Complex & Intelligent Systems (2026) 12:193
https://doi.org/10.1007/s40747-026-02318-z

ORIGINAL ARTICLE

Enhancing emotion recognition through three modalities fusion and contrastive learning with the Mamba architecture

Qianjun Shuai[1] · Xiaohao Chen[1] · Feng Hu[1] · Yongqiang Cheng[2]

Received: 19 September 2024 / Accepted: 31 March 2026

Abstract

Multimodal emotion recognition extracts emotional information from sequential multimodal data and classifies emotion tendencies. Current multimodal fusion methods based on artificial intelligence mainly rely on Transformers to extract features and integrate different data types. Despite their strength in learning global information, Transformers face challenges due to their quadratic complexity. Recent advances in state space models, especially the Mamba architecture, provide a promising solution by achieving global awareness with linear complexity. However, the potential of Mamba for information fusion in multimodal domains remains largely untapped. This paper introduces an innovative and efficient multimodal fusion and contrastive learning method called Fusion Mamba and Contrastive Learning, which leverages artificial intelligence for implementation and application in emotion recognition tasks. To effectively extract distinct features, the unimodal Mamba architecture is used to enhance unimodal representations. For comprehensive information fusion, the Mamba block is extended to handle dual inputs, forming a novel module called the Fusion Mamba block. This forms the basis for an architecture that incorporates three different modalities and three branches. Additionally, contrastive learning and interaction-level auxiliary classification constraints are jointly optimized to boost performance. The effectiveness of our approach, which highlights the application of artificial intelligence, is validated through experiments on three public datasets. Both quantitative and qualitative evaluations show that our method achieves state-of-the-art performance with 32.2% faster inference. Extensive ablation studies further confirm the effectiveness of the Mamba architecture in multimodal tasks.

Keywords Artificial intelligence · Emotion eecognition · Mamba · Contrastive learning · Multimodal fusion

Introduction

In recent times, the significance of human–computer interaction has grown, leading to extensive research in emotion recognition [1]. Given the variety of information, multimodal data (such as human utterances with linguistic, auditory, and visual elements) have become key resources for systems to comprehend human emotions [2]. This research area primarily aims at improving and integrating sequential emotional data from different modalities. Earlier studies mainly focused on creating sophisticated models to extract and merge emotional features from three distinct types of data (linguistic, acoustic, and visual) [3]. Lately, some attention-based [4, 5] and Transformer-based methods [6, 7] have been introduced to handle modality interactions and extract multimodal features. These methods have demonstrated improved performance compared to traditional frameworks such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). While models based on Transformers or their combinations with other architectures excel in global modeling, they incur significant computational costs due to the quadratic growth in resource requirements caused by the self-attention mechanism's token count [8].

Recent advancements in state space models (SSMs), especially Mamba [9], offer an effective solution by achieving global awareness with linear complexity. Mamba has

✉ Xiaohao Chen
xiaohaochen@cuc.edu.cn

Qianjun Shuai
sqj@cuc.edu.cn

Feng Hu
fenghu@cuc.edu.cn

Yongqiang Cheng
yongqiang.cheng@sunderland.ac.uk

[1] School of Information and Communication Engineering, Communication University of China, Beijing 100024, China

[2] Faculty of Technology, Sunderland SR6 0DD, UK

Published online: 05 May 2026

demonstrated its efficacy in tasks that need long-term dependency modeling, such as natural language processing, due to its adaptable input and global information modeling capabilities. This approach maintains linear complexity, reducing computational costs and enhancing inference speed [10]. Recently, various Mamba variants have also achieved promising outcomes in computer vision tasks, including image classification [11] and medical image fusion [12]. Simultaneously, in the visual domain, some multimodal image fusion based on the Mamba architecture has also achieved success [13].

Inspired by Mamba's success in different modalities [14, 15], we were motivated to apply Mamba to the emotion recognition task across tri-modal visual, textual, and auditory data. At present, there is limited research on the potential of SSM to integrate these three modalities, which is vital for the multimodal field.

