



Text-centered cross-sample fusion network for multimodal sentiment analysis

Qionghao Huang¹ · Jili Chen¹ · Changqin Huang¹ · Xiaodi Huang² · Yi Wang¹

Received: 29 March 2024 / Accepted: 15 July 2024 / Published online: 30 July 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

Abstract

Significant advancements in multimodal sentiment analysis tasks have been achieved through cross-modal attention mechanisms (CMA). However, the importance of modality-specific information for distinguishing similar samples is often overlooked due to the inherent limitations of CMA. To address this issue, we propose a **Text-centered Cross-sample Fusion Network (TeCaFN)**, which employs cross-sample fusion to perceive modality-specific information during modal fusion. Specifically, we develop a cross-sample fusion method that merges modalities from distinct samples. This method maintains detailed modality-specific information through the use of adversarial training combined with a task of pairwise prediction. Furthermore, a robust mechanism using a two-stage text-centric contrastive learning approach is developed to enhance the stability of cross-sample fusion learning. TeCaFN achieves state-of-the-art results on the CMU-MOSI, CMU-MOSEI, and UR-FUNNY datasets. Moreover, our ablation studies further demonstrate the effectiveness of contrastive learning and adversarial training as the components of TeCaFN in improving model performance. The code implementation of this paper is available at <https://github.com/TheShy-Dream/MSA-TeCaFN>.

Keywords Multimodal sentiment analysis · Contrastive learning · Multimodal fusion · Modality-specific information

1 Introduction

The burgeoning short-form video industry has led to an exponential growth in multimodal data, which includes text, audio, and video components [1–3]. As people increasingly

express their emotions, opinions, and thoughts across various media, there is a heightened need for sophisticated multimodal sentiment analysis tools to interpret these diverse modes of expression with precision [4]. Multimodal sentiment analysis has a wide range of practical applications. For instance, in social media monitoring, it helps brands understand public sentiment towards their products or services by analyzing images, videos, and text to track human interest and upgrade new designs [5]. In film studies, it could be used to track the evolution of characters' emotions throughout a movie to provide deeper insights into narrative structures and character development. For trend analysis in the context of sequential video frames, multimodal sentiment analysis can be applied to detect and analyze emotional changes over time. For instance, by applying multimodal sentiment analysis to a series of video frames, healthcare professionals can observe changes in a patient's facial expressions and body language over time, which can be crucial for understanding the progression of their emotional well-being to personalize care plans and improve patient satisfaction and outcomes [6]. Most multimodal sentiment analysis frameworks primarily focus on integrating different modalities, known as multimodal fusion [5]. Traditional RNN-based approaches

Communicated by Bing-kun Bao.

✉ Qionghao Huang
qhhuang@m.scnu.edu.cn

Jili Chen
irelia@zjnu.edu.cn

Changqin Huang
cqhuang@zju.edu.cn

Xiaodi Huang
xhuang@csu.edu.au

Yi Wang
wangyi@zjnu.edu.cn

¹ Zhejiang Key Laboratory of Intelligent Education Technology and Application, Zhejiang Normal University, Jinhua 321004, Zhejiang, China

² School of Computing, Mathematics and Engineering, Charles Sturt University, Albury, NSW 2640, Australia

like TFN [7] and MFN [8] concentrate on capturing temporal information and employ straightforward methods like concatenation or summation of modality-specific tensors. With the advent of transformers, models such as MuT [9] and MAG-BERT [10] have been developed to facilitate more effective interactions between modalities using transformers. Notably, MISA [11] and Self-MM [12] incorporate auxiliary tasks to bridge the divide between different modalities. Most methodologies are underpinned by the hypothesis, supported by numerous researchers [13, 14], that cross-modal attention mechanisms are capable of intuitively identifying the inter-connections between a given modality and others [5, 15].

One technique for enhancing cross-modal attention is to provide modalities with potential modal clues that capture desirable properties. Some recent work [16–19] advocate for the primacy of text as a modality, asserting its unique fusion capabilities not found in other modalities. They propose text enhancement to emphasize the importance of text-driven cross-modal fusion, leading to improved accuracy. However, within the framework of cross-modal attention fusion, a potential learning bias may exist. The cross-modal attention mechanism operates by measuring the relevance of features from one modality to a query in another, often text [20, 21]. This process inherently biases the integration towards samples highly similar to the textual query, potentially leading to the convergence of feature selection. As a consequence, modality-specific information from non-text modalities that do not closely align with the textual features may be overlooked or discarded, resulting in a loss of the original modality's distinctiveness.

Sentimental trend analysis in the context of continuous video frames is particularly challenging due to the abundance of similar samples. These frames often contain repetitive visual and audio cues, such as identical facial expressions or consistent speech patterns, which can make it difficult for models to distinguish between genuine sentiment changes and mere noise. The task requires sophisticated machine learning models capable of recognizing the subtle differences in sentiment even within these similar frames, emphasizing the importance of robust feature extraction and modality-specific information to capture the true sentimental trends. As illustrated in the left part of Fig. 1, during the sentiment semantic alignment (with text modality as the anchor, for instance) of different modalities, the subtle differences (modality-specific information) between the visual modality of two highly similar samples, S1 and S2, could be vanished once aligned with the text modality's sentiment semantics. The right part in Fig. 1 takes a text-centric cross-modal attention based model as instance. Sample 1 (S1) and Sample 2 (S2) share the same textual positive intensity but differ in emotional strength. Sample 1 is labeled as 'Positive, 2.0' and Sample 2 as 'Positive, 2.2', with a minor intensity difference of 0.2. During text-centered cross-modal attention

alignment and fusion, the model might neglect the nuanced differences in the video modality due to identical textual emotional expressions. This can result in an alignment that needs to account for these subtle variations within the video modality. Consequently, some cross-modal attention-based models (CMA-FN) incorrectly predict both Sample 1 and Sample 2 as 'Positive, 2.1', even though the actual label for Sample 1 is 'Positive, 2.0' and the actual label for Sample 2 is 'Positive, 2.2'. Hence, when samples exhibit a high degree of similarity, and their sentiments are challenging to differentiate based solely on modal consistency, it is crucial to consider fine-grained indicators from the modality-specific information for their differentiation. Ideally, modal fusion should not only amalgamate consistent information across modalities to minimize redundancy but also extract modality-specific information to enhance precision [11, 12].

When considering the retention of modality-specific information during modal fusion, an important question arises: *Q1: How can we utilize strategies within cross-modal attention mechanisms to ensure that modality-specific information is preserved for more precise predictions?* As depicted in Fig. 1, even though the textual and auditory modalities convey the same sentiment, a slight change in the visual modality leads to a 0.2 intensity difference between Sample 1 and Sample 2. It is crucial to devise a method to model these minor differences (modality-specific information) to distinguish highly similar samples. One potential solution is to employ a cross-sample learning method to capture these minor differences and effectively encapsulate modality-specific information. Nevertheless, the samples collected are from complex, real-world multimodal data, and the minute differences among similar samples are incredibly varied. Thus, developing a cross-sample learning method that can accurately model subtle variations and effectively preserve modality-specific information is a significant challenge, especially considering the inherent learning biases in cross-modal attention alignment. Furthermore, cross-sample learning methods often grapple with instability due to the pronounced variability among samples. Without a well-devised method for sample selection and feature learning constraint, the learning results can be highly erratic, with a risk of learning predominantly noise [22]. This brings us to an essential question, *Q2: How can we develop a sturdy mechanism within a cross-modal attention framework to enhance the stability of cross-sample learning methods and prevent the interference of unnecessary noise?* One crucial factor in implementing a cross-sample learning technique that captures modality-specific information between highly similar samples is to establish an appropriate similarity metric to filter out highly similar samples for cross-sample learning. Another vital factor is that some samples may fail to find similar samples for learning. In such cases, some poor features

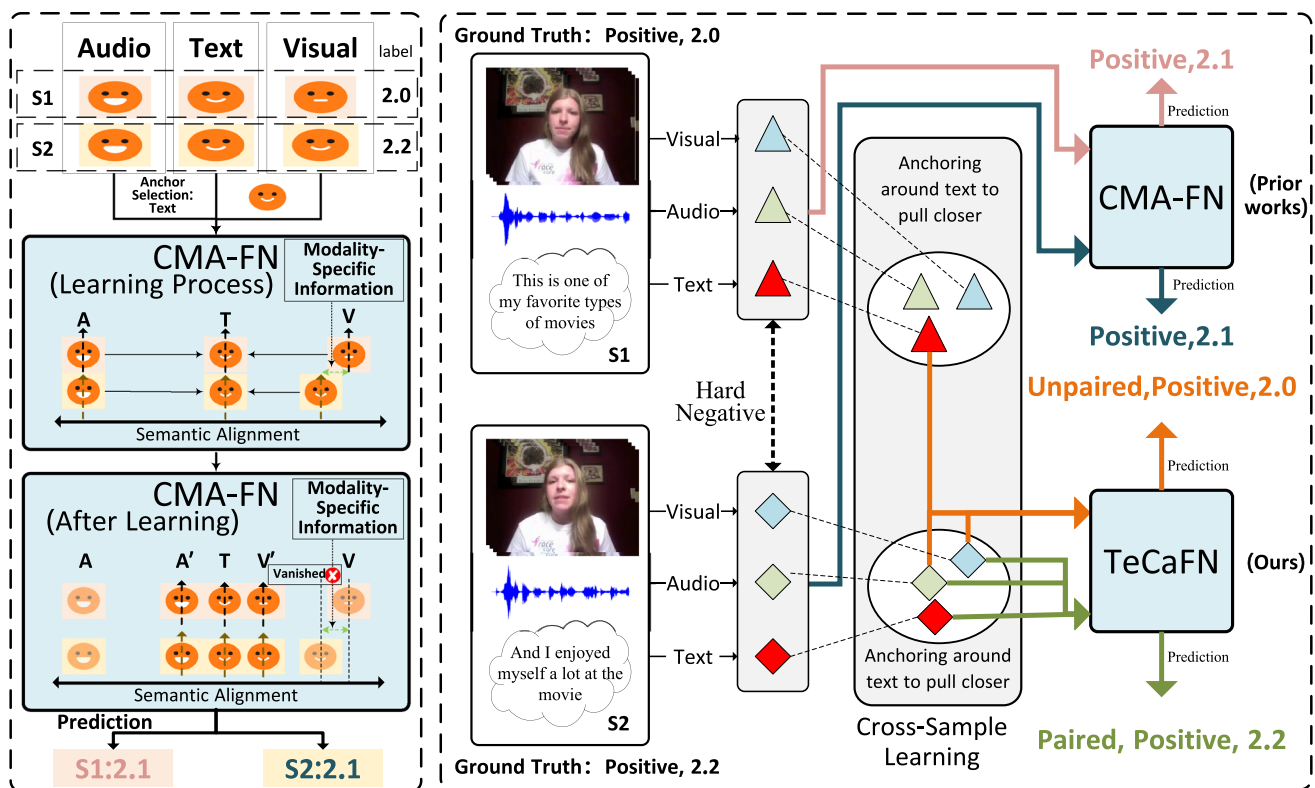


Fig. 1 Left: Illustration of learning biases of cross-modal attention fusion networks. Right: Comparisons between prior cross-modal attention based models and our TeCaFN with two samples from CMU-MOSI. S1 denotes Sample 1, while S2 for Sample 2. Hard negatives: Instances in other modalities that closely resemble the current ones. CMA-FN denotes previous CMA-based works like TETFN [17], which often fail to detect subtle emotional differences in hard negatives due to the neglect of modality-specific information preservation with their learning biases. In contrast, TeCaFN utilizes a cross-sample learning technique to retain modality-specific information,

