weight, as described in Eq. (23).

$$g_{tv} = 1 - Sigmoid(\mathbf{M}_{tv}) \tag{22}$$

$$h'_v = norm(h_v + h_t \cdot g_{tv}) \tag{23}$$

After the gated interaction, to further promote the integration and understanding of cross-modal emotional information, as shown in Fig. 6, we employ a cross-modal translation approach to achieve deep interaction between modalities. First, we treat the text as the query in the cross-modal translation, with the nonverbal modality serving as the key and value, and perform the computation as per Eq. (24). Subsequently, based on the modality pair, the query, key, and value required for the cross-modal translation are calculated using Eq. (25), thereby mapping the information from the nonverbal modality into the text modality, resulting in the feature $h_{\tilde{m}t}$ after the interaction of the nonverbal modality.

$$Q_t = Linear(h_t) = \mathbf{W}_q^t h_t \tag{24}$$

$$K_{\widetilde{m}\to t} = Linear(h'_{\widetilde{m}}) = \mathbf{W}_k^{\widetilde{m}\to t} h'_{\widetilde{m}}$$

$$V_{\widetilde{m}\to t} = Linear(h'_{\widetilde{m}}) = \mathbf{W}_v^{\widetilde{m}\to t} h'_{\widetilde{m}}$$

$$h_{\widetilde{m}t} = Softmax\left(\frac{Q_t \cdot K_{\widetilde{m}\to t}^{\mathrm{T}}}{\sqrt{d_t}}\right) V_{\widetilde{m}\to t} \tag{25}$$

where $\tilde{m} \in \{v, a\}$ denotes the nonverbal modality.

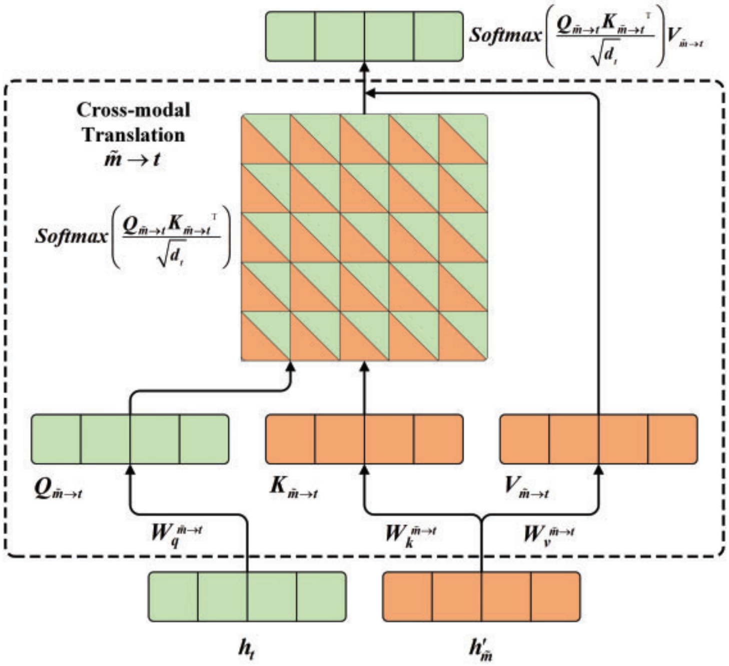
$Softmax\left(\frac{Q_{\tilde{m}\to t}K_{\tilde{m}\to t}^{\mathrm{T}}}{\sqrt{d_t}}\right)V_{\tilde{m}\to t}$
Cross-modal Translation
$\tilde{m} \to t$
$Softmax\left(\frac{Q_{\tilde{m}\to t}K_{\tilde{m}\to t}^{\mathrm{T}}}{\sqrt{d_t}}\right)$
$Q_{\tilde{m}\to t}$
$K_{\tilde{m}\to t}$
$V_{\tilde{m}\to t}$
$W_q^{\tilde{m}\to t}$
$W_k^{\tilde{m}\to t}$
$W_v^{\tilde{m}\to t}$
h_t
$h'_{\tilde{m}}$

Figure 6: Schematic of text-centered cross-modal translation

3.6 Fusion and Prediction

After the inter-sample interaction fusion, we concatenate the three modal features according to the method outlined in Eq. (26), resulting in a multimodal joint representation F.

$$F = \left(h_t \parallel h_{vt} \parallel h_{at}\right) \tag{26}$$

Subsequently, as shown in Eq. (27), we pass the representation through the Multilayer Perceptron (MLP) layer for sentiment prediction, obtaining the model's sentiment prediction result $\mathbf{Y}_{Pre}$.

$$\mathbf{Y}_{Pre} = \left(\mathbf{W}_{o2}\Psi\left(\mathbf{W}_{o1}F + b_{o1}\right) + b_{o2}\right) \tag{27}$$

where $\parallel$ denotes the concatenation operation, and Ψ represents the ReLU activation function. The model's prediction is a continuous value ranging from -3 to $+3$, representing the sentiment range from extremely negative to extremely positive.

3.7 Optimization Objectives

In our model, the loss function consists of two components: the task loss $Loss_{task}$ and the mutual information loss $Loss_{mi}$. Optimizing these two losses corresponds to different objectives: minimizing the task loss aims to optimize the overall model parameters, while minimizing the mutual information loss seeks to reduce the excessive dependence of intermediate features on the input features, thereby extracting higher-quality modal features.

For the mutual information loss, we use the calculation method defined in Eq. (20). For the task loss, considering the wide applicability of mean squared error (MSE) in regression tasks, its sensitivity to errors, and its good differentiability, we select MSE as the task loss function to guide model optimization. Therefore, the overall loss function of the model can be expressed in the form of Eq. (28).

$$\begin{aligned} Loss &= Loss_{task} + \omega_{mi} Loss_{mi} \\ &= MSE\left(\mathbf{Y}_{True}, \mathbf{Y}_{Pre}\right) + \omega_{mi} Loss_{mi} \end{aligned} \tag{28}$$

where ω_{mi} is the weight hyperparameter for the mutual information loss.

Finally, to present a clearer overview of the complete workflow of the TGICP model, we provide a systematic summary of its key steps in Algorithm 1.

Algorithm 1: Text-gated interaction network with inter-sample commonality perception (TGICP)

```
Input: $D = \{(I_t, I_v, I_a)\}$, search space = $\{\alpha_{range}, \omega_{mirange}, \tau_{srange}, B_{range}\}$.
Output: Prediction sentiment score $\mathbf{Y}_{Pre}$.
for learning rate $\alpha$, $\omega_{mi}$, $\tau_s$ in search space do
    Train model on training set.
    for each epoch do
        for mini-batch $\{(t_i, v_i, a_i), \mathbf{Y}_{True\,i}\}_{i=1}^{B}$ from $D$ do
            Enhancing nonverbal modalities to obtain high-level features as Eqs. (2)–(7).
            Map the features of each modality to the same feature space as Eq. (8).
            for $m$ in $\{t, v, a\}$, do
                Construct the intra-modal similarity sample adjacency matrix $\mathbf{M}_{FC}^{m}$
                as Eqs. (9) and (10).
```

(Continued)

Algorithm 1 (continued)

```
                Calculate the edge weight matrix of the adjacency matrix $\mathbf{M}^{m}_{Att}$ and obtain the weighted adjacency matrix $\mathbf{A}_m$ as Eqs. (11) and (12).
                Extract the common information $\mathbf{G}_m$ among intra-modal similar samples as Eqs. (13)–(15).
                Enhance the original sample features using the common information as Eqs. (16)–(19).
            end for
            Calculate the information disparity $\mathbf{M}_{t\tilde{m}}$ between the text and nonverbal modalities as Eq. (21).
            Enhance the nonverbal modality using the text modality based on the information disparity as Eqs. (22) and (23).
            Perform cross-modal interaction guided by the text modality as Eqs. (24) and (25).
            Perform multimodal fusion and sentiment prediction as Eqs. (26) and (27).
        end for
    end for
    Validate on validation set.
    Record validation performance (e.g., accuracy, loss).
    Select optimal hyperparameters $\alpha^*$, $\omega^*_{mi}$, $\tau^*_s$, $B^*$ with best validation performance.
end for
```

4 Experiments

We provide a detailed description of the experimental setup and conduct a comprehensive evaluation of the proposed TGICP model on several mainstream publicly available datasets, validating its performance in the multimodal sentiment analysis task. Additionally, through a series of well-designed ablation experiments, we thoroughly analyze the effectiveness of each key module in the model and assess their contributions to the overall performance.