Therefore, we propose a MER model based on Mamba. It consists of three stages: Intra-Mamba feature extraction, Inter-Mamba modality interaction and fusion, and a tri-modal contrastive learning module. The Intra-Mamba feature extraction stage comprises convolutional layers and multiple stacked Mamba blocks, leveraging CNN's excellent early-stage processing capabilities for visual tasks and Mamba's efficiency in extracting long-distance features. In the Inter-Mamba modality interaction and fusion stage, inspired by previous cross-attention mechanisms, we design a dual-modality Mamba fusion module (FM). Since FM is designed for interactions between two modalities. We suggest a parallel structure with three branches (TFM) for three modality pairs: ta, tv, and av, where t, a, and v stand for textual, auditory, and visual modalities, respectively. This parallel structure allows our method to extract inter-modal features while considering all relevant modalities. Finally, we introduce an unsupervised multimodal contrastive learning method to ensure that each parallel TFM is effectively trained.

The contribution of this work can be summarized as follows:

- To the best of our knowledge, our approach is among the pioneering works utilizing Mamba for tri-modal emotion recognition, demonstrating efficiency and effectiveness compared to traditional Transformers;
- We designed a parallel structure incorporating a tri-branch Mamba fusion module to effectively integrate information from different modalities. This parallel structure allows for unified modality interactions while extracting inter-modal features;
- We introduced a multimodal contrastive learning module to alleviate the modality gap in the outputs of the parallel structure. This module leverages contrastive learning-based modality-invariant features;
- We demonstrated our approach's effectiveness using three public datasets, achieving sota performance in multimodal emotion recognition (MER). Furthermore, extensive ablation studies affirm the Mamba architecture's efficacy in multimodal tasks.

Related work

In this section, we briefly review previous related works about multimodal emotion recognition, state space model, and contrastive learning.

Multimodal emotion recognition

MER aims to infer human emotions by analyzing information from three modalities present in videos: language, visual, and auditory. There are three main approaches to MER: fusion-based methods [16–18], decoupling-based methods [19–22], and cross-modal attention methods [23–25].

Fusion-based methods develop advanced strategies for multimodal integration. Recent approaches highlight that integrating different modalities is crucial in MER. One effective method is the unrestricted exchange of information between multiple modalities. Zadeh et al. [17] introduced the tensor fusion network (TFN), which progressively aggregates unimodal, bimodal, and trimodal interactions. Zadeh et al. [18] proposed the memory fusion network (MFN), using temporal information for modality interactions in a long-short term memory (LSTM) manner. Liu et al. [16] developed the low-rank multimodal fusion (LMF) method, using weighted low-rank matrix factorization to reduce excessive parameters in fusion models. However, multimodal feature fusion often faces challenges due to inherent heterogeneity and redundant information between different modalities.

An effective way to address the inherent heterogeneity and redundancy between different modalities is through feature decoupling or disentanglement. Decoupling-based strategies promote more effective fusion of multimodal representations. DMD [19] deconstructed the representation of each modality into two parts in an autoregressive manner, namely the modalityirrelevant/-exclusive spaces, reducing redundant information and thereby improving the effectiveness of multimodal fusion. Building on this, Weichen Dai et al. [22] proposed the multimodal information disentanglement (MInD) method to tackle the underutilization of heterogeneous modality information. The primary strategy involves decomposing the features of each modality into three distinct components through information optimization, which together provide a comprehensive view of the given input.

As Transformer networks [26] become more prevalent in natural language processing and computer vision,

researchers have utilized specific self-attention mechanisms to identify correlations between different modalities [23]. Transformers' self-attention mechanism captures more global information than traditional attention methods. This mechanism inherently generates attention maps from two independent streams (queries (Q) and keys (K) in Transformer structures), naturally applying to multimodal data processing. Consequently, most Transformer-based multimodal methods use Q and K from two modalities to compute a joint attention map for information exchange. This is known as inter-modal attention. MulT [27] introduced the Multimodal Transformer, focusing on interactions between multimodal sequences across different time steps and aligning streams from one modality to another.

While Transformer-based networks have inherent advantages for multimodal processing, their quadratic time complexity and computational resource demands associated with self-attention mechanisms make them inefficient for multimodal tasks and less scalable for tri-modal data. Specifically, cross-attention mechanisms can compute correlations for at most two modalities simultaneously. In this study, to utilize all modalities, we created a parallel structure with three branches, each designed to transfer information between two modalities.