thereby capturing these nuanced differences. Essentially, TeCaFN first aligns modalities with corresponding texts and then maintains the textual information constant, while other modalities may be fused with hard negatives. The ground truth is paired with the sample providing the text modality. In Sample 1, the text is fused with the hard negative, resulting in an “Unpaired” output (denoted as the orange line), while the text of Sample 2 is fused with itself, yielding a “Paired” output (denoted as the green line). With these configurations, TeCaFN can discern a subtle difference of sentiment intensity between Sample 1 and Sample 2 (colour figure online)

might be learned from these dissimilar samples, introducing unnecessary noise. Therefore, aligning and integrating the features obtained after cross-sample learning is also critical in ensuring the stability of cross-sample feature learning. These two aspects are critically significant in creating a sturdy mechanism within a cross-modal attention framework to strengthen the stability of cross-sample learning methods.

In addressing *Q1*, we must enhance the signals of emotional nuances in similar samples. As illustrated in Fig. 1, cross-sample learning involves a text-centric fusion process that integrates text with other modalities. Typically, the text from a given sample is combined with that of a highly similar sample from a different modality but not with itself. This fusion process incorporates adversarial training to extract the nuanced, modality-specific details across different modalities. Subsequently, a pairwise classification task, distinguishing between ‘Paired’ or ‘Unpaired’, is applied to the merged features to detect subtle differences

within highly similar modalities. In addressing *Q2*, we develop a two-stage contrastive learning approach within a cross-modal attention framework to solidify the cross-sample fusion process. The first stage utilizes unimodal contrastive learning to rank unimodal features by similarity, filtering out analogous samples for cross-sample fusion. The second stage implements bimodal contrastive learning to counteract the instability caused by poorly aligned similar samples with subpar features. Given the critical importance of text in sentiment prediction [16–18], a text-centric approach is applied in both stages of contrastive learning.

Put it together, we propose a novel network, **Text-centered Cross-sample Fusion Network (TeCaFN)**, that integrates the modules above and functionalities. Comparative experimental results on the CMU-MOSI, CMU-MOSEI, and UR-FUNNY datasets demonstrate that our proposed TeCaFN method attains the new state-of-the-art in multimodal sentiment analysis.

The contributions of this paper can be outlined as follows:

- We propose a novel text-centered cross-sample fusion network for multimodal sentiment analysis, which can capture modality-specific information within a cross-modal attention mechanism, enhancing the ability to distinguish between highly similar samples and thereby improving performance.
- We devise a two-stage contrastive learning strategy to facilitate the cross-sample fusion process, contributing to the stability of the proposed TeCaFN's predictions.
- We conduct extensive experiments on CMU-MOSI, CMU-MOSEI, and UR-FUNNY, with results showcasing that our proposed TeCaFN attains state-of-the-art performance across nearly all accuracy metrics.

The upcoming sections are assigned as follows: Sect. 2 offers a comprehensive overview of the latest developments in multimodal sentiment analysis. Section 3 delves into the specifics of our proposed model. Section 4 outlines the detailed procedures employed in our experiments. Lastly, Sect. 5 concludes this paper and presents future directions for exploration.

2 Related work

This section provides an overview of related work on multimodal sentiment analysis methods, mainly modality-specific information learning, and text-centered fusion methods.

2.1 Multimodal sentiment analysis

Multimodal sentiment analysis evaluates sentiments within data from multiple sources, primarily text, visual, and audio. Different modalities perceive the same event from unique perspectives, so the processes of feature extraction, feature fusion, and multimodal interaction across these modalities are crucial.

Zadeh et al. [7] propose the tensor fusion network to aggregate unimodal, bimodal, and trimodal information. Moreover, they use multiple LSTMs and design memory fusion network [8] for single-view and cross-view modeling, respectively. Majumder et al. [23] use GRU to model both modalities layer by layer to allow full interaction between the two modalities to extract information. Along with the breakthrough in unimodal with the transformer proposed by Google, sentiment analysis has begun to process modal information with the structure of the transformer. Tsal et al. [9] extend cross-attention to cross-modal attention, which enables different modal interactions to align and merge information from other

modalities based on attention weights. Rahman et al. [24] design adaption gates to enhance the centralized coding capabilities of BERT and XLNET for different modalities. Han et al. [25] combine mutual information maximization with modal interaction theory to minimize the depletion of crucial task-relevant information. Kim et al. [26] put all modalities after the fusion gate into a unified BERT encoder and use the pretraining task of BERT to fuse the modal features.

Previous work has made impressive achievements, but these works have focused on the extraction and fusion of unimodal features, and they have not paid attention to the importance of retaining modality-specific information [27]. So, we focus on preserving modality-specific information during multimodal fusion for better prediction.

2.2 Modality-specific information learning in MSA

Different modalities provide distinct types of information. For example, text can provide semantic information, speech can convey tone and intonation, and images can capture facial expressions and body language. Modality-specific features capture unique information from each modality, providing a more comprehensive and rich representation for sentiment analysis.

Hazarika et al. [11] propose MISA, which learns modality-specific knowledge by using modality-specific encoders for each modality, applying an orthogonal loss function to ensure non-redundant information between vectors and a reconstruction loss function to avoid learning trivial vectors. Peng et al. [28] devise a framework that captures modality-specific information by employing a separable tensor fusion network. This framework further conducts a Tucker decomposition on the weight tensor to extract modality-specific weights. To learn the modality-specific representations, Yu et al. [12] develop a module for unimodal label generation based on self-supervised strategies for each modality. Zhang et al. [29] design a network for feature fusion based on weights to diminish modality noise, paired with a modality-specific feature generator to preserve the unique characteristics of each modality while capturing multimodal interaction details. Ando et al. [30] employ a gated decoder to derive utterance-level embeddings from each modality, enhancing the representation with modality-specific information. Bo et al. [31] develop modality translation method based on Seq2Seq architecture.

The strategies mentioned above design distinct tasks to retain modality-specific information. The modules above are dedicated to capturing modal variability at the expense of modal coherence. So we design cross-sample fusion

to balance modal coherence and the capture of modality-specific information.

2.3 Text-centered fusion in MSA

Various modalities interpret a single event through distinct lenses, each offering unique insights into the underlying sentiment. Text, in particular, provides a direct and instinctive reflection of sentiment, leading numerous researchers to regard it as the foundational modality in multimodal sentiment analysis.

Sun et al. [16] control the text-based outer product matrix to learn valuable features by optimizing the Canonical Correlation Analysis (CCA) loss. Mai et al. [32] design gating mechanisms to enhance linguistic representation and correct textual information through other modalities. Attention-based methods are widely used in NLP, CV and multimedia [33–36]. They provide new methods for multimodal fusion and powerful text encoders for unimodal learning, such as BERT [34] and RoBERTa [37]. With the support of powerful text encoders, more work is being done to treat textual modalities as the core modalities in multimodal sentiment analysis. Wu et al. [38] undertake a cross-modal prediction task to investigate the shared and individual semantics within each non-textual modality. Wang et al. [17] use text-orient attentional mechanisms and cross-modal transformers to inject text-modal information into other modalities to enhance the representational capabilities of other modalities. Huang et al. [18] utilize text-centered cross-modal attention for unimodal and fused modalities to correlate the tri-modal data.

However, being text-centric does not equate to relying solely on text. Prior research has often overemphasized the importance of text, neglecting the value of other modalities. To our knowledge, extracting modality-specific information is essential in multimodal learning. Therefore, in this paper, we not only highlight the guiding role of text for other modalities but also introduce adversarial samples to retain modality-specific information for each modality.

3 Methodology

This section outlines the multimodal sentiment analysis task and introduces a comprehensive overview of Text-centered Cross-sample Fusion Network (TeCaFN).

3.1 Task definition

Multimodal sentiment analysis involves automated identification and comprehension of sentiments conveyed within video clips incorporating text, video, and audio. Each modality processes simultaneously in TeCaFN end-to-end, after which the corresponding modal encoders will be updated by gradient backpropagation. The model generates a sentiment score of -3 to 3 as output. Sentimental polarity can be inferred from sentiment scores according to positive and negative situations. Other notations are referenced in Table 1.

Table 1 Notations

Notation	Description	Notation	Description
t	Text modality	v	Vision modality
a	Audio modality	\mathbb{M}	The set of modalities
m	A specific modality	l_m	The length of m
d_m	The feature dimension of m	I_m^0	The model input of m with positional embeddings
$[m_1; \dots; m_l]$	The list of modality tokens	$LN(\cdot)$	The operation of layer normalization
$SeAt(\cdot)$	The operation of self attention	$FFN(\cdot)$	The feed-forward network function
X_m	The presentation of m after unimodal feature extraction	$g(\cdot)$	The 1D temporal convolution function
$MHA(\cdot)$	The operation of multi-head attention	$head_i$	The i -th head matrix in multi-head attention
$exp(\cdot)$	The exponential function	$FC(\cdot)$	The fully connected layer
Q	The query in multi-head attention	K	The key in multi-head attention
V	The value in multi-head attention	$CoAt_{m \rightarrow t}(\cdot)$	The cross-modal attention with t as Q and m as KV
$Pool(\cdot)$	The operation of mean pooling	\mathcal{L}_{task}	The task loss function
z_s	The fusion result	y	The ground truth
\hat{y}	The output of the TeCaFN	$MAE(\cdot)$	The mean absolute loss function
\mathcal{L}_{tcon}	The unimodal text centered contrastive learning loss	\mathcal{L}_{ap}	The pairwise prediction loss
\mathcal{L}_{bicon}	The bimodal contrastive learning loss	\mathcal{L}_{main}	The main loss function

3.2 Overall architecture

In this section, we elaborate on the specifics of the modules within TeCaFN, specifically the unimodal feature extraction, the text-centered unimodal contrastive learning module, the text-centered cross-sample fusion module, and the fused modal bidirectional contrastive learning module. The model architecture diagram is illustrated in Fig. 2.