4.1 Datasets

Our study is conducted on two important multimodal sentiment analysis benchmark datasets: CMU-MOSI [41] and CMU-MOSEI [42].

CMU-MOSI is one of the most commonly used datasets in the field of multimodal sentiment analysis, containing data from three modalities. The dataset includes 93 videos, with each video divided into up to 62 video segments. Each video segment contains a spoken sentence, and the emotional intensity ranges from [−3, +3], where −3 represents the most negative sentiment, and +3 represents the most positive sentiment.

CMU-MOSEI is an extended version of CMU-MOSI, featuring greater speaker diversity and a broader range of topics. This dataset contains 23,453 video segments, with emotional intensity labels identical to those in CMU-MOSI, ranging from [−3, +3] to represent emotional intensity. The video segments are sourced from 5000 videos, covering 1000 different speakers and 250 topics.

4.2 Experimental Settings

The proposed TGICP model is trained using the Adam optimizer. The model incorporates an early stopping mechanism, with a maximum of 4 training epochs and a patience value of 6. This means that if there is no performance improvement over 6 consecutive epochs, early stopping will be triggered. If no improvement occurs after 4 consecutive early stops, the training process will halt. Each time early stopping is triggered, the learning rate is halved, and the best model from the previous training epoch is restored to ensure the stability and effectiveness of the training process. Detailed hyperparameter settings are provided in Table 1, where τ_s represents the feature similarity threshold and ω_{mi} denotes the weight of the mutual information loss function.

Table 1: Hyperparameter settings in the experiment

Datasets	Batch size	Learning rate	τ_s	ω_{mi}	Drop out
CMU-MOSI	16	1e-5	0.8	0.8	0.1
CMU-MOSEI	64	1e-5	0.8	0.4	0.1

At the same time, to comprehensively evaluate the model's performance, we conducted a statistical analysis of the model's parameter size and computational efficiency. The total number of parameters in the model is 125.606 million, of which approximately 110 million parameters come from the pre-trained text feature extraction model SentiLARE, while the remaining parameters (excluding SentiLARE) amount to around 15 million. The test results on an RTX 3080 GPU show that the model's computational complexity is 4.867 G Floating Point Operations (FLOPs). The training time on the CMU-MOSI dataset is approximately 34 min, with a GPU memory usage of 6 GB during training.

4.3 Evaluation Metrics

To facilitate comparison with other studies, we adopted the evaluation methods used in previous research and employed the same performance metrics to assess the model's performance. Specifically, we used Mean Absolute Error (MAE) to measure the error between the predicted and true values, and the Pearson correlation coefficient (Corr) to evaluate the linear correlation between the predicted and true values. In the sentiment analysis task, we first used binary classification accuracy (Acc-2) to measure the accuracy of coarse-grained sentiment analysis, followed by seven-class accuracy (Acc-7) to evaluate the accuracy of fine-grained sentiment analysis. Additionally, we employed the weighted F1 score (F1-Score) to provide a comprehensive assessment of the model's precision and recall, offering a holistic view of the model's performance.

4.4 Baselines

We compared the performance of the proposed model with several classic and state-of-the-art MSA models across multiple metrics. The models compared and their brief descriptions are as follows:

TFN [7] generates a multidimensional tensor by computing the Cartesian product between modalities, enabling explicit capture of the interactions between single-modal, bi-modal, and tri-modal features.

LMF [9] improves upon TFN by using low-rank tensors for efficient multimodal fusion, significantly reducing computational complexity.

MulT [2] is based on an attention mechanism that maps the features of one modality to those of another, focusing on cross-modal interactions across time steps.

MISA [3] projects each modality into a modality-invariant subspace, aiming to learn the shared representation of each modality within that subspace.

Self-MM [10] generates unimodal labels through self-supervised tasks, thereby learning the similarities and differences between modalities.

TETFN [28] achieves consistent representations between modalities through a text-enhanced Transformer fusion network and maintains modality-specific differences through the generated unimodal labels.

SGIR [43] significantly reduces computational complexity and enhances noise resistance by facilitating indirect interactions between the central hub features and the modality-specific private features.

TMBL [44] processes multimodal data by introducing a bimodal and trimodal binding mechanism. It leverages an improved Transformer architecture to promote cross-modal interactions when extracting modality-invariant and modality-specific features, and enhances feature interactions through a fine-grained convolutional module.

FNENet [45] reduces the granularity gap between linguistic and non-linguistic modalities by vectorizing the non-linguistic modality. It then integrates non-linguistic information into a pre-trained language model using a sequence fusion mechanism.

SIMSUF [46] dynamically selects the dominant modality and performs feature enhancement and fusion based on this modality, thereby obtaining an effective multimodal fusion representation.

MCL-MCF [47] gradually reduces the representational differences between different modalities through multi-level contrastive learning, and enhances high-level semantic features by utilizing multi-layer convolutional fusion, thereby alleviating the heterogeneity issue in multimodal data.

4.5 Results and Comparison

We compared the proposed TGICP model with various classical and state-of-the-art baseline methods on the core datasets for multimodal sentiment analysis, namely CMU-MOSI and CMU-MOSEI. The experimental results are shown in Table 2.

Table 2: The results of TGICP on the CMU-MOSI and CMU-MOSEI datasets. The baseline results marked with ♦ are from [23]. Metrics marked with ↓ indicate that smaller values are better, while metrics marked with ↑ indicate that larger values are better. The optimal values are highlighted in bold

Models	**CMU-MOSI**					**CMU-MOSEI**				
	MAE↓	**Corr↑**	**F1-score% ↑**	**Acc-2% ↑**	**Acc-7% ↑**	**MAE↓**	**Corr↑**	**F1-score% ↑**	**Acc-2% ↑**	**Acc-7% ↑**
TFN♦	0.901	0.698	-/80.7	-/80.8	34.9	0.593	0.7	-/82.1	-/82.5	50.2
LMF♦	0.917	0.695	-/82.4	-/82.5	33.2	0.623	0.677	-/82.1	-/82.0	48
MulT	0.871	0.698	-/82.8	-/83.0	40	0.58	0.703	-/82.3	-/82.5	51.8
MISA	0.783	0.761	81.7/83.60	81.8/83.40	42.3	0.555	0.756	83.8/85.3	83.6/85.50	52.2
Self-MM	0.713	0.798	84.42/85.95	84.0/85.98	–	0.53	0.765	82.53/85.30	82.81/85.17	–
TETFN	0.717	0.8	83.83/86.07	84.05/86.10	–	0.551	0.748	84.18/85.27	84.25/85.18	–
SGIR	0.735	0.782	-/85.25	-/85.33	44.81	0.532	0.762	-/85.3	-/85.45	53.39
TMBL	0.867	0.762	82.41/84.29	81.78/83.84	36.3	0.545	0.766	**84.87**/85.92	84.23/85.84	52.4
FNENet	0.69	0.805	83.45/85.5	83.53/85.52	**48.25**	0.535	0.765	84.3/86.13	84.14/86.3	53.98
SIMSUF	0.709	0.802	-/85.98	-/86.08	45.72	0.529	0.772	-/86.12	-/86.23	53.68
MCL-MCF	0.692	0.799	84.7/87.2	84.9/87.3	–	0.536	0.767	84.4/86.3	84.2/86.4	–
TGICP(ours)	**0.688**	**0.809**	**85.6/87.58**	**85.53/87.55**	44.28	**0.522**	**0.776**	84.11/**86.77**	**84.53/86.69**	**54.02**