State space model

The SSM concept was first presented in the S4 [28] model, offering a distinctive architecture that effectively models global information, surpassing traditional convolutional neural networks and Transformers. Building on S4, the S5 [29] model reduced the complexity to a linear level, and H3 [30] applied it to language modeling tasks. Mamba enhanced the SSM with an input activation mechanism, resulting in faster inference and better performance compared to similarly sized Transformers. Initially showing great promise in NLP tasks, Mamba later attracted significant attention in computer vision. Notable contributions include Vision Mamba [31], VMamba [32], and Pan-Mamba [33].

Recently, some research has extended Mamba to the visual multimodal domain. FusionMamba [8] integrated Mamba blocks into two U-shaped networks: Spatial U-Net and Spectral U-Net, achieving efficient image fusion. CFMW [34] proposed a cross-modal Mamba with weather removal capability, enhancing detection accuracy under adverse weather conditions. Audio Mamba [35] employed a self-attention-free, fully SSM-based audio classification model, achieving comparable or better performance compared to Transformers. Coupled Mamba [36] applies Mamba modules to multimodal emotion analysis by integrating visual, audio, and textual features. This method adopts a direct multimodal fusion strategy, ensuring tight information integration. However, during the fusion process, differences between modalities may affect the overall fusion performance. To address this, we propose a phased interaction strategy, where pairwise modality fusion (V-T, V-A, T-A) is first conducted to better align information from different modalities, followed by a higher-level three-modal fusion to enhance the balance and effectiveness of multimodal integration.

Contrastive learning

Contrastive learning aims to bring similar instances closer and push dissimilar ones apart in the feature space. In recent years, contrastive learning has shown remarkable success in both computer vision and natural language processing fields. A prominent example is CLIP [37] (Contrastive Language-Image Pre-training), which trains visual and textual representations jointly to achieve superior multimodal understanding. Another example is LanguageBind [38], which employs contrastive learning to align language representations with emotion information, achieving significant improvements.

Akbari et al. [39] proposed a framework for cross-modal contrastive learning, using a transformer-based architecture to enhance the learning of emotion features across multiple modalities. Inspired by the prior advances in multimodal contrastive learning, this paper employs contrastive learning as a key component to enhance our MER model.

Therefore, leveraging previous developments in multimodal contrastive learning, this paper adopts contrastive learning to provide robust feature alignment for MER.

Methodology

Preliminaries

State space model

The SSM is a continuous system that converts a 1D input $x(t) \in \mathbb{R}$ into an output $y(t) \in \mathbb{R}$ through an intermediate hidden state $h(t) \in \mathbb{R}^N$. This can be mathematically described using ordinary differential equations (ODEs):

$$\begin{aligned} h'(t) &= \mathbf{A}h(t) + \mathbf{B}x(t), \\ y(t) &= \mathbf{C}h(t). \end{aligned} \tag{1}$$

In these equations, $A \in \mathbb{R}^{N\times N}$ is the evolution parameter, while $B \in \mathbb{R}^{N\times 1}$ and $C \in \mathbb{R}^{1\times N}$ are projection parameters. This indicates that SSM demonstrates global awareness, as the current output depends on all prior input data.

When A, B, and C are constants, Eq. (1) defines a linear time-invariant (LTI) system, as mentioned in S4 [28]. If they vary, the system becomes a linear time-varying (LTV) system, as described in Mamba [9]. LTI systems cannot inher-

ently perceive input content, whereas LTV systems with input awareness are designed to do so. This significant distinction allows Mamba to overcome SSM limitations.

Discretization

For SSM to be applied in deep learning (DL), it must be discretized. A timescale parameter $\Delta \in \mathbb{R}$ is introduced to convert the continuous parameters A and B into their discrete counterparts, denoted as A and B. Using zero-order hold (ZOH) as the transformation method, the discrete parameters are expressed as follows:

$$\begin{aligned} \bar{\mathbf{A}} &= \exp(\Delta \mathbf{A}), \\ \bar{\mathbf{B}} &= (\Delta \mathbf{A})^{-1}(\exp(\Delta \mathbf{A}) - \mathbf{I}) \cdot \Delta \mathbf{B} \approx \Delta \mathbf{B}. \end{aligned} \tag{2}$$

Thus, the discrete form of Eq. (1) becomes:

$$\begin{aligned} h_t &= \bar{\mathbf{A}} h_{t-1} + \bar{\mathbf{B}} x_t, \\ y_t &= \mathrm{C} h_t. \end{aligned} \tag{3}$$

In practice, x_t is a feature vector of size C, and Eq. (3) operates independently on each feature.