3.2.1 Unimodal feature extraction

Different modalities contain unique information, necessitating the use of diverse feature extraction methods. In accordance with [39], position embedding is added to each modality sequence, enabling the model to grasp the temporal information of various modalities. The BERT model serves as the textual feature extractor [34]. The first token of the BERT tokenizer is denoted as t_{cls} , the last token is denoted as t_{sep} , and positional embeddings for text, video, and audio are denoted as t^{pos} , v^{pos} , and a^{pos} , respectively:

$$\begin{aligned} I_t^0 &= [t_{cls}; t_1; \dots; t_l; t_{sep}] + t^{pos}, \\ I_v^0 &= [v_1; \dots; v_l] + v^{pos}, \\ I_a^0 &= [a_1; \dots; a_l] + a^{pos}. \end{aligned} \quad (1)$$

In the case of video and audio modalities, to enhance their semantic representations for fusion, a L -layer transformer encoder is initially employed. This makes it more adaptable for mapping to various vector spaces. Subsequently, a single-layer unidirectional LSTM is used to capture temporal information.

$$\hat{I}_m^i = \text{SeAt}(\text{LN}(I_m^{i-1})) + \text{LN}(I_m^{i-1}), \quad m \in \{a, v\}, \quad (2)$$

$$I_m^i = \text{FFN}(\text{LN}(\hat{I}_m^i)) + \text{LN}(\hat{I}_m^i), \quad m \in \{a, v\}, \quad (3)$$

$$\hat{X}_m = \begin{cases} \text{BERT}(I_m^L), & m = t, \\ \text{LSTM}(I_m^L), & m \in \{a, v\}, \end{cases} \quad (4)$$

where \hat{I}_m^i denotes the modality representation of audio or vision after the i -th layer of self-attention and I_m^i denotes the modality representation of $m \in \{a, v\}$ after i -th layer of feed-forward, and i is ranging from 1 to L .

To guarantee comprehensive awareness among individual elements within the input sequence, a temporal convolution layer [9] is employed to remap the dimension for each modality:

$$X_m = \text{Conv1D}(\hat{X}_m, k_m), \quad m \in \{t, a, v\}, \quad (5)$$

where k_m represents the convolution kernel of the temporal convolution layer and X_m represents the output of modality m after 1D temporal convolution.

3.2.2 Text-centered unimodal contrastive learning for similarity ranking (Q2)

For effective learning of modality-specific information across samples, especially those that are highly similar, it is crucial to filter out similar samples. Given that text provides the most direct sentiment feedback [17], we utilize pairs of text-based modalities, denoted as (t, m) , to learn a joint embedding for similarity ranking. Given text t_i and any other modality m_i , then the depth-wise Conv1Ds with pooling layers are used to map the two modalities to the same dimension: $q_i = f_1(t_i)$ and $k_i = f_2(m_i)$ where f_1, f_2 are corresponding Conv1D mappers. The unimodal encoders and embeddings are optimized using the InfoNCE [40] loss:

$$\mathcal{L}_{t,m} = -\log \frac{\exp(q_i k_i^\top / \tau)}{\exp(q_i k_i^\top / \tau) + \sum_{j \neq i} \exp(q_j k_j^\top / \tau)}, \quad (6)$$

where τ is a scalar temperature coefficient controlling softmax smoothness. In a minibatch, matched modality pairs are indexed as i and termed positive samples, and unmatched modal pairs are negative pairs indexed as j [41]. InfoNCE aims to bring q_i and k_i closer in the embedding space. This process ensures that the other modalities will be closer to the corresponding textual modalities. In practice, an asymmetric text-centered loss $\mathcal{L}_{tccn} = \mathcal{L}_{t,a} + \mathcal{L}_{t,v}$ is exploited. Then, the normalized similarity matrix generated by unimodal contrastive learning is used to mine similar samples (i.e., hard negatives):

$$S_{t,m} = \frac{q_i k_i^\top}{\|q_i\|_2 \|k_i\|_2}, \quad m \in \{a, v\}, \quad (7)$$

where $S_{t,m} \in \mathbb{R}^{N \times N}$ is the normalized similarity matrix reflecting the similarity of the modality t and modality m in the batch. Given the normalized similarity matrix $S_{t,m}$ that quantifies the degree of similarity between modalities t (text) and m (non-text), the process of extracting hard negatives involves identifying the top num most similar non-text samples to a given text sample t . This is achieved by sorting the similarity scores associated with each text sample t in descending order and selecting the first num entries as the hard negatives. These hard negatives are samples from modality m that are highly similar to the text but are not the correct corresponding matches, thus providing challenging examples for the learning algorithm to distinguish from true positives. The pseudocode for mining similar samples is outlined in Algorithm 1.

Algorithm 1: Hard Negative Mining Pseudocode, NumPy-like

```

#  $I_m$ : minibatch of modality  $m$  [batchsize,  $l_m, d_m$ ]
#  $I_t$ : minibatch of textual modality [batchsize,  $l_t, d_t$ ]
#  $f_m$ : encoder for modality  $m$ 
#  $f_t$ : encoder for textual modality
#  $W_m$ : learned proj of modality  $m$  to embed [ $d_m, d_c$ ]
#  $W_t$ : learned proj of modality  $t$  to embed [ $d_t, d_c$ ]
# num: the number of hard negatives

# Unimodal Feature Extraction
 $X_m = f_m(I_m)$  #[batchsize,  $l_m, d_m$ ]
 $X_t = f_t(I_t)$  #[batchsize,  $l_t, d_t$ ]

# Multimodal Embedding Learning [batchsize,  $d_c$ ]
 $I_m = \text{l2norm}(\text{np.dot}(\text{np.mean}(X_m, \text{axis}=1), W_m), \text{axis}=1)$ 
 $I_t = \text{l2norm}(\text{np.dot}(\text{np.mean}(X_t, \text{axis}=1), W_t), \text{axis}=1)$ 

# Scaled Pairwise Cosine Similarities  $S_{t,m}$  [batchsize, batchsize]
logits =  $\text{np.dot}(I_t, I_m.T)$ 

# Obtain the indices of hard negative samples [batchsize, num]
indices =  $\text{np.argsort}(\text{logits}, \text{axis}=1)[:, -\text{num}:]$ 

```

3.2.3 Text-centered cross-sample fusion for capturing modality-specific information (Q1)

TeCaFN utilizes adversarial training between modalities to retain modality-specific information, in conjunction with a task of pairwise prediction. In this module, text modality is functioned to guide other modalities during adversarial training. Specifically, keeping the text unchanged, we identify the *num* most hard negatives (similar samples) for the audio and visual modalities. Next, the positive key, i.e., the visual or audio modality that matches the text, and the hard negatives are put into the replacement gate.

The replacement gate returns one of the hard negatives of modality m with probability p and the original matching modality m with probability $1 - p$. The similar modalities are sampled from the uniform distribution of the hard negatives. At the same time, two-dimensional pseudo-targets are generated, with 1 indicating a matching modal pair and 0 indicating a mismatching modal pair for modality m . The replacement gate is only utilized at training, and during validation and testing, only in-sample modal fusion is implemented. A detailed table of replacement probabilities can be viewed in Table 2.

TeCaFN uses the text modality as the query and the modality m after the replacement gate as the key and value to design text-centered cross-sample attention to fuse the modal efficiently. The following equations explain how text-centered cross-sample attention fuses text modality t and any other modality m :

$$MHA(Q, K, V) = \text{Concat}(\text{head}_1 \cdots \text{head}_h)W_o, \quad (8)$$

$$\text{head}_i = \text{Softmax}\left(\frac{(QW_i^Q)(KW_i^K)^T}{\sqrt{d_k}}\right)(VW_i^V), \quad (9)$$

$$\text{CoAt}_{m \rightarrow t} = MHA(X_t, X_m, X_m), \quad m \in \{a, v\}, \quad (10)$$

where $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$ and $W_o \in \mathbb{R}^{hd \times d_{\text{model}}}$ represent the projection matrix of the different branches in the multi-head attention, h represents the num of heads and d_k represents the dimension of cross-modal attention. We use the feed-forward layer to make cross-modal features more flexible. The L -layer cross-modal attention can be expressed as:

$$\hat{h}_{tm}^i = \text{CoAt}_{m \rightarrow t}(\text{LN}(h_{tm}^{i-1})) + \text{LN}(h_{tm}^{i-1}), \quad m \in \{a, v\}, \quad (11)$$

$$h_{tm}^i = \text{FFN}(\text{LN}(\hat{h}_{tm}^i)) + \text{LN}(\hat{h}_{tm}^i), \quad m \in \{a, v\}, \quad (12)$$

where \hat{h}_{tm}^i denotes the cross-modal hidden representation of text and modality m after the i -th layer of cross-attention and h_{tm}^i denotes the cross-modal hidden representation of text and modality m after the i -th layer of feed-forward network, and i is ranging from 1 to L .