CMU-MOSI: The experimental results on the CMU-MOSI dataset demonstrate that our model outperforms traditional models and existing state-of-the-art (SOTA) methods across key metrics, including MAE, Corr, Acc-2, and F1. Specifically, compared to TETFN, which also performs cross-modal fusion based on text, our model improves by 0.029 in MAE (0.688 vs. 0.717) and 1.45% in Acc-2 (87.55% vs. 86.10%). Additionally, except for Acc-7, our model shows improvements on all evaluation metrics, with Corr and F1 increasing by

0.002 (0.809) and 0.38% (87.58%), respectively, further validating its superior performance in cross-modal sentiment analysis.

CMU-MOSEI: The experimental results on the CMU-MOSEI dataset demonstrate that the TGICP model performs exceptionally well in multimodal sentiment analysis tasks. Specifically, it outperforms the state-of-the-art method SIMSUF in both MAE (0.522 vs. 0.529) and Corr (0.776 vs. 0.772), while also improving Acc-2 and F1 scores by 0.29% and 0.47%, respectively. These improvements indicate that TGICP enhances sentiment prediction accuracy, correlation, and classification balance, further validating its advantages in unimodal feature enhancement and cross-modal interaction control.

In addition, to verify the reliability of the performance improvement of the proposed TGICP model, we conducted a Student's t-test on the CMU-MOSI dataset, selecting two important baseline models and a latest model with available source code. The results are shown in Table 3.

Table 3: The student's t-test results on CMU-MOSI are presented, with the variance of each metric indicated by the data on the right side of ±. Metrics marked with ↓ indicate that lower values are preferred, while those marked with ↑ indicate that higher values are preferred

Model	MAE↓	Acc-2% ↑	p(MAE) vs. TGICP↓	p(Acc-2) vs. TGICP↓
MISA	0.783 ± 0.0022	83.4 ± 0.224	0.0038	0.009
Self-MM	0.713 ± 0.0028	85.98 ± 0.130	0.0292	0.0151
TMBL	0.867 ± 0.0028	83.84 ± 0.025	0.0024	0.0067

As can be seen from the experimental results, in terms of the MAE metric, the p-values of all models compared to TGICP are less than 0.05, indicating a significant difference in MAE values between these models and the TGICP model. Specifically, the p-value for the MISA model is 0.00375, for the Self-MM model is 0.02916, and for the TMBL model is 0.00239, all of which indicate a significant difference in MAE compared to the TGICP model.

Regarding the Acc-2 metric, the p-values of all models compared to TGICP are also less than 0.05, indicating a significant difference in Acc-2 values between these models and the TGICP model. The p-values are particularly low, with the MISA model at 0.00009, the Self-MM model at 0.015129, and the TMBL model at 0.00067, further emphasizing the significant differences in Acc-2 compared to the TGICP model.

Overall, TGICP demonstrates a significant performance advantage over models such as MISA, Self-MM, and TMBL in both MAE and Acc-2 metrics. This indicates that our work has achieved a significant performance improvement in the multimodal sentiment analysis task compared to important and state-of-the-art models.

4.6 Ablation Study

1) Evaluation on components in TGICP: We conducted a detailed performance evaluation of the two main modules and their submodules in the TGICP model on the CMU-MOSI dataset, with the experimental results shown in Table 4. The experiments revealed that removing the ICP module led to the model being unable to obtain a more refined multimodal representation, while removing the TGI module resulted in the model's inability to fully leverage emotional information from nonverbal modalities during cross-modal interactions. The removal of any module or submodule significantly degraded the overall performance of the model. These results underscore the importance and indispensability of these modules within the entire model. Further analysis showed that the ICP module had the greatest overall impact on model performance, as its key role was to provide the model with more emotionally rich and complete sample features.

Table 4: Evaluation of the performance of each module of the model on CMU-MOSI

Description	MAE↓	Corr↑	Acc-2% ↑	F1-score% ↑
None modual	0.978	0.675	77.7/79.12	77.8/79.15
w/o ICP	0.705	0.806	81.34/83.38	81.37/83.35
w/o TGI	0.695	0.801	84.11/85.82	84.14/85.81
w/o NIE	0.691	0.808	83.53/85.21	83.56/85.20
TGICP	0.688	0.809	85.53/87.55	85.6/87.58

To gain a deeper understanding of the contribution of the ICP module, we also conducted separate evaluations of the two submodules within the ICP module: NIE and CEO. The experimental results show that removing the NIE module led to the loss of the commonality relationship between frames in nonverbal modalities. This loss resulted in insufficient capture of potential commonality information during the common feature extraction process, leading to a significant decline in model performance. A detailed analysis of the CEO module and its role in the model will be further discussed below.

2) Evaluation on CEO: In order to better leverage the common features extracted from the data, we conducted experimental evaluations on several commonly used methods from previous research using the CMU-MOSI dataset. The experimental methods include: adding common features to the original features (Add), concatenating the common features with the original features (Cat), and directly use common features (UCF) to replace original features for subsequent operations. The experimental results of these methods are shown in Table 5.

Table 5: Evaluation of the CEO module performance on CMU-MOSI

Description	MAE↓	Corr↑	Acc-2% ↑	F1-score% ↑
Add	0.745	0.798	82.65/84.45	82.65/84.11
Cat	0.726	0.796	81.34/83.69	81.45/83.72
UCF	0.694	0.81	82.22/84.6	82.3/84.61
CEO	0.688	0.809	85.53/87.55	85.6/87.58

The experimental results indicate that among the three common methods for utilizing common features, directly use common features (UCF) to replace original features achieved the best performance. Specifically, replacing the original features with common features significantly improves the model's performance in sentiment analysis tasks, as this method allows for a more direct utilization of common information, thereby optimizing feature representation. However, despite the excellent performance of the UCF method, there remains a noticeable gap compared to the results obtained using our proposed CEO module. By integrating a contextual sentiment optimization strategy, the CEO module can more efficiently consolidate the common features, further enhancing model performance. These experimental findings fully demonstrate the innovation and advancement of the CEO module in utilizing common features. Compared to traditional feature fusion methods, the CEO module can more accurately extract and leverage common features, significantly improving the model's sentiment recognition ability, and providing an effective direction for further research in this field.

3) Evaluation on missing modality: To assess the contribution of different nonverbal modalities to the model and the effectiveness of cross-modal information extraction in the presence of multimodal data, we conducted ablation experiments on the CMU-MOSI dataset, analyzing the model's performance under different input modality combinations. The experimental results are shown in Table 6. Since we employed a text-centered cross-modal interaction approach, the experiments considered only combinations of text and nonverbal modalities.