Selective scan

In Mamba, the parameters vary with input, making it impossible to reformulate Eq. (3) into a convolutional form, thus hindering SSM parallelization. Mamba addresses this with a selective scan mechanism, integrating three classical techniques: kernel fusion, parallel scan, and recomputation. This mechanism allows Mamba to achieve high processing speed with relatively low memory usage.

Overview

The processed embeddings of the three modalities are sequentially fed into the Intra-Mamba and Tri-modal Mamba Fusion modules, where the Tri-modal Mamba Fusion module is primarily structured around the Fusion Mamba Block. It employs a hierarchical fusion strategy, first enabling direct modality interaction in the middle fusion stage, followed by further integration of the fused representations in the late fusion stage to enhance cross-modal understanding. Subsequently, sample-based tri-modal contrastive learning and interaction-level auxiliary classification constraints applied to cross-modal representations are employed. Finally, the fused features are passed to the classifier to predict emotion tendencies. The specific process is shown in Fig. 1. To improve readability, Table 1 summarizes the main symbols and tensor shapes used in our framework.

Modality encoder

The modality encoder aims to obtain modality-enhanced sequential representations. It mainly consists of two parts: Temporal Convolution and Intra-Mamba.

(1) Temporal convolution. To ensure that three unimodal sequence representations lie in the same space, we feed them into a 1D convolutional layer:

$$\begin{aligned} F_a &= 1\mathbf{D}_a(F_a^{cova}; \theta_a^{1D}), \\ F_v &= \mathbf{1D}_v(F_v^{facet}; \theta_v^{1D}), \\ F_t &= 1\mathbf{D}_v(F_t^{bert}; \theta_t^{1D}), \end{aligned} \tag{4}$$

where θ_m^{1D} denotes the parameters of a one-dimensional convolutional network.

(2) Intra-modal Mamba. We introduce the intra-modal encoder to further extract unimodal features, reducing complexity compared to traditional Transformers, as shown in Fig. 2. Specifically, the Intra-modal Mamba employs a standard Mamba block structure. Here's a detailed explanation: The input feature sequence $\mathbf{H}_m \in \mathbb{R}^{B\times N\times d}$ is first layer-normalized to obtain $\mathbf{H}'_m$. Then, in two independent branches, two multilayer perceptrons (MLP) project $\mathbf{H}'_m$ to x and z. In the first branch, z is activated by a SiLU function, acting as a gating factor for f, resulting in f'. Finally, f' passes through an MLP layer and a residual connection, yielding the output $\mathbf{H}_{m\to m}$ for higher-level feature extraction. The detailed process is illustrated in Algorithm 1. We take $\mathbf{H}_m$ as input:

$$\mathbf{H}_{m\to m} = \text{Intra} - \text{Mamba}(\mathbf{H}_m) \in \mathbb{R}^{B\times N\times d} \tag{5}$$

where $\mathbf{H}_m \in \mathbb{R}^{B\times N\times d}$, $\mathbf{H}_{m\to m} \in \mathbb{R}^{B\times N\times d}$, $m \in \{t, a, v\}$.

Algorithm 1 Intra-Mamba

Input: Feature sequence $H_m : (B, N, d)$
Output: Feature sequence $\mathbf{H}_{m\to m} : (B, N, d)$
$H'_m : (B, N, d) \leftarrow \text{LayerNorm}(H_m)$
$x : (B, N, d') \leftarrow \text{MLP}_x(H'_m)$
$z : (B, N, d') \leftarrow \text{MLP}_z(H'_m)$
$x' : (B, N, d') \leftarrow \text{SiLU}(\text{Conv}(x))$
$\bar{A} : (B, N, d', D)$, $\bar{B} : (B, N, d', D)$, $C_s : (B, N, d', D) \leftarrow \text{Disc}(x')$
$f : (B, N, d') \leftarrow \text{SSM}(\bar{A}, \bar{B}, C_s)(x')$
$f' : (B, N, d') \leftarrow f \odot \text{SiLU}(z)$
$\mathbf{H}_{m\to m} : (B, N, d) \leftarrow \text{MLP}(f') + H_m$
return $\mathbf{H}_{m\to m}$