One branch in the cross-sample fusion module is dedicated to predicting the pairwise situation of modalities, assessing whether the replacement gate substitutes the initial video or audio modality with hard negatives. We take this branch through the fully connected layer and output a

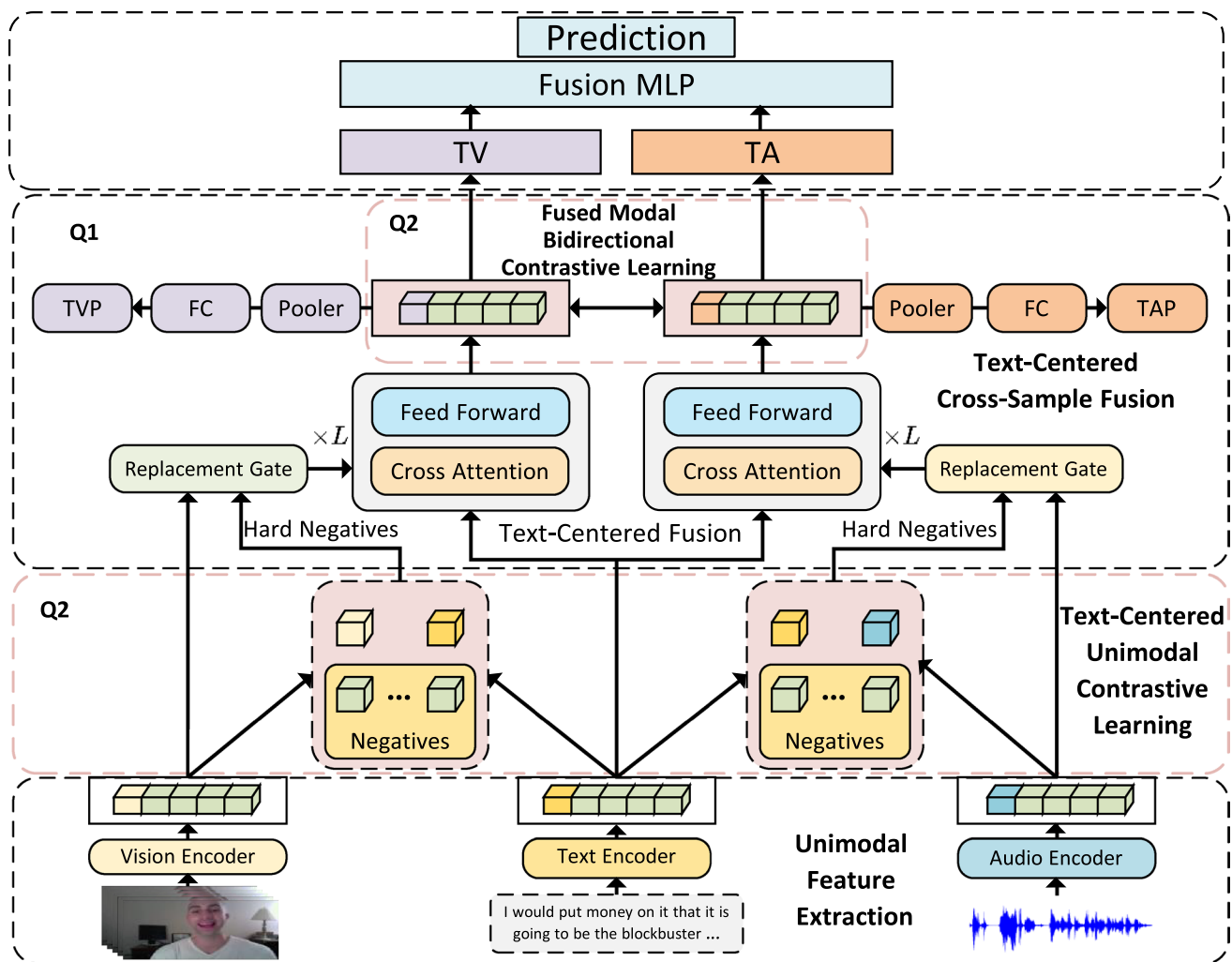


Fig. 2 The model diagram of our proposed TeCaFN. Cubes represent modality vectors. Non-green cubes represent multi-modal vectors in the sample being processed, while green cubes represent multi-modal vectors from other samples in the same batch. TV and TA respectively

represent the fusion results of text with video, and text with audio. Q1 and Q2 denote the modules that are specifically designed to address the first and second questions raised in the introduction section (colour figure online)

two-dimensional vector after mean pooling, which corresponds to the two-dimensional pseudo-target in the previous section, and use cross-entropy to optimize it:

$$\hat{y}_{imp} = FC(Pool(h_{im}^L)), \quad m \in \{a, v\}, \quad (13)$$

$$\mathcal{L}_{imp} = \mathbb{E}_{t,m} H(y_{imp}, \hat{y}_{imp}), \quad m \in \{a, v\}, \quad (14)$$

where \hat{y}_{imp} denotes the branch output which is a two-dimensional vector, H denotes the cross-entropy loss function, and y_{imp} represents the two-dimensional pseudo-target generated in the replacement gate. Hard negatives tend to be semantically similar, with only fine-grained differences. So hard negatives allow the model to perceive more fine-grained features and capture more modality-specific information. In the implementation, $\mathcal{L}_{ap} = \mathcal{L}_{tp} + \mathcal{L}_{tap}$ are designed to optimize text-to-visual and text-to-audio pairwise predictions.

Table 2 Replacement probability

Situation	Prob	Label
Text origin-vision origin-audio	$(1-p)^2$	(1, 1)
Text vision hard negative origin-audio	$p(1-p)$	(0, 1)
Text origin-vision audio hard negative	$p(1-p)$	(1, 0)
Text vision hard negative audio hard negative	p^2	(0, 0)

3.2.4 Fused modal bidirectional contrastive learning for stable modality-specific information mining process (Q2)

The text-centered cross-sample fusion module may cause instability in the uniform embedding space for the fused modalities. During unimodal contrastive learning, the modalities from the same samples are pulled closer and modalities from different samples farther are pushed apart in embedding space. This makes the vector similarity of different samples always far from each other during the cross-sample fusion, even under hard negative conditions. Furthermore, the three modalities involved in the fusion may come from three different samples due to the presence of replacement gates. In addition, some samples may fail to find similar samples for learning, which introduces unnecessary noise raises the instability of cross-sample learning. The text might be unable to query useful information from other modalities to obtain consistent predictions.

Upon the aforementioned analysis, we have discovered that each sample does not necessarily fuse with the most suitable one, hence the position of the fused modalities in the embedding space is sub-optimal. To correct for the bias, we focus on the fact that the text remains unchanged. Using the text modality as a guide, we can optimize the positioning of the fused modalities in the embedding space. We pull the distances between the fused modalities with the same text modality closer in the feature space, making them more inclined to reach consistent conclusions. By adjusting the fused modalities bidirectionally in the feature space, we can thereby mitigate the instability issues brought about by cross-sample fusion. Specifically, we introduce the fused modal bidirectional contrastive learning module. The output of the cross-modal fusion module is denoted as h_{tm}^L . Similar to unimodal contrastive learning, depth-wise Conv1Ds with pooling layers are employed to map the two cross-modal vectors to the same dimension: $z_i = g_1(h_{tv_i}^L)$ and $u_i = g_2(h_{ta_i}^L)$ where g_1, g_2 are corresponding Conv1D mappers. The cross-modal encoders and embeddings are optimized using the InfoNCE [40] loss:

$$\mathcal{L}_{tv,ta} = -\log \frac{\exp(z_i u_i^T / \tau)}{\exp(z_i u_i^T / \tau) + \sum_{j \neq i} \exp(z_j u_j^T / \tau)}, \quad (15)$$

where $\mathcal{L}_{tv,ta}$ is the loss function for text-to-visual and text-to-audio pairwise predictions, z_i and u_i are the embeddings of the cross-modal vectors mapped to the same dimension through Conv1D mappers, τ is the temperature parameter that scales the dot product of the embeddings. The fraction inside the logarithm represents the similarity between the embeddings of the same sample (positive pair) over the

sum of similarities of all different sample pairs (negative pairs), excluding the positive pair. The negative log likelihood function is used to minimize the loss, which encourages the model to increase the similarity for positive pairs and decrease it for negative pairs.

Here, TeCaFN employs contrastive learning from different views. Within a mini-batch, the modal fusion vectors using the same text will be pulled closer while the modal fusion vectors with the different text will be pushed farther. In practice, the loss can be expressed as:

$$\mathcal{L}_{bicon} = \mathcal{L}_{tv,ta} + \mathcal{L}_{ta,tv} \quad (16)$$

where $\mathcal{L}_{tv,ta}$ and $\mathcal{L}_{ta,tv}$ represent the contrastive losses for text-to-visual and text-to-audio modalities respectively. The total loss is then minimized during training, which encourages the model to learn modality-specific representations that are similar for the same text but different for different texts. This approach effectively enhances the discriminative power of the learned features and improves the performance of the multi-modal sentiment analysis task.

3.3 Objective function and prediction

The fusion result z_s is processed through an MLP with activation functions for final predictions:

$$\hat{y} = \tanh(FC(z_s)), \quad (17)$$

where \hat{y} is the final output of the whole model for training, validation, and testing.

We consider the multimodal sentiment analysis task as a regression task, employing mean absolute error (MAE) to gauge the loss between predicted values \hat{y} and actual values y :

$$\mathcal{L}_{task} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|, \quad (18)$$

where n denotes the batch size.

Finally, we summarise all the losses above, including task loss (\mathcal{L}_{task}), text-centered unimodal contrastive learning loss (\mathcal{L}_{tcon}), pairwise prediction loss (\mathcal{L}_{ap}), bimodal symmetric contrastive learning loss (\mathcal{L}_{bicon}) and constitute the final loss function as follows:

$$\begin{aligned} \mathcal{L}_{con} &= \mathcal{L}_{tcon} + \mathcal{L}_{bicon}, \\ \mathcal{L}_{main} &= \mathcal{L}_{task} + \gamma_1 \mathcal{L}_{con} + \gamma_2 \mathcal{L}_{ap}, \end{aligned} \quad (19)$$

where γ_1 is the hyperparameter for controlling contrastive learning loss, γ_2 is the hyperparameter for controlling pairwise prediction loss. The pseudocode of our algorithm is outlined in Algorithm 2.

Algorithm 2: Text-centered Cross-sample Fusion Network (TeCaFN)

Input : $D = \{(\mathbb{M}_t, \mathbb{M}_a, \mathbb{M}_v), Y\}$, γ_1, γ_2 , learning rate $\eta_{main}, \eta_{bert}, N_{epochs}$
Output: Prediction \hat{y} # *sentiment score*

for each epoch do

for mini-batch $\{(t_i, a_i, v_i), Y_i\}_{i=1}^B$ **from** D **do**

 Add positional embeddings to each modality as (1)

 Encode I_m^i into \hat{X}_m^i as (2-4)

 Map X_m^i into F_m^i using Conv1D as (5)

if training then

 Execute unimodal contrastive learning and calculate \mathcal{L}_{tcon} as (6)

 Mine hard negatives as (7)

 Execute cross-sample $CoAt_{a \rightarrow t}$ and $CoAt_{v \rightarrow t}$ as (8-12)

 # *fusion with hard negatives based on probability*

 Calculate pairwise prediction loss L_{ap} as (13-14)

 Execute bimodal contrastive learning and calculate \mathcal{L}_{bicon} as (15-16)

 Produce predictions \hat{y} as (17)

 Calculate task loss \mathcal{L}_{task} as (18)

 Sum sub-loss over to get main loss \mathcal{L}_{main} as (19)

 Mini-batch gradient descent

 Update model parameters

end if

else

 Compute $CoAt_{a \rightarrow t}$ and $CoAt_{v \rightarrow t}$ as (8-12)

 Produce predictions \hat{y} as (17)

end if

end for

end for

3.4 Discussion on the role of hard negatives in cross-sample attention

Traditional methods fuse the three modalities through cross-modal attention mechanism within a single sample [9, 25], yet this approach blindly align and fuse different modalities, resulting in the loss of modality-specific information [42], this paper proposes a cross-sample fusion approach. The cross-sample fusion allows the network to learn and complement missing modality-specific information to make final predictions.