Table 6: Evaluation of performance with modalities missing on CMU-MOSI

Modalities	MAE↓	Corr↑	Acc-2% ↑	F1-score% ↑
T+V	0.687	0.809	84.4/86.43	84.46/86.44
T+A	0.708	0.807	82.36/84.15	82.4/84.12
T+V+A	0.688	0.809	85.53/87.55	85.6/87.58

The results indicate that, compared to the auditory modality, the visual modality contributes more significantly to the model's predictions. This suggests that visual information plays an essential role in sentiment recognition tasks, providing the model with more contextual cues to help capture emotional features more accurately. On the other hand, the three-modal combination (i.e., text, visual, and auditory modalities) outperforms the two-modal combinations (such as text and visual or text and auditory). This result further demonstrates that our proposed model can effectively extract cross-modal information in a multimodal data environment, integrating the strengths of each modality to improve the accuracy of sentiment analysis.

4) Evaluation on different modalities' splice order: To evaluate the impact of modality concatenation order during the feature fusion phase on model performance, we conducted experiments with different concatenation orders on the CMU-MOSI dataset. The experimental results are shown in Table 7. The results indicate that the T+V+A (Text + Visual + Auditory) order performed the best among all concatenation orders, achieving optimal performance across multiple evaluation metrics.

Table 7: Evaluation of the impact of feature concatenation order on CMU-MOSI

Splice order	MAE↓	Corr↑	Acc-2% ↑	F1-score% ↑
T+A+V	0.688	0.813	82.65/84.6	82.67/84.57
V+T+A	0.702	0.801	82.8/85.06	82.87/85.06
V+A+T	0.706	0.799	82.51/84.45	82.54/84.45
A+T+V	0.691	0.811	83.09/85.37	83.19/85.4
A+V+T	0.69	0.807	83.97/86.13	84.01/86.12
T+V+A	0.688	0.809	85.53/87.55	85.6/87.58

Specifically, the T+V+A order maximizes the retention of each modality's strengths during the fusion of features, while simultaneously enhancing the efficiency of cross-modal interaction. Textual information, as the core modality, provides the semantic foundation, while the visual and auditory modalities complement it by enriching the multi-dimensional expression of emotional information. This concatenation order effectively improves the model's ability to recognize complex emotional patterns and significantly outperforms other concatenation orders across multiple performance metrics. This result validates that the concatenation order chosen at the feature fusion stage is both reasonable and effective.

5) Evaluation on similarity threshold: During the selection of similar samples, we set the feature similarity threshold hyperparameter to control which samples can be selected as similar samples. To assess the impact of different thresholds on model performance, we conducted detailed experiments on the CMU-MOSI dataset. The experimental results are shown in Fig. 7. When the feature similarity threshold was set to 0.8, all evaluation metrics, except for the correlation (Corr) metric, achieved their optimal values, and the Corr value was also very close to the optimal value of 0.81.

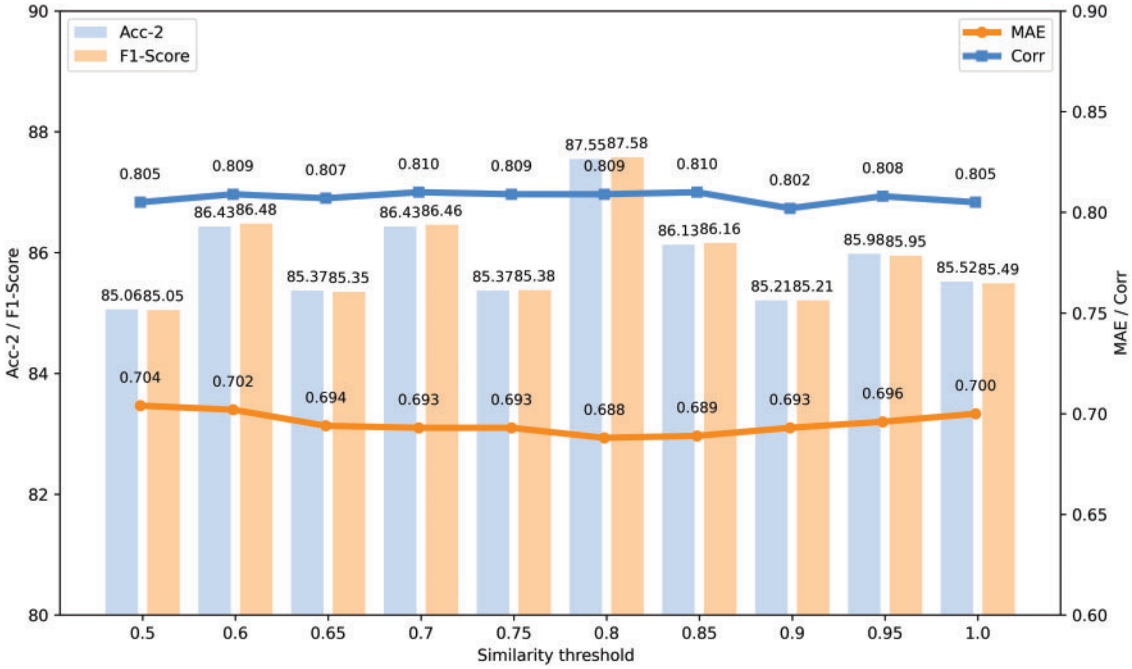

Figure 7: The impact of different feature similarity thresholds on model performance

The experimental results show that when the feature similarity threshold is set to 0.8, the model is able to select similar samples while minimizing the introduction of noise, thereby improving model performance. Both excessively high and low thresholds lead to inaccurate selection of similar samples, which negatively impacts the model's training effectiveness. A higher threshold may result in overly stringent filtering, limiting the number of valid samples, while a lower threshold could introduce excessive irrelevant noise, leading to instability in model training. Overall, selecting 0.8 as the feature similarity threshold is the most appropriate choice, as it effectively reduces noise while selecting similar samples within the modality, thus enhancing the model's performance in multimodal sentiment analysis tasks.

6) Evaluation on mutual information loss weight τ_s: To assess the impact of different mutual information loss weights on model performance, we conducted detailed experiments on the CMU-MOSI dataset, adjusting the weight of the mutual information loss and observing its effect on model performance. The experimental results are shown in Fig. 8. When the mutual information loss weight was set to 0.8, the model achieved the best performance. Under other weight settings, while the model performance showed some fluctuations, it did not surpass the optimal performance achieved with a weight of 0.8.