Tri-modal fusion Mamba

Currently, the Mamba architecture seldom handles multimodal information due to the lack of mechanisms similar to cross-attention. Although some work has made progress

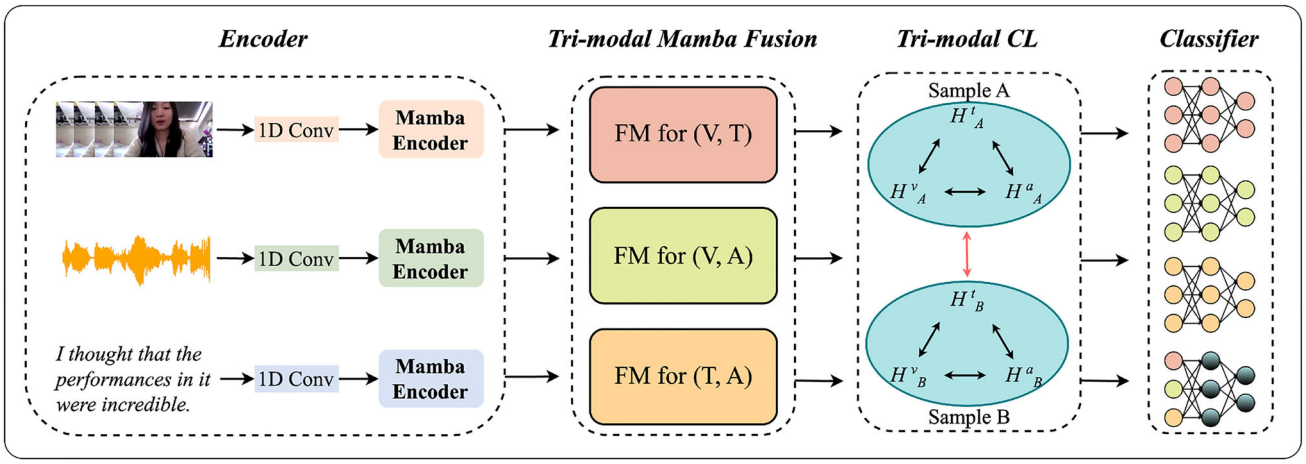


Fig. 1 The overall structure of our proposed model. The processed embeddings of the three modalities are sequentially fed into the intra-Mamba and Tri-modal Mamba Fusion modules to extract multimodal features. Subsequently, a sample-based tri-modal contrastive learning and interaction-level auxiliary classification constraints are applied. The fused features are then passed to the classifier to predict emotion tendencies

Table 1 Main symbols and tensor shapes in our framework

Symbol	Shape	Module	Description
$\mathbf{H}_m$	$B \times N \times C$	Modality encoder	Raw unimodal sequence feature
$\mathbf{H}_{m\to m}$	$B \times N \times d$	Intra-Mamba	Intra-modal enhanced representation
$\mathbf{H}_{vt}$	$B \times N \times 2d$	Fusion Mamba	Visual–text interaction feature
$\mathbf{H}_{va}$	$B \times N \times 2d$	Fusion Mamba	Visual–audio interaction feature
$\mathbf{H}_{at}$	$B \times N \times 2d$	Fusion Mamba	Audio–text interaction feature
$\mathbf{H}_{TFM}$	$B \times N \times 6d$	Tri-modal fusion	Final fused tri-modal representation
$\widehat{\mathbf{H}}^m$	$B \times N \times d$	Contrastive MLP	Projected feature for contrastive learning
$\mathbf{Y}$	$B \times 1$	Classifier	Main prediction output
$\mathbf{Y}_{mn}$	$B \times 1$	Auxiliary classifier	Interaction-level auxiliary output