Building upon this, we employ hard negatives to enhance cross-sample fusion considering three main aspects: Firstly,

hard negatives simulate a noisy real-world scenario, demanding text to extract modality-specific information from the noisy representations of other modalities. Secondly, we can amplify modality-specific information by combining hard negatives with the pairwise prediction task, which enhance the discrimination ability of similar samples. Thirdly, hard negatives bear a high resemblance to the original samples, indicating that the fusion of text with hard negatives maintains multimodal semantic consistency. This is crucial for maintaining stability during training. To sum up, training models on hard negatives helps enhance model robustness, stability and generalization.

4 Experiments

This section presents comprehensive information regarding our experimentation, including datasets, baselines that TeCaFN is compared with in unaligned scenarios, evaluation metrics, the ablation study, the analysis of parameter sensitivity, and experiment findings.

Table 3 Dataset splits for CMU-MOSI, CMU-MOSEI, and UR-FUNNY

Dataset	Train	Valid	Test	All
CMU-MOSI	1284	229	686	2199
CMU-MOSEI	16,326	1871	4659	22,856
UR-FUNNY	7614	980	994	9588

4.1 Datasets

We perform experiments on two publicly available datasets, CMU-MOSI [43] and CMU-MOSEI [44], and UR-FUNNY [45], to assess the effectiveness of the proposed TeCaFN for MSA. Our approach to dataset splitting aligns with the original dataset scheme, detailed in Table 3.

CMU-MOSI. This dataset encompasses a diverse collection of 2199 video clips from various sources, such as movie reviews, interviews, and monologues. Every video clip in the dataset is accurately labeled with sentiment intensity scores ranging from -3 to 3 with negative values representing varying degrees of negativity, 0 denoting neutrality, and positive values indicating varying degrees of positivity.

CMU-MOSEI. This dataset extends and improves upon CMU-MOSI, sourced from monologue videos on YouTube, containing about 3228 videos and 23,453 sentences, involving 1000 narrators across 250 topics. The sentiment scoring in CMU-MOSEI follows the same annotation approach as CMU-MOSI.

UR-FUNNY. This dataset is a multimodal collection focused on humor detection. It comprises video clips, audio, and text annotations, capturing a wide range of humorous expressions. It employs a binary annotation scheme to categorize the content into two classes: humorous and non-humorous.

4.2 Evaluation metrics

Drawing from prior research [25], the evaluation metrics used for CMU-MOSI and CMU-MOSEI are as follows:

Mean absolute error (MAE): This measures the average of the differences between the predicted sentiment scores and the actual scores.

Pearson correlation (Corr): This assesses the strength and direction of the linear relationship between the predicted and actual sentiment scores.

Classification accuracy (ACC): This is used to evaluate the model's ability to classify sentiment scores into categories: **ACC-7:** It divides sentiment scores into seven equal intervals from -3 to 3 and evaluates fine-grained accuracy. **ACC-2:** It uses two settings for binary classification: Positive/Negative (P/N): Scores greater than 0 are positive, less than 0 are negative. Non-negative/Negative (NN/N): Scores less than 0 are negative, 0 or greater are non-negative.

F1 Score (F1): This is a balance between precision (the accuracy of positive predictions) and recall (the fraction of actual positives correctly predicted), providing a single measure of a model's accuracy.

According to previous studies [11, 45], for UR-FUNNY, we use ACC-2 to measure the accuracy rate of binary classification predictions.

4.3 Baselines

Mainstream state-of-the-art MSA models have been chosen to exhibit the performance of the TeCaFN model.

TFN [7]: TFN (2017) use tensor-based operations to capture complex interactions and relationships between modalities.

LMF [46]: LMF (2018) creates a module for efficient multimodal fusion without compromising performance by using low-rank weight tensors.

MFM [47]: MFM (2019) factorizes multimodal representations into multimodal discriminators and modality-specific generators to capture information within and between modalities.

MuT [9]: MuT(2019) incorporates transformer layers and cross-modal attention mechanisms to process multiple modalities simultaneously.

ICCN [16]: ICCN (2020) utilizes an outer-product for extracting cross-modal information.

MISA [11]: MISA (2020) leverages utterance-level representations, each modality is projected into two distinct subspaces, contributing to input reconstruction and employed in task prediction through fusion.

MAG-BERT [24]: MAG-BERT (2020) designs adaptation gates to enhance the processing of information from diverse modalities by BERT and XLNET.

Self-MM [12]: Self-MM (2021) extracts similarity information from multimodal tasks and dissimilarity information from unimodal tasks through a self-supervised label generator.

MMIM [25]: MMIM (2021) preserves task-relevant information through hierarchical maximization of mutual information.

BIMHA [48]: BIMHA (2022) develops a multi-head attention architecture founded on bimodal information orientation to extract separate and coherent information.

TETFN [17]: TETFN (2023) employs text-enhanced cross-modal attention to aggregate the semantic information within fused representations.

MTMD [49]: MTMD(2023) implements unimodal and multimodal momentum distillation for modal interaction.

HCIL [50]: HCIL (2024) enhances sentiment representations through four types of interactions.

CRNet [51]: CRNet (2024) promotes multimodal representation using gradient-based methods.

TMBL [52]: TMBL (2024) uses multimodal binding techniques to discover modality similarities.

Table 4 Results on CMU-MOSI, CMU-MOSEI, and UR-FUNNY

Models	CMU-MOSI				CMU-MOSEI				UR-FUNNY	
	MAE↓	Corr↑	ACC-7↑	ACC-2↑	F1↑	MAE↓	Corr↑	ACC-7↑	ACC-2↑	F1↑
TFN (2017) [7]	0.901	0.698	34.9	−80.8	−80.7	0.593	0.700	50.2	−82.5	−82.1
LMF (2018) [46]	0.917	0.695	33.2	−82.5	−82.4	0.623	0.677	48.0	−82.0	−82.1
MFM (2019) [47]	0.877	0.706	35.4	−81.7	−81.6	0.568	0.717	51.3	−84.4	−84.3
MuT (2019) [9]	0.861	0.711	−	81.5/84.1	80.6/83.9	0.580	0.703	−	−82.5	−82.3
ICCN (2020) [16]	0.862	0.714	39.0	−83.0	−83.0	0.565	0.713	51.6	−84.2	−84.2
MISA (2020) [11]	0.804	0.764	−	80.79/82.10	80.77/82.03	0.568	0.724	−	82.59/84.23	82.67/83.97
MAG-BERT (2020) [24]	0.727	0.781	43.62	82.37/84.43	82.50/84.61	0.543	0.755	52.67	82.51/84.82	82.77/84.71
Self-MM (2021) [12]	0.712	0.795	45.79	82.54/84.77	82.68/84.91	0.529	0.767	53.46	82.68/84.96	82.95/84.93
MMIM (2021) [25]	0.700	0.800	46.65	84.14/86.06	84.00/85.98	0.526	0.772	<u>54.24</u>	82.24/85.97	82.66/85.94
BIMHA (2022) [48]	0.925	0.671	36.44	78.57/80.3	78.5/80.03	0.559	0.731	52.11	84.07/83.96	83.35/83.5
TETFN (2023) [17]	0.717	0.800	−	84.05/86.10	83.83/86.07	0.551	0.748	−	84.25/85.18	84.18/85.27
MTMD (2023) [49]	0.705	0.799	<u>47.5</u>	84.0/86.0	83.9/86.0	0.531	0.767	53.7	<u>84.8/86.1</u>	84.9/85.9
HCIL (2024) [50]	0.703	0.810	−	84.25/86.07	84.18/86.01	0.532	0.768	−	82.56/85.97	82.68/85.29
CRNet (2024) [51]	0.712	0.797	47.4	−86.4	−86.4	0.541	0.771	53.8	−86.2	−86.1
TMBL (2024) [52]	0.867	0.762	36.3	81.78/83.84	82.41/84.29	0.545	0.766	52.4	84.23/85.84	84.87/85.92
TeCaFN (ours)	0.684	0.800	47.81	84.99/86.89	84.88/86.84	0.526	0.772	55.39	85.04/86.27	85.04/86.01
Rank	1	2	1	1/1	1/1	1	1	1	1/1	1/2

The highest values are marked in bold type. In F1 scores and ACC-2, the left side of the “/” signifies “non-negative/negative (NN/N),” while the right side signifies “positive/negative (PN).” ↑ denotes improvement with higher values, and ↓ signifies improvement with lower values. “−” indicates unreported metrics in the respective paper. Underlined results indicate the previous state-of-the-art (SOTA) performance

4.4 Basic settings

This section presents the fundamental configuration of the experiment, which encompasses the feature extraction methodology and the hyperparameter configurations.

Feature extraction: To ensure a fair comparison, a commonly used feature extraction method similar to MMIM [25] is utilized. Specifically, the pre-trained BERT is employed as the text feature extractor. Additionally, COVAREP [53] is used to extract speech features important for sentiment analysis, including pitch, formants, spectral features (MFCCs), speech rate, emotional features, and resonance peaks. For the visual modality, OpenFace [54] is selected to extract facial attributes such as head pose estimation and gaze tracking, among others.

Experiment setup: The Adam optimizer is utilized to optimize our model, setting the learning rate at $\{1e-3, 2e-3, 3e-3\}$. Early stopping is applied if model accuracy does not improve for 10 epochs. The hyperparameter ranges are as follows: batch size in $\{32, 64, 128\}$, γ_1, γ_2 in $\{0.1, 0.2, 0.4\}$, gradient clipping in $\{1, 2, 4, 8\}$ and unimodal hidden size in $\{16, 32, 64\}$. All experiments are conducted on an NVIDIA Tesla V100 (32 G GPU).

4.5 Results and comparison

Table 4 displays the performance of our proposed TeCaFN model on both the CMU-MOSI and CMU-MOSEI datasets, along with a comparison to other baseline models.