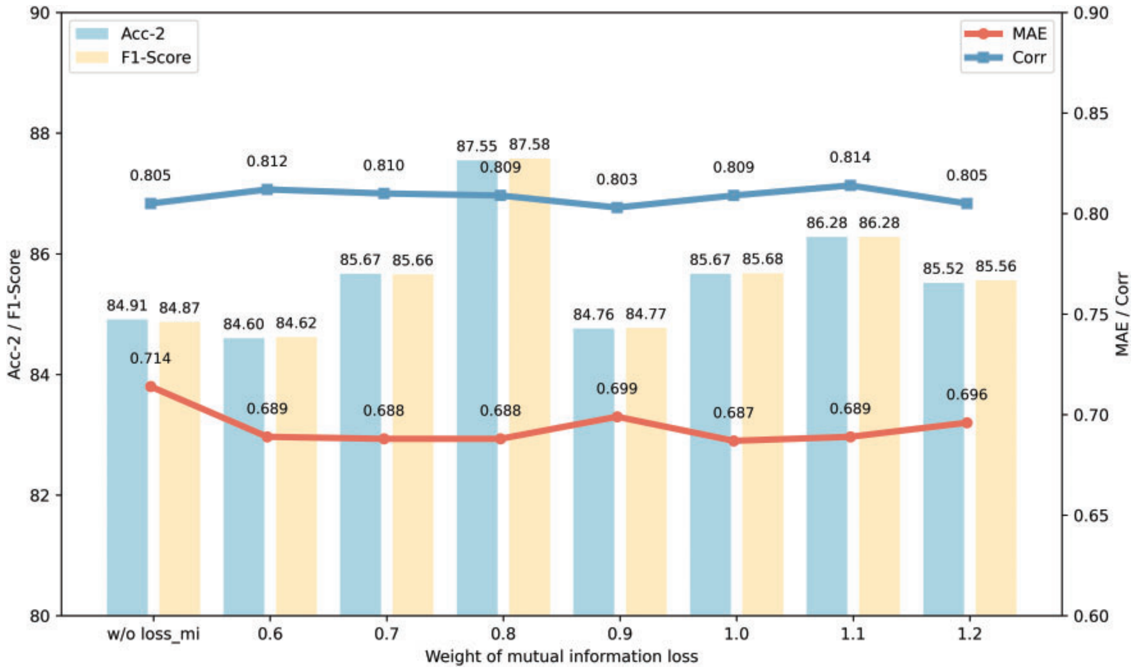

Figure 8: The impact of different mutual information loss weights τ_s on model performance

These results indicate that the setting of the mutual information loss weight has a significant impact on model performance. An appropriate weight can balance the dependencies between features within the modality and the extraction of common features, thus enhancing the model's performance in multimodal sentiment analysis. Specifically, when the weight is 0.8, the model effectively reduces the dependencies between common features of modality-specific samples and the original samples, thereby enhancing the independence of common features and allowing them to more accurately represent the shared information between modalities. Additionally, the experiment further demonstrates that introducing mutual information loss as a regularization method helps better extract common features, suppress irrelevant features, and significantly improve model performance.

7) Evaluation on batch size: During the model training process, we employed Mini-Batch Gradient Descent for training. To evaluate the impact of different batch sizes on model training effectiveness, we conducted experiments on the CMU-MOSI dataset. The experimental results are shown in Fig. 9. After comparing five common batch sizes, we found that when the batch size was set to 16, the model achieved the best overall performance, with outstanding results across multiple key metrics.

Specifically, when the batch size is set to 16, the model's Mean Absolute Error (MAE) and Correlation (Corr) are close to their optimal values, while Accuracy (Acc-2) and F1 score achieve their best performance. This indicates that a batch size of 16 provides a balanced training effect, ensuring high model precision while effectively capturing subtle differences in the data, particularly in sentiment analysis tasks. Additionally, smaller batch sizes may lead to unstable gradient updates during training, introducing noise and oscillation, which in turn affects convergence speed. On the other hand, larger batch sizes, while reducing such fluctuations, may slow down the training process and increase the risk of getting stuck in local optima. Therefore, choosing 16 as the batch size not only helps the model converge more stably during training but also improves its performance in sentiment analysis tasks.

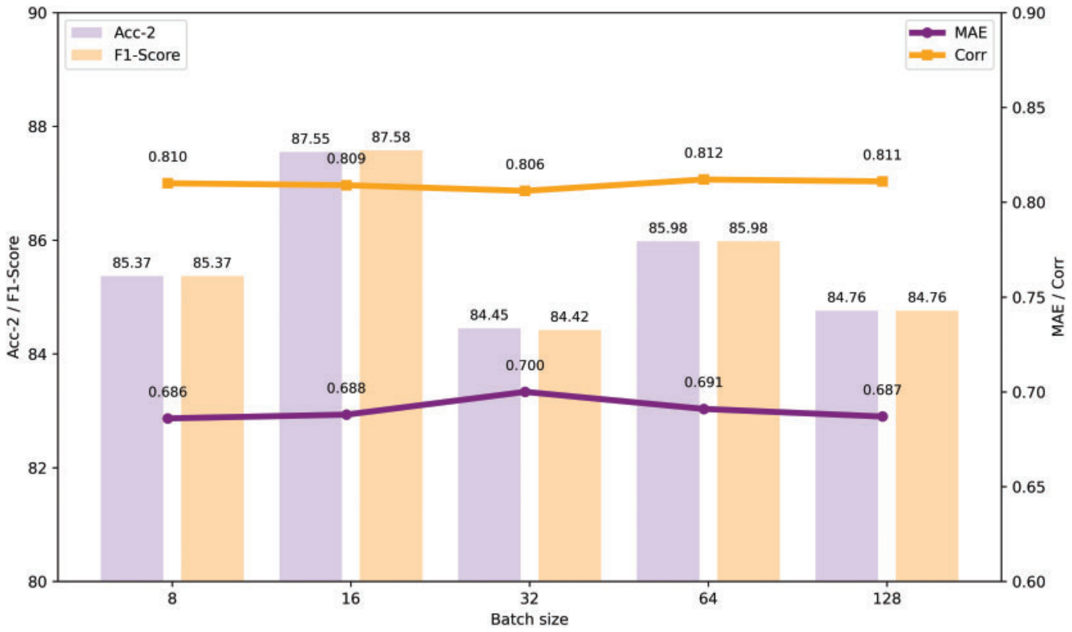

Figure 9: The impact of different batch sizes on model performance

In conclusion, a batch size of 16 is a suitable choice for the CMU-MOSI dataset, as it ensures high training accuracy while minimizing significant training oscillations, thereby accelerating model convergence and enhancing its overall performance.

8) Evaluation on imbalance dataset: To evaluate the model's robustness under imbalanced data distributions, we tested the model's binary classification accuracy by randomly removing varying proportions of negative, neutral, and positive samples from the CMU-MOSI training set. The results are shown in Fig. 10. The experimental results indicate that as the proportion of randomly removed samples from one class increases (from 0% to 40%), the degree of data imbalance gradually intensifies, and the model's binary classification accuracy changes accordingly. Under the imbalanced scenarios caused by a reduction in the number of negative and positive samples, the model's performance exhibits a noticeable decline. However, in the case of a reduction in neutral samples, despite an overall decline in performance, slight fluctuations were observed during the decrease. This could be due to the relatively small number of neutral samples, with the model's weaker prediction capability for boundary samples among the neutral samples. When these boundary samples are removed, the model's overall performance experiences a slight rebound. Overall, although the model's performance shows a slight downward trend as the degree of data imbalance increases, it still performs well. This suggests that the model demonstrates both sensitivity and a certain level of robustness when facing imbalanced data distributions.