in bimodal integration, there is still a gap in tri-modal fusion. Therefore, we have designed a Fusion Mamba (FM) block. The structure is illustrated in Fig. 3. To fully integrate information from each modality, we have designed a tri-modal fusion module (TFM) consisting of three parallel bimodal Mamba fusion modules. The computation equations are defined as follows:

$$\begin{aligned}\mathbf{H}_{vt} &= \mathrm{FM}\left(\mathbf{H}_{v\to v}, \mathbf{H}_{t\to t}\right) \in \mathbb{R}^{B\times N\times 2d},\\ \mathbf{H}_{va} &= \mathrm{FM}\left(\mathbf{H}_{v\to v}, \mathbf{H}_{a\to a}\right) \in \mathbb{R}^{B\times N\times 2d},\\ \mathbf{H}_{at} &= \mathrm{FM}\left(\mathbf{H}_{a\to a}, \mathbf{H}_{t\to t}\right) \in \mathbb{R}^{B\times N\times 2d}.\end{aligned} \tag{6}$$

$$\mathbf{H}_{TFM} = \mathrm{cat}[\mathbf{H}_{vt}, \mathbf{H}_{va}, \mathbf{H}_{at}] \tag{7}$$

The inputs to FM are the different modality embeddings $\mathbf{H}_{m\to m}$ obtained from the Modality Encoder. First, each modality undergoes a simple shallow fusion to obtain fused embeddings. Then, each modality fusion module constructs three branches, with each branch receiving features from two modalities. Similarly, these branches are layer-normalized, convolved, activated by SiLU, and passed through SSM to obtain outputs. Finally, the outputs are concatenated to obtain the final fused features. For more details, refer to Algorithm 2.

Contrastive learning

To ensure the effective execution of multimodal Mamba, we integrated a contrastive learning component following the three FM modules. First, we map $\mathbf{H}_{vt}$, $\mathbf{H}_{va}$, and $\mathbf{H}_{at}$ to a shared space, obtaining $\hat{H}^m$. The calculation is defined as follows:

$$\widehat{H}^m = \mathrm{MLP}(\mathbf{H}_{mn}) \tag{8}$$

where $m \in \{a, v, t\}$, $n \in \{a, v, t\}$, and $\mathrm{MLP}(*)$ is a linear layer.

Algorithm 2 Fusion Mamba Block

Input: Feature sequences $H_m : (B, N, d)$; $H_n : (B, N, d)$
Output: Feature sequence $\mathbf{H}_{m\to n} : (B, N, 2d)$
$H_f : (B, N, 2d) \leftarrow \text{Cat}(H_m, H_n)$
for $I \in \{m, n, f\}$ **do**
  $H^{I'} : (B, N, d) \leftarrow \text{LayerNorm}(H_I)$
  $x_I : (B, N, d') \leftarrow \text{MLP}_{x_I}(H^{I'})$
  $x'_I : (B, N, d') \leftarrow \text{SiLU}(\text{Conv}(x_I))$
  $\bar{A} : (B, N, d', D)$, $\bar{B} : (B, N, d', D)$, $C_s : (B, N, d', D) \leftarrow \text{Disc}(x'_I)$
  $y_I : (B, N, d') \leftarrow \text{SSM}(\bar{A}, \bar{B}, C_s)(x'_I)$
end for
$z : (B, N, d') \leftarrow \text{MLP}_z(H_f)$
$y'_m : (B, N, d') \leftarrow y_m \odot \text{SiLU}(z)$
$y'_n : (B, N, d') \leftarrow y_n \odot \text{SiLU}(z)$
$y'_f : (B, N, d') \leftarrow y_f \odot \text{SiLU}(z)$
$\mathbf{H}_{m\to n} : (B, N, 2d) \leftarrow \text{MLP}_F(y'_m + y'_n + y'_f) + H_f$
return $\mathbf{H}_{m\to n}$