Results on CMU-MOSI. TeCaFN is optimal in most accuracies. Specifically, there is a huge improvement over the traditional tensor-based multimodal fusion approach, such as TFN [7], LMF [46], MFM [8] in all metrics. The MAE decreases from 0.901 to 0.684, accompanied by a 4% improvement in classification accuracy. Additionally, the experimental outcomes surpass those achieved by MulT [9], MAG-BERT [24], and BIMHA [48], which employ attentional mechanisms to address unaligned data. TeCaFN increased Corr from 0.781 to 0.800 and ACC-7 from 43.62 to 47.81. ICCN [16], MISA [11], and Self-MM [12] use different methods to fuse multimodal information, compared to these methods, TeCaFN outperforms them in all metrics. Compared to MMIM, TeCaFN emphasizes using textual modality cues to guide other modalities, resulting in substantial enhancements in MAE, ACC-2, and F1 scores. TETFN [17] also uses the textual modality as the core modality for cross-modal attention. However, TeCaFN outperforms all metrics except Corr because of the different auxiliary tasks. In addition, MTMD [49], HCIL [50], CRNet [51] and TMBL [52] use different techniques to promote multimodal representation learning and multimodal interaction. Unlike previous work, TeCaFN uses adversarial learning to learn modality-specific features by providing alignment before cross-modal attention and reducing its learning of irrelevant information. Therefore, compared to recent state-of-the-art, TeCaFN outperforms MTMD by 0.31% on ACC-7, exceeds MMIM by 0.016 on MAE, surpasses HCIL [50] and CRNet

Table 5 Results of the ablation study on CMU-MOSI and CMU-MOSEI

Models	CMU-MOSI					CMU-MOSEI				
	MAE↓	Corr↑	ACC-7↑	ACC-2↑	F1↑	MAE↓	Corr↑	ACC-7↑	ACC-2↑	F1↑
TeCaFN	0.684	0.800	47.81	84.99/86.89	84.88/86.84	0.526	0.772	55.39	85.04/86.27	85.04/86.01
\mathcal{L}_{con}										
w/o \mathcal{L}_{ta}	0.721	0.784	46.06	84.11/85.21	84.09/85.23	0.540	0.763	53.42	84.50/85.83	84.53/85.57
w/o \mathcal{L}_{tv}	0.734	0.794	43.15	82.65/84.60	82.54/84.56	0.538	0.759	53.51	81.00/85.39	81.51/85.33
w/o $\mathcal{L}_{ta} + \mathcal{L}_{tv}$	0.712	0.796	46.79	83.09/85.37	82.96/85.31	0.539	0.761	53.34	81.15/85.11	81.95/85.05
w/o \mathcal{L}_{bicon}	0.719	0.795	45.77	81.78/83.38	81.70/83.36	0.531	0.769	53.87	83.77/86.24	84.04/86.11
\mathcal{L}_{ap}										
w/o \mathcal{L}_{tap}	0.740	0.787	44.17	82.94/85.21	82.82/85.16	0.534	0.766	53.94	84.65/85.94	84.66/85.67
w/o \mathcal{L}_{tvp}	0.742	0.786	43.88	82.22/84.60	82.09/84.56	0.534	0.768	54.26	80.51/85.25	81.15/85.29
w/o $\mathcal{L}_{tap+tvp}$	0.749	0.775	45.04	82.80/84.76	82.61/84.63	0.538	0.773	52.78	83.73/85.37	83.70/85.03
<i>Hard negative</i>										
w/o $Hardneg_{ta}$	0.715	0.793	46.50	83.38/85.37	83.23/85.28	0.529	0.774	54.32	81.03/85.39	81.61/85.41
w/o $Hardneg_{tv}$	0.715	0.797	43.00	82.22/84.60	82.07/84.54	0.535	0.769	52.78	83.11/ 86.63	83.50/ 86.58
w/o $Hardneg_{ta,tv}$	0.729	0.793	42.71	82.51/84.76	82.35/84.68	0.530	0.764	54.56	82.94/85.86	83.25/85.75
<i>Multimodal</i>										
w/o V	0.780	0.760	45.92	80.90/83.69	80.40/83.34	0.565	0.754	52.11	84.01/84.15	83.73/83.62
w/o A	0.794	0.762	41.83	81.63/83.99	81.43/83.88	0.559	0.751	52.07	79.74/85.00	80.40/85.01
w/o V, A	0.878	0.763	34.99	81.05/83.08	80.99/83.09	0.613	0.700	50.48	70.98/79.09	72.32/79.44

The bold numbers represent the best results for the corresponding metrics

[51] and 0.4–0.8% on ACC-2 and F1-score. It also performs well in other metrics.

Results and comparison on CMU-MOSEI. TeCaFN achieves the best performance in most accuracy categories. The CMU-MOSEI dataset contains more complex and diverse data, than most traditional methods, such as those based on tensor fusion [7, 8, 46], traditional attentional mechanisms [9, 24, 48], or simple auxiliary tasks [11, 11, 12] do not perform well on CMU-MOSEI. TeCaFN also outperforms them in all metrics. Even though TeCaFN does not use a large encoder, its performance on the CMU-MOSEI dataset still far exceeds that of MMIM and TETFN. Thanks to adversarial training, compared with MTMD [49], HCIL [50], CRNet [51] and TMBL [52], ACC-2 and F1 improve by an average of 0.1–0.2%, Corr improves by 0.005, ACC-7 improves by 0.89%, implying that inter-modal alignment followed by fusion can aggregate multimodal information effectively.

Results and comparison on UR-FUNNY. TeCaFN has achieved a new state-of-the-art (SOTA) performance on the UR-FUNNY dataset, marking a 0.18% improvement over the previous SOTA value. Despite the UR-FUNNY dataset only annotated with binary classification labels instead of fine-grained sentimental scores, TeCaFN is still capable of making accurate predictions through its adversarial training process and reliance on modality-specific information.

In summary, TeCaFN can learn good feature representations on CMU-MOSI, CMU-MOSEI, and UR-FUNNY through the guiding role of textual modality, cross-modal attention, and modality-specific information. It achieves more significant improvements in the vast majority of metrics with no obvious drawbacks.

4.6 Ablation study

TeCaFN consists of three main modules: text-centered unimodal contrastive learning, text-centered cross-sample fusion, and fused modal bidirectional contrastive learning.

We disassemble the structure of TeCaFN to confirm the validity of our proposed modules and methods on both the CMU-MOSI and CMU-MOSEI datasets. The result of the ablation experiments can be found in Table 5.

The effect of text-based unimodal contrastive learning.

We sequentially remove audio-to-text contrastive learning (denoted as w/o \mathcal{L}_{ta}), and vision-to-text contrastive learning (denoted as w/o \mathcal{L}_{tv}) and both (denoted as w/o $\mathcal{L}_{ta} + \mathcal{L}_{tv}$) to validate the effect of the unimodal text-based contrastive learning module. No matter which part of the contrastive learning loss is removed, the results do not improve. In addition to this, when aligning only two modalities and ignore the third, the effect may not be as good as if we do not implement any alignment at all. However, this drop is not caused by the insignificant feature of the third modality but by a bias in alignment. When we only bring two modalities closer together while keeping the third unchanged, then according to the distance measure of cross-modal attention, the attention of the third modality will decrease. When we only remove text and visual modality contrastive learning, the result drops the most, even more than the no contrastive learning, with an average 3% drop in accuracy on the classification task, implying that unimodal contrastive learning of text and vision is more important. Text and audio are inherently more similar regarding information, so removing the unimodal comparison of text and audio from the experiments does not cause a distinct drop in experimental results. The factual results show that the text-centered unimodal contrastive learning to embed different modalities into a unified space significantly impacts the extraction of modality-specific information and multimodal fusion.

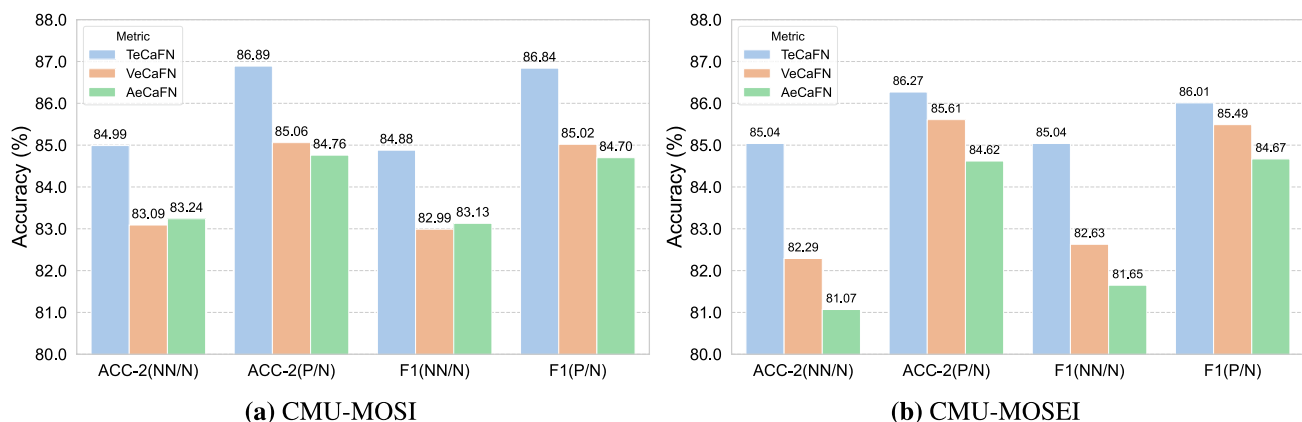


Fig. 3 Comparison of models centered on different modalities

The effect of text-centered cross-sample fusion. We successively remove the replacement gate for audio modalities (denoted by \mathcal{L}_{tap}) and visual modalities (denoted as \mathcal{L}_{rvp}) and both (denoted as $\mathcal{L}_{tap+rvp}$) to demonstrate how the modality replacement contributes to improving model performance. We find the modality replacement gate is very significant to our model. The ACC-2 drops by 2–3%, and the ACC-7 drops by 2–3% on average on both datasets without it. Besides, when implementing the modality replacement for only two modalities and ignore the third modality, the conclusion is consistent with the previous one. The results are not even as good as when we do not perform modality replacement. Because modal replacement is an adversarial task that affects the ability of the entire network to perceive modalities. In addition, visual replacement plays a more significant role than audio replacement. The results show that the text-centered cross-sample fusion helps improve task accuracy.

The effect of fused modal bidirectional contrastive learning. We remove the fused modal bidirectional contrastive learning after cross-modal attention (denoted as \mathcal{L}_{bicon}) in this section. The second stage of contrastive learning after fusing the modalities equally improves the model performance on both datasets, especially on CMU-MOSI, because it is a smaller dataset, making it more challenging to find suitable negative samples for cross-sample fusion. The adversarial training can easily disrupt the judgments of the sentiment, so the re-alignment of fused modal is necessary for the entire learning process. This also confirms the importance of fused modal bidirectional contrastive learning in overcoming noise and instability in cross-sample fusion.