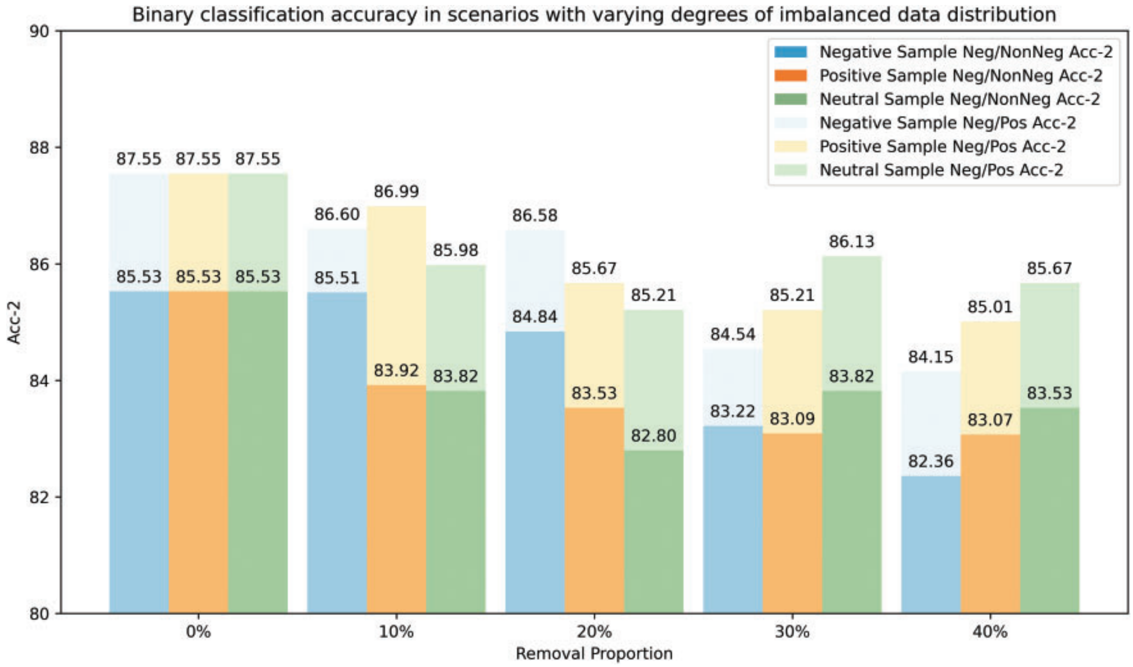

Figure 10: The model's binary classification accuracy under different levels of data distribution imbalance

4.7 Results of t-SNE Analysis for Sentiment Clustering

To demonstrate the model's ability to learn emotional features at different training optimization stages, we present the multimodal joint feature emotion clustering t-SNE plots for four training stages in Fig. 11. Points in the figure that are close to yellow represent a positive emotional bias, while points near purple indicate a negative emotional bias, and the green region represents neutral emotions. By mapping the joint features of the samples to colors, we can visually observe the model's representation distribution across different emotional polarities, thereby illustrating the model's learning effectiveness with respect to emotional features at different stages.

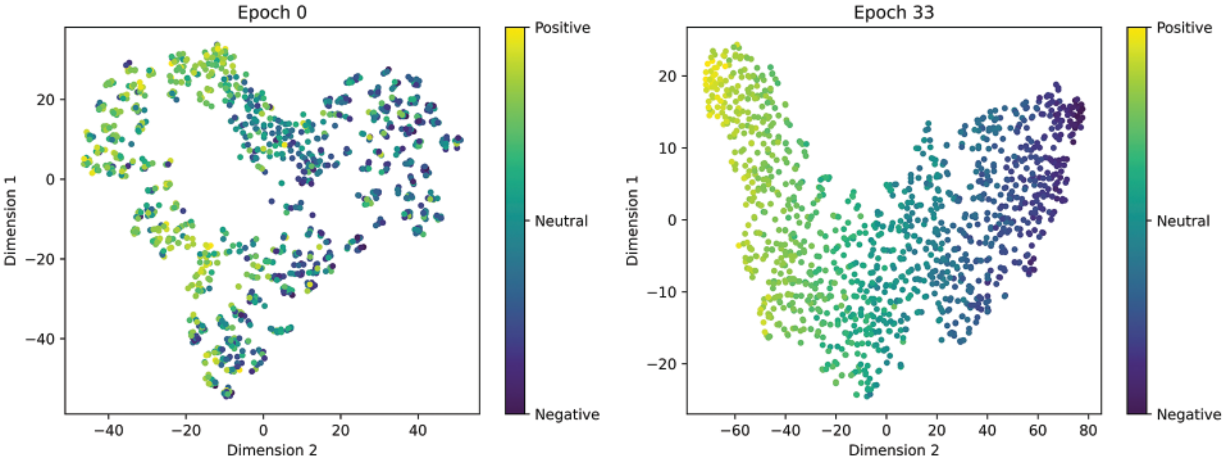

Figure 11: (Continued)

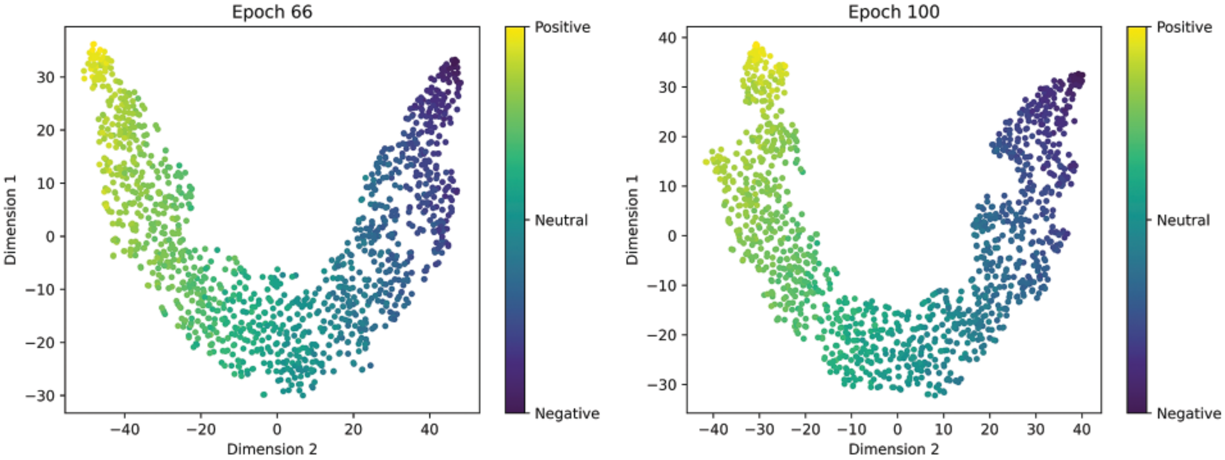

Figure 11: Results of t-SNE analysis for sentiment clustering at different learning stages: With the optimization of model training, the emotion clustering performance significantly improved. The distribution of emotional samples became more organized, and the overlap between categories was notably reduced

At the initial stage (Epoch 0), the feature space is chaotic, with high overlap among samples of different emotions, indicating that the model has not yet developed an effective ability to distinguish emotions. As training progresses (Epoch 33), an initial clustering trend begins to appear, but there is still noticeable confusion in the neutral emotional category, showing that the model has started to capture basic emotional features.

In the later stages of training (Epoch 66), the feature space is significantly optimized, with improved separation between categories and reduced overlap, confirming that the model has acquired a more precise emotional discrimination ability. At the final stage (Epoch 100), an ideal continuous emotional distribution is observed, with tight intra-class cohesion and natural inter-class transitions, which fully validates the model's outstanding multi-modal emotional understanding and classification performance. These results convincingly demonstrate the model's capability to accurately capture emotional features and highlight its excellent multi-modal emotional comprehension, providing direct evidence of the model's effectiveness in emotional classification.

5 Conclusion

This paper presents an innovative multimodal sentiment analysis model, TGICP. The model mines the common information of modality-specific similar samples through a unimodal inter-sample graph, and introduces a dual-perspective commonality embedding optimization layer. This layer enhances the role of common features while preserving the uniqueness of samples, thereby improving the model's generalization ability and fully leveraging the advantages of each modality. To address the issue of asymmetric information between modalities, the model incorporates a dynamic gating adjustment mechanism. This mechanism dynamically modulates the interaction process by calculating the information discrepancy between textual and non-textual modalities, thus enhancing the model's ability to learn from non-textual modalities.