The effect of hard negatives. In this case, we do not mine similar samples for cross-modal attention via the similarity matrix of ta (denoted as w/o $Hardneg_{ta}$), the similarity matrix of tv (denoted as w/o $Hardneg_{tv}$), or neither (denoted as w/o $Hardneg_{ta,tv}$). Results show that the hard negative operation would also affect the result of the experiment. When we do not use the hard negatives, a cross-sample fusion of different samples interferes with the normal progress of training, and this interference cannot be fully serviced even with the textual modality not being replaced. In addition, the hard negative allows the model to capture more fine-grained features, which is also essential for multimodal sentiment analysis tasks. So hard negative not only makes training more stable but also makes model training more efficient.

Table 6 Effect of p

p	CMU-MOSI		CMU-MOSEI	
	MAE↓	ACC-2↑	MAE↓	ACC-2↑
0.1	0.706	83.97/85.98	0.522	81.37/85.69
0.2	0.740	82.36/83.84	0.524	84.42/85.94
0.3	0.732	83.24/85.52	0.529	84.31/85.53
0.4	0.708	83.53/85.06	0.526	85.04/86.27
0.5	0.704	84.69/ 87.04	0.537	79.61/84.65
0.6	0.757	81.34/84.15	0.538	82.31/85.99
0.7	0.684	84.99 /86.89	0.565	84.04/84.15
0.8	0.769	81.92/83.84	0.529	80.21/85.31
0.9	0.771	81.92/84.30	0.537	83.34/85.97

The bold numbers represent the best results for the corresponding metrics

Fig. 4 Loss tracing

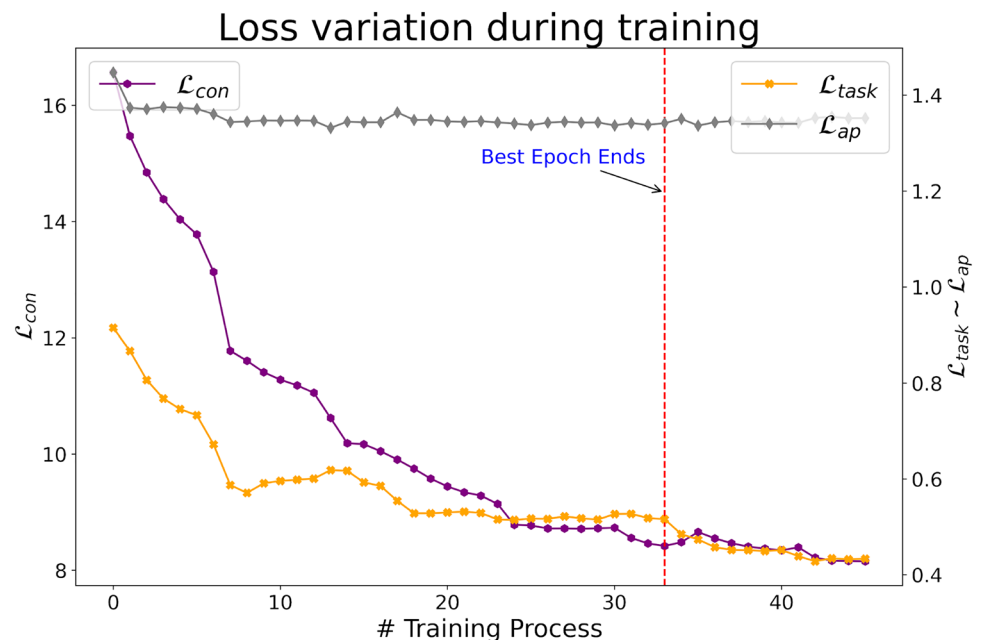


Table 7 Effect of num

num	CMU-MOSI		CMU-MOSEI	
	MAE↓	ACC-2↑	MAE↓	ACC-2↑
3	0.684	84.99/86.89	0.526	85.04/86.27
5	0.747	81.05/82.93	0.526	83.26/ 86.30
10	0.735	82.65/84.76	0.525	83.04/85.75
15	0.693	83.97/85.82	0.536	85.19/85.31
30	0.720	84.55/86.43	0.527	83.54/85.77

















The bold numbers represent the best results for the corresponding metrics

The effect of the combination of modalities. We sequentially remove visual modality (denoted as w/o V), audio modality (denoted as w/o A), and both (denoted as w/o V, A) to validate the effect of the combination of modalities. Removing either the visual or acoustic modality, or both, consistently leads to a decline in performance. This underscores the necessity of non-verbal signals (visual and acoustic) in addressing MSA, illustrating how text, acoustic, and visual elements complement each other. This also indicates the significance of modality-specific information. As more modalities are integrated, the perception of sentiments becomes more accurate.

The effect of treating text as a core modality. All previous ablation experiments are based on TeCaFN, but do other modalities have the same effectiveness as text modality? To answer this question, we design the audio-centered AeCaFN and the vision-centered VeCaFN to be compared with TeCaFN. It is worth mentioning that AeCaFN and VeCaFN are identical to TeCaFN in terms of architecture and training methodology, except for the core modality. Figure 3 displays the experimental outcomes.

Results show that neither AeCaFN nor VeCaFN is as effective as TeCaFN on both datasets. The results of AeCaFN and VeCaFN are similar to the CMU-MOSI dataset. In addition, they are 1.7–2.3% lower than TeCaFN in classification accuracies. On the CMU-MOSEI dataset, VeCaFN outperforms AeCaFN by about 1%. However, they are 3–4% lower than TeCaFN in the discrimination of the non-negative and negative problem and 1–2% lower than TeCaFN in the discrimination of the positive/negative problem. This means that the textual modality has the most vital ability to align and fuse the other modalities among the textual, speech, and visual modalities, which is common sense and validates our previous hypothesis. Vision and audio have similar abilities to align and fuse other modalities on small datasets. However, vision tends to align and fuse data more effectively on complex datasets.

Table 8 Comparison of the output of TeCaFN with other MSA models

High-similarity samples			MMIM [25]		TeFNA [18]		TeCaFN(ours)		Ground Truth
			\hat{y}	$ \hat{y} - y \downarrow$	\hat{y}	$ \hat{y} - y \downarrow$	\hat{y}	$ \hat{y} - y \downarrow$	
	His fight sequences are very neat		2.1950	0.1950	2.1964	0.1964	2.0681	0.0681	2.0000
	And he delivers a lot of intensity		1.3000	0.3000	1.4707	0.1293	1.6289	0.0289	1.6000
	Sound of cars the action and humor were mostly flat		-1.5259	0.2741	-1.6370	0.1630	-1.7389	0.0611	-1.8000
	But but I just did didnt laugh		-2.0840	0.0840	-2.2743	0.2743	-1.9874	0.0126	-2.0000
	It's rated PG-thirteen, (uhh) and again starring ...		-0.2220	0.2220	-0.4247	0.4247	-0.0479	0.0479	0.0000
	(uhh){clears throat} And so we keep seeing a lot of...		-0.1284	0.2049	-0.4895	0.1562	-0.3061	0.0272	-0.3333
	Now, that employee that sold you that bag of popcorn ...		0.5360	0.1307	0.7581	0.0911	0.6910	0.0243	0.6667
	It is a belief that an organization operates behind ...		0.3042	0.0291	0.4553	0.1220	0.3289	0.0044	0.3333

The bold numbers represent the best results for the corresponding metrics

4.7 Loss tracing

To better demonstrate how the multitasks contribute to each other, we trace each component of the loss function. Figure 4 illustrates the results of loss tracing.

The average loss is plotted at fixed intervals every 15 training steps, with specific labeling applied to each of the proposed losses. It is observed that the contrastive learning loss \mathcal{L}_{con} and the task loss \mathcal{L}_{task} consistently decrease for the most part of the training process. The pairwise prediction loss \mathcal{L}_{ap} remains relatively stable throughout the training process, showing only a minor decrease. All sub-losses can be optimized jointly, achieving the optimal value under unified supervision.

4.8 Parameter sensitivity

We set up a series of experiments to examine the sensitivity of the parameters. Two important hyperparameters are involved in our proposed module: p denotes the probability that the modality is replaced with hard negatives in the replacement gate, and num denotes that there are **num** hard negative samples in the replacement gate.

Table 6 depicts the effect of different replacement probabilities on comparative experiments (0.1–0.9, steps of 0.1). The optimal value of p for the CMU-MOSI dataset is 0.7, whereas for the CMU-MOSEI dataset, it is 0.4. Furthermore, the model performance varies based on the count of adversarial attacks. Appropriate adversarial attacks can help the model learn modal features better. The results suggest that our model can learn multimodal representations from adversarial training.

Table 7 illustrates the outcomes of comparative experiments regarding the sensitivity to the count of hard negatives. The model attains optimal performance on both datasets when num is set to 3. However, the accuracy decreases with an increase in hard negatives, affecting both MAE and ACC-2 adversely. This means that the selection criteria for hard negatives are crucial. This criterion impacts the granularity of modality-specific information learned by the model. Besides, these negative samples should be sufficiently similar to positive ones. Otherwise, the training results will become ineffective.

4.9 Case analysis

To demonstrate the role of modality-specific features for fine-grained sentiment recognition, we compare our model with MMIM [25] and TeFNA [18]. For a fair comparison, we randomly choose four pairs of similar samples distributed in different sentimental intervals. The pairs of similar samples are from different clips of the same video and the absolute value of the ground truth of the similar samples is < 0.4 .

In Table 8, the first column displays similar pairs, while the subsequent columns represent the outputs of respective models and ground truth. We use mean absolute error (MAE) to evaluate the model performance, where \hat{y} denotes the model output and $|\hat{y} - y|$ denotes the absolute error with ground truth which is the smaller, the better. Optimal values are marked in bold.

As expected, in each sample, the text provided coarse-grained information about the sentiment, while the other modalities provided modality-specific complemented information that made the model predictions more accurate. Case 1 and case 2 illustrate how the combined use of adverbs, adjectives, pauses in speech, tone of voice, and facial expressions collectively influence their sentiment score. Case 3 and case 4 show that when emotions tend to be neutral, we can still find clues in the modalities, such as the change of speech speed and micro-expressions to perceive their sentiments. Results show that our model can predict sentimental scores most accurately in different intervals and perceive more fine-grained sentimental information.

5 Conclusion

Conventional models that rely on cross-modal attention often face challenges in extracting modality-specific information, especially when it comes to differentiating between similar samples. This is primarily due to the inherent biases present in the learning process. To overcome these limitations, we introduce TeCaFN, which employs a unique cross-sample fusion technique that amalgamates modalities from separate samples. This innovative approach preserves intricate modality-specific details by leveraging adversarial training in conjunction with a pairwise prediction task. Moreover, we have devised a robust two-stage mechanism centered around text-centric contrastive learning. This mechanism significantly enhances the stability of the learning process involved in cross-sample fusion. Through extensive experiments, TeCaFN has demonstrated superior performance over existing methods in various multimodal sentiment analysis tasks. The findings assess the efficacy of modality-specific information in multimodal sentiment analysis and underscore the importance of prioritizing text as a primary modality. Future work can be carried out based on efficient mining of negative samples or designing auxiliary tasks more adapted to adversarial learning.

Author contributions Q.H. completed conceptualization, methodology, software, validation, writing and editing. J.C. completed conceptualization, methodology, formal analysis, software, validation, visualization and writing the original draft. C.H. completed validation, visualization, writing and editing. Y.W. completed validation, visualization, writing and editing. X.H. completed methodology,

validation, visualization, writing and editing. All authors reviewed the manuscript.

Funding This work was supported by the Pioneer “and Leading Goose” R&D Program of Zhejiang (No. 2022C03106), the National Natural Science Foundation of China (No. 62207028, 62337001), partially by Zhejiang Provincial Natural Science Foundation (No. LY23F020009), and Scientific Research Fund of Zhejiang Provincial Education Department (No. 2023SCG367), and Special Research Project of Zhejiang Normal University on Serving Provincial Strategic Planning and Promoting the Construction of Common Prosperity.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

- Shenoy, A., Sardana, A.: Multilogue-net: a context aware rnn for multi-modal emotion detection and sentiment analysis in conversation (2020). arXiv preprint [arXiv:2002.08267](https://arxiv.org/abs/2002.08267)
- Fu, J., Mao, Q., Tu, J., Zhan, Y.: Multimodal shared features learning for emotion recognition by enhanced sparse local discriminative canonical correlation analysis. *Multimed. Syst.* **25**(5), 451–461 (2019)
- Huang, Q., Huang, C., Wang, X., Jiang, F.: Facial expression recognition with grid-wise attention and visual transformer. *Inf. Sci.* **580**, 35–54 (2021)
- Luo, Y., Wu, R., Liu, J., Tang, X.: Balanced sentimental information via multimodal interaction model. *Multimed. Syst.* **30**(1), 1–9 (2024)
- Das, R., Singh, T.D.: Multimodal sentiment analysis: a survey of methods, trends, and challenges. *ACM Comput. Surv.* **55**(13s), 1–38 (2023)
- Shaik, T., Tao, X., Li, L., Xie, H., Velásquez, J.D.: A survey of multimodal information fusion for smart healthcare: mapping the journey from data to wisdom. *Inf Fusion* **102**, 102040 (2023)
- Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.-P.: Tensor fusion network for multimodal sentiment analysis. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1103–1114 (2017)
- Zadeh, A., Liang, P.P., Mazumder, N., Poria, S., Cambria, E., Morency, L.-P.: Memory fusion network for multi-view sequential learning. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, pp. 5634–5641 (2018)
- Tsai, Y.-H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.-P., Salakhutdinov, R.: Multimodal transformer for unaligned multimodal language sequences. In: *Proceedings of the Conference. Association for Computational Linguistics. Meeting*, vol. 2019, p. 6558. NIH Public Access (2019)
- Gan, Z., Chen, Y.-C., Li, L., Zhu, C., Cheng, Y., Liu, J.: Large-scale adversarial training for vision-and-language representation learning. *Adv. Neural Inf. Process. Syst.* **33**, 6616–6628 (2020)
- 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*, pp. 1122–1131 (2020)
- Yu, W., Xu, H., Yuan, Z., Wu, J.: Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 10790–10797 (2021)
- Chen, Q., Huang, G., Wang, Y.: The weighted cross-modal attention mechanism with sentiment prediction auxiliary task for multimodal sentiment analysis. *IEEE/ACM Trans. Audio Speech Lang. Process.* **30**, 2689–2695 (2022)
- Wang, D., Liu, S., Wang, Q., Tian, Y., He, L., Gao, X.: Cross-modal enhancement network for multimodal sentiment analysis. *IEEE Trans. Multimed.* **25**, 4909–4921 (2022)
- Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., Hussain, A.: Multimodal sentiment analysis: a systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Inf. Fusion* **91**, 424–444 (2023)
- Sun, Z., Sarma, P., Sethares, W., Liang, Y.: Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 8992–8999 (2020)
- Wang, D., Guo, X., Tian, Y., Liu, J., He, L., Luo, X.: Tetfn: a text enhanced transformer fusion network for multimodal sentiment analysis. *Pattern Recognit.* **136**, 109259 (2023)
- 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.* **269**, 110502 (2023)
- Luo, Y., Wu, R., Liu, J., Tang, X.: A text guided multi-task learning network for multimodal sentiment analysis. *Neurocomputing* **560**, 126836 (2023)
- Wei, X., Zhang, T., Li, Y., Zhang, Y., Wu, F.: Multi-modality cross attention network for image and sentence matching. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10941–10950 (2020)
- Xu, X., Wang, T., Yang, Y., Zuo, L., Shen, F., Shen, H.T.: Cross-modal attention with semantic consistence for image-text matching. *IEEE Trans. Neural Netw. Learn. Syst.* **31**(12), 5412–5425 (2020)
- Chen, Z., Zhang, C., Zhang, B., He, Y.: Triplet contrastive learning framework with adversarial hard-negative sample generation for multimodal remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **62**, 3354304 (2024)
- Majumder, N., Hazarika, D., Gelbukh, A., Cambria, E., Poria, S.: Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowl.-Based Syst.* **161**, 124–133 (2018)
- Rahman, W., Hasan, M.K., Lee, S., Zadeh, A., Mao, C., Morency, L.-P., Hoque, E.: Integrating multimodal information in large pre-trained transformers. In: *Proceedings of the Conference. Association for Computational Linguistics. Meeting*, vol. 2020, p. 2359. NIH Public Access (2020)
- 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*, pp. 9180–9192 (2021)
- Kim, K., Park, S.: Aobert: all-modalities-in-one BERT for multimodal sentiment analysis. *Inf. Fusion* **92**, 37–45 (2023)
- Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: vision and language representation learning with momentum distillation. *Adv. Neural Inf. Process. Syst.* **34**, 9694–9705 (2021)
- Peng, W., Hong, X., Zhao, G.: Adaptive modality distillation for separable multimodal sentiment analysis. *IEEE Intell. Syst.* **36**(3), 82–89 (2021)
- Zhang, J., Wu, X., Huang, C.: Adamow: multimodal sentiment analysis based on adaptive modality-specific weight fusion network. *IEEE Access* **11**, 48410–48420 (2023)
- Ando, A., Masumura, R., Takashima, A., Suzuki, S., Makishima, N., Suzuki, K., Moriya, T., Ashihara, T., Sato, H.: On the use of

- modality-specific large-scale pre-trained encoders for multimodal sentiment analysis. In: 2022 IEEE Spoken Language Technology Workshop (SLT), pp. 739–746. IEEE (2023)
31. Yang, B., Shao, B., Wu, L., Lin, X.: Multimodal sentiment analysis with unidirectional modality translation. *Neurocomputing* **467**, 130–137 (2022)
 32. 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.* **29**, 1424–1437 (2021)
 33. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: transformers for image recognition at scale (2020). arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929)
 34. Kenton, J.D.M.-W.C., Toutanova, L.K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, pp. 4171–4186 (2019)
 35. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
 36. Dixit, C., Satapathy, S.M.: A customizable framework for multimodal emotion recognition using ensemble of deep neural network models. *Multimed. Syst.* **29**(6), 3151–3168 (2023)
 37. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: a robustly optimized BERT pretraining approach. arXiv e-prints, 1907 (2019)
 38. 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: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 4730–4738 (2021)
 39. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 5998–6008 (2017)
 40. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding (2018). arXiv preprint [arXiv:1807.03748](https://arxiv.org/abs/1807.03748)
 41. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3733–3742 (2018)
 42. Huang, J., Pu, Y., Zhou, D., Cao, J., Gu, J., Zhao, Z., Xu, D.: Dynamic hypergraph convolutional network for multimodal sentiment analysis. *Neurocomputing* **565**, 126992 (2024)
 43. Zadeh, A., Zellers, R., Pincus, E., Morency, L.-P.: Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intell. Syst.* **31**(6), 82–88 (2016)
 44. Zadeh, A.B., Liang, P.P., 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), pp. 2236–2246 (2018)
 45. Hasan, M.K., Rahman, W., Zadeh, A., Zhong, J., Tanveer, M.I., Morency, L.-P., et al.: Ur-funny: a multimodal language dataset for understanding humor (2019). arXiv preprint [arXiv:1904.06618](https://arxiv.org/abs/1904.06618)
 46. Liu, Z., Shen, Y., Lakshminarasimhan, V.B., Liang, P.P., Zadeh, A.B., Morency, L.-P.: Efficient low-rank multimodal fusion with modality-specific factors. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2247–2256 (2018)
 47. Tsai, Y.-H.H., Liang, P.P., Zadeh, A., Morency, L.-P., Salakhutdinov, R.: Learning factorized multimodal representations. In: International Conference on Representation Learning, pp. 1–20 (2019)
 48. Wu, T., Peng, J., Zhang, W., Zhang, H., Tan, S., Yi, F., Ma, C., Huang, Y.: Video sentiment analysis with bimodal information-augmented multi-head attention. *Knowl.-Based Syst.* **235**, 107676 (2022)
 49. Lin, R., Hu, H.: Multi-task momentum distillation for multimodal sentiment analysis. *IEEE Trans. Affect. Comput.* **15**, 549–565 (2023)
 50. Fu, Y., Zhang, Z., Yang, R., Yao, C.: Hybrid cross-modal interaction learning for multimodal sentiment analysis. *Neurocomputing* **571**, 127201 (2024)
 51. Shi, H., Pu, Y., Zhao, Z., Huang, J., Zhou, D., Xu, D., Cao, J.: Co-space representation interaction network for multimodal sentiment analysis. *Knowl.-Based Syst.* **283**, 111149 (2024)
 52. Huang, J., Zhou, J., Tang, Z., Lin, J., Chen, C.Y.-C.: Tmbl: transformer-based multimodal binding learning model for multimodal sentiment analysis. *Knowl.-Based Syst.* **285**, 111346 (2024)
 53. Degottex, G., Kane, J., Drugman, T., Raitio, T., Scherer, S.: COVAREP—a collaborative voice analysis repository for speech technologies. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 960–964. IEEE (2014)
 54. Baltrušaitis, T., Robinson, P., Morency, L.-P.: Openface: an open source facial behavior analysis toolkit. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–10. IEEE (2016)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.