The experiments demonstrate that TGICP significantly improves performance on two mainstream multimodal sentiment analysis datasets, but there is still room for improvement. Currently, the cross-sample graph does not fully consider the spatiotemporal relationships of non-text modalities and the word dependencies within the text modality, which may impact representation learning. Additionally, the common feature enhancement process may lead to excessive smoothing in the feature space, potentially introducing unnecessary errors in multiclass tasks. Moreover, due to the lack of large-scale pre-trained corpora with emotional annotations (similar to the English Yelp Dataset Challenge) and linguistic knowledge resources (similar to the English SentiWordNet) in the Chinese domain, SentiLARE can only be applied to English data, which limits the applicability of our model when dealing with non-English data.

Future work will focus on designing more refined methods for extracting emotional information within each modality to enhance representation learning and optimizing the common feature introduction mechanism to avoid the issue of feature space smoothing. Additionally, ongoing exploration of model optimization directions will aim to improve its robustness and applicability in a wider range of scenarios. Once again, although we have improved the temporal information modeling of non-text modalities by designing the NIE module to address the imperfections caused by BiLSTM, and achieved good performance, this improvement has not fundamentally solved the problem. Therefore, in our follow-up research, we will consider adopting other more sophisticated methods for handling temporal features to further optimize the model's ability to process temporal information. At the same time, we will actively explore pathways for obtaining and constructing large-scale pre-trained corpora in Chinese, along with targeted fine-tuning of SentiLARE; we will also assess and introduce more versatile language models as alternatives, achieving performance improvements through model fusion and parameter optimization.

Furthermore, given the current study's limitations in user-generated content privacy protection, subsequent research will employ techniques such as data de-identification and privacy-preserving enhancement algorithms to build a secure and compliant research framework, advancing the work while ensuring user privacy.

Finally, to comprehensively evaluate the performance of TGICP in complex labeling scenarios and real-world application environments, we plan to introduce real-world datasets containing multi-level semantic annotations and fine-grained emotional labels. This will allow us to systematically explore the model's capabilities in complex semantic understanding and scene adaptability. Additionally, we will simulate real application environments to design experimental setups, combining quantitative evaluation metrics and qualitative analysis methods to thoroughly validate the model's robustness and generalization ability, providing reliable performance evaluation for practical applications.

Acknowledgement: This paper was supported by the Natural Science Foundation of Henan, the Henan Provincial Science and Technology Research Project, the Jiangsu Provincial Scheme Double Initiative Plan, the XJTLU, the Science and Technology Innovation Project of Zhengzhou University of Light Industry, Higher Education Teaching Reform Research and Practice Project of Henan Province.

Funding Statement: This paper was supported by the Natural Science Foundation of Henan under Grant 242300421220, the Henan Provincial Science and Technology Research Project under Grants 252102211047 and 252102211062, the Jiangsu Provincial Scheme Double Initiative Plan JSS-CBS20230474, the XJTLU RDF-21-02-008, the Science and Technology Innovation Project of Zhengzhou University of Light Industry under Grant 23XNKJTD0205, and the Higher Education Teaching Reform Research and Practice Project of Henan Province under Grant 2024SJGLX0126.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Erlin Tian and Zuhe Li; methodology, Shuai Zhao; software, Shuai Zhao; validation, Min Huang and Yushan Pan; formal analysis, Min Huang; resources, Erlin Tian; data curation, Yihong Wang; writing—original draft preparation, Shuai Zhao; writing—review and editing, Zuhe Li and Yushan Pan; visualization, Yihong Wang; project administration, Erlin Tian; funding acquisition, Erlin Tian and Zuhe Li. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: All data generated or analyzed during this study are available from the corresponding author on reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Li Z, Huang Z, Pan Y, Yu J, Liu W, Wang H, et al. Hierarchical denoising representation disentanglement and dual-channel cross-modal-context interaction for multimodal sentiment analysis. Expert Syst Appl. 2024;252(5):124236. doi:10.1016/j.eswa.2024.124236.
2. Tsai Y-HH, Bai S, Liang PP, Kolter JZ, Morency L-P, Salakhutdinov R. Multimodal transformer for unaligned multimodal language sequences. In: Korhonen A, Traum D, Màrquez L, editors. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019 Jul; Florence, Italy. Florence, Italy: Association for Computational Linguistics; 2019. p. 6558–69.
3. Hazarika D, Zimmermann R, Poria S. MISA: modality-invariant and -specific representations for multimodal sentiment analysis. In: Proceedings of the 28th ACM International Conference on Multimedia; 2020 Oct; Seattle, WA, USA. New York, NY, USA: Association for Computing Machinery; 2020. p. 1122–31.
4. Wu Y, Lin Z, Zhao Y, Qin B, Zhu L-N. A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis. In: Zong C, Xia F, Li W, Navigli R, editors. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021; 2021 Aug; Online. Stroudsburg, PA, USA: Association for Computational Linguistics; 2021. p. 4730–8.
5. Liu Z, Cai L, Yang W, Liu J. Sentiment analysis based on text information enhancement and multimodal feature fusion. Pattern Recognit. 2024;156(11):110847. doi:10.1016/j.patcog.2024.110847.
6. Baltrušaitis T, Ahuja C, Morency L-P. Multimodal machine learning: a survey and taxonomy. IEEE Trans Pattern Anal Mach Intell. 2019;41(2):423–43. doi:10.1109/TPAMI.2018.2798607.
7. Zadeh A, Chen M, Poria S, Cambria E, Morency L-P. Tensor fusion network for multimodal sentiment analysis. In: Palmer M, Hwa R, Riedel S, editors. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing; 2017 Sep; Copenhagen, Denmark. Stroudsburg, PA, USA: Association for Computational Linguistics; 2017. p. 1103–14.
8. Zadeh A, Liang PP, Poria S, Vij P, Cambria E, Morency L-P. Multi-attention recurrent network for human communication comprehension. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence; 2018 Apr; New Orleans, LA, USA. Menlo Park, CA, USA: AAAI Press; 2018. 692 p.
9. Liu Z, Shen Y, Lakshminarasimhan VB, Liang PP, Zadeh BA, Morency L-P. Efficient low-rank multimodal fusion with modality-specific factors. In: Gurevych I, Miyao Y, editors. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2018 Jul; Melbourne, VIC, Australia. Stroudsburg, PA, USA: Association for Computational Linguistics; 2018. p. 2247–56.
10. Yu W, Xu H, Yuan Z, Wu J. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. Proc AAAI Conf Artif Intell. 2021 May;35(12):10790–7. doi:10.1609/aaai.v35i12.17289.
11. Poria S, Chaturvedi I, Cambria E, Hussain A. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In: Proceedings of the 2016 IEEE 16th International Conference on Data Mining (ICDM); 2016 Dec 12–15; Barcelona, Spain. p. 439–48.
12. Zadeh A, Liang PP, Mazumder N, Poria S, Cambria E, Morency L-P. Memory fusion network for multi-view sequential learning. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence; 2018 Apr 2–7; New Orleans, LA, USA. 691 p. doi:10.1609/aaai.v32i1.12021.
13. Majumder N, Hazarika D, Gelbukh A, Cambria E, Poria S. Multimodal sentiment analysis using hierarchical fusion with context modeling. Knowl-Based Syst. 2018;161(2):124–33. doi:10.1016/j.knosys.2018.07.041.

14. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 Dec 4–9; Long Beach, CA, USA. Red Hook, NY, USA: Curran Associates Inc.; 2017. p. 6000–10.
15. Lv F, Chen X, Huang Y, Duan L, Lin G. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 19–25; Nashville, TN, USA. Piscataway, NJ, USA: IEEE; 2021. p. 2554–62.
16. Rahman W, Hasan MK, Lee S, Bagher Zadeh AA, Mao C, Morency L-P, et al. Integrating multimodal information in large pretrained transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020 Jul; Online. Stroudsburg, PA, USA: Association for Computational Linguistics; 2020. p. 2359–69.
17. Han W, Chen H, Poria S. 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; 2021 Nov; Online. Stroudsburg, PA, USA: Association for Computational Linguistics; 2021. p. 9180–92.
18. Hou JM, Omar N, Tiun S, Saad S, He Q. TF-BERT: tensor-based fusion BERT for multimodal sentiment analysis. Neural Netw. 2025;185(6):107222. doi:10.1016/j.neunet.2025.107222.
19. Chen Y, Wang N, Zhang Z. DarkRank: accelerating deep metric learning via cross sample similarities transfer. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence; 2018; New Orleans, LA, USA. 348 p. doi:10.1609/aaai.v32i1.11783.
20. Huang Y, Lin Z. I2SRM: intra- and inter-sample relationship modeling for multimodal information extraction. In: Proceedings of the 5th ACM International Conference on Multimedia in Asia; 2024; Tainan, Taiwan. New York, NY, USA: Association for Computing Machinery; 2024. 28 p.
21. Li M, Wei Y, Zhu Y, Wei S, Wu B. Enhancing multimodal depression detection with intra- and inter-sample contrastive learning. Inf Sci. 2024;684:121282. doi:10.1016/j.ins.2024.121282.
22. Zhang H, Xu H, Long F, Wang X, Gao K. Unsupervised multimodal clustering for semantics discovery in multimodal utterances. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2024 Aug; Bangkok, Thailand. Stroudsburg, PA, USA: Association for Computational Linguistics; 2024. p. 18–35.
23. Huang Q, Chen J, Huang C, Huang X, Wang Y. Text-centered cross-sample fusion network for multimodal sentiment analysis. Multimed Syst. 2024;30(4):228. doi:10.1007/s00530-024-01421-w.
24. Sun Z, Sarma P, Sethares W, Liang Y. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. Proc AAAI Conf Artif Intell. 2020;34(5):8992–9. doi:10.1609/aaai.v34i05.6431.
25. Mai S, Xing S, Hu H. Analyzing multimodal sentiment via acoustic- and visual-LSTM with channel-aware temporal convolution network. IEEE/ACM Trans Audio Speech Lang Process. 2021;29:1424–37. doi:10.1109/TASLP.2021.3068598.
26. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); 2019 Jun; Minneapolis, MN, USA. Stroudsburg, PA, USA: Association for Computational Linguistics; 2019. p. 4171–86.
27. Liu Z, Lin W, Shi Y, Zhao J. A robustly optimized BERT pre-training approach with post-training. In: Li S, Sun M, Liu Y, Wu H, Liu K, Che W, et al., editors. Proceedings of the 20th Chinese National Conference on Computational Linguistics; 2021 Aug; Huhhot, China. Beijing, China: Chinese Information Processing Society of China; 2021. p. 1218–27.
28. Wang D, Guo X, Tian Y, Liu J, He LH, Luo X. TETFN: a text enhanced transformer fusion network for multimodal sentiment analysis. Pattern Recognit. 2023;136(2):109259. doi:10.1016/j.patcog.2022.109259.
29. Huang C, Zhang J, Wu X, Wang Y, Li M, Huang X. TeFNA: text-centered fusion network with crossmodal attention for multimodal sentiment analysis. Knowl Based Syst. 2023;269(4):110502. doi:10.1016/j.knosys.2023.110502.
30. Li Z, Liu P, Pan Y, Yu J, Liu W, Chen H, et al. Text-dominant multimodal perception network for sentiment analysis based on cross-modal semantic enhancements. Appl Intell. 2024;55(3):188. doi:10.1007/s10489-024-06150-1.

31. Ke P, Ji H, Liu S, Zhu X, Huang M. Sentiment-aware language representation learning with linguistic knowledge. In: Webber B, Cohn T, He Y, Liu Y, editors. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA, USA: Association for Computational Linguistics; 2020. p. 6975–88.
32. Toutanova K, Klein D, Manning CD, Singer Y. Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology—Volume 1. Edmonton, AB, Canada: Association for Computational Linguistics; 2003. p. 173–80.
33. Baccianella S, Esuli A, Sebastiani F. SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: Calzolari N, Choukri K, Maegaard B, Mariani J, Odijk J, Piperidis S, et al., editors. Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). Valletta, Malta: European Language Resources Association (ELRA); 2010.
34. Cheong JH, Jolly E, Xie T, Byrne S, Kenney M, Chang LJ. Py-feat: python facial expression analysis toolbox. Affect Sci. 2023 Dec 1;4(4):781–96. doi:10.1007/s42761-023-00191-4.
35. Degottex G, Kane J, Drugman T, Raitio T, Scherer S. A collaborative voice analysis repository for speech technologies. In: Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2014; Florence, Italy. p. 960–4.
36. Wang D, Liu S, Wang Q, Tian Y, He L, Gao X. Cross-modal enhancement network for multimodal sentiment analysis. IEEE Trans Multimed. 2023;25:4909–21. doi:10.1109/TMM.2022.3183830.
37. Mai S, Xing S, He J, Zeng Y, Hu H. Multimodal graph for unaligned multimodal sequence analysis via graph convolution and graph pooling. ACM Trans Multimed Comput Commun Appl. 2023;19(2):54–24. doi:10.1145/3542927.
38. Tishby N, Zaslavsky N. Deep learning and the information bottleneck principle. In: 2015 IEEE Information Theory Workshop (ITW); 2015; Jerusalem, Israel. p. 1–5. doi:10.1109/ITW.2015.7133169.
39. Lambert J, Sener O, Savarese S. Deep learning under privileged information using heteroscedastic dropout. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018; Salt Lake City, UT, USA. p. 8886–95.
40. van Den OA, Li Y, Vinyals O. Representation learning with contrastive predictive coding. arXiv:1807.03748. 2018.
41. Zadeh A, Zellers R, Pincus E, Morency L-P. MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv:1606.06259. 2016. doi: 10.48550/arXiv.1606.06259.
42. Zadeh AA, Liang PP, Poria S, Cambria E, Morency L-P. 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 (Volume 1: Long Papers); 2018; Melbourne, VIC, Australia. p. 2236–46.
43. Ding Z, Lan G, Song Y, Yang Z. SGIR: star graph-based interaction for efficient and robust multimodal representation. IEEE Trans Multimed. 2024;26:4217–29. doi:10.1109/TMM.2023.3321404.
44. Huang J, Zhou J, Tang Z, Lin J, Chen CY-C. TMBL: transformer-based multimodal binding learning model for multimodal sentiment analysis. Knowl Based Syst. 2024;285(1):111346. doi:10.1016/j.knosys.2023.111346.
45. Zheng C, Peng J, Wang L, Zhu L, Guo J, Cai Z. Frame-level nonverbal feature enhancement based sentiment analysis. Expert Syst Appl. 2024;258:125148. doi:10.1016/j.eswa.2024.125148.
46. Huang J, Ji Y, Qin Z, Yang Y, Shen HT. Dominant single-modal supplementary fusion (SIMSUF) for multimodal sentiment analysis. IEEE Trans Multimed. 2024;26(11):8383–94. doi:10.1109/TMM.2023.3344358.
47. Fan C, Zhu K, Tao J, Yi G, Xue J, Lv Z. Multi-level contrastive learning: hierarchical alleviation of heterogeneity in multimodal sentiment analysis. IEEE Trans Affect Comput. 2025;16(1):207–22. doi:10.1109/TAFFC.2024.3423671.