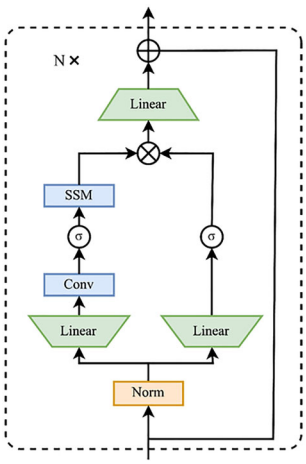


Fig. 2 Intra-Mamba architecture

Next, we construct positive and negative pairs for contrastive learning. Specifically, we build sample pairs from features of different fused modalities, such as text-audio, text-visual, and audio-visual pairs. Positive sample pairs consist of corresponding features from the same sample, while negative pairs consist of any two features from unmatched samples. Assuming a training batch contains two samples $\{S_A, S_B\}$, $\hat{H}^a_A$, $\hat{H}^v_A$, and $\hat{H}^t_A$ form positive pairs with each other. Additionally, any features among $\hat{H}^a_A$, $\hat{H}^v_A$, or $\hat{H}^t_A$ and $\hat{H}^a_B$, $\hat{H}^v_B$, $\hat{H}^t_B$ from sample B constitute negative pairs.

Referring to previous works on contrastive learning [40], we construct the following loss function for contrastive learning on X and Y pairs:

$$\mathcal{L}_{CL}(X, Y) = -\log \frac{\exp(\sin(X_i, Y_i))}{\sum_{j=1}^{N} \mathbf{1}_{[j\neq i]} \exp\left(\sin(X_i, Y_j)\right)}, \tag{9}$$

where $(X, Y) \in ((\hat{H}^a, \hat{H}^t), (\hat{H}^v, \hat{H}^t), (\hat{H}^a, \hat{H}^v))$, and X is the anchor point, X_i, Y_i denote the positive pair, the X_i, Y_j denote negative pairs, N is the sample size, and sim is the cosine similarity.

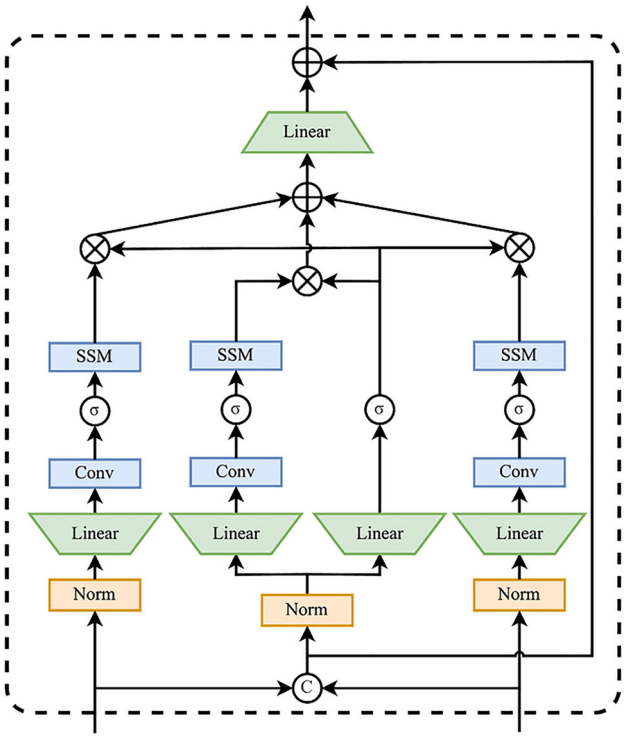


Fig. 3 Fusion Mamba architecture

Furthermore, to perform more comprehensive contrastive learning, we swap the positions of X and Y, and recompute the loss $L_{CL}(Y, X)$. Thus, the final loss for contrastive learning is defined as:

$$\mathcal{L}_{CLT}(X, Y) = \frac{1}{2}(\mathcal{L}_{CL}(X, Y) + \mathcal{L}_{CL}(Y, X)). \tag{10}$$

Finally, we improve the performance of the Tri-Modal Fusion Mamba by minimizing this loss.

Classifier

Given the cross-modal interaction representations $\mathbf{H}_{m\to n}$ generated by the multimodal fusion Mamba, we perform adaptive feature fusion to obtain the final multimodal representation $\mathbf{H}_{TFM}$. Specifically, $\mathbf{H}_{m\to n}$ denotes the intermediate representation after information from modality m is injected into modality n.

The fused representation is projected to the prediction space as:

$$\mathbf{Y} = \text{MLP}(\mathbf{H}_{TFM}), \tag{11}$$

where MLP denotes a two-layer fully connected network.

Table 2 The compositions of the number of samples in the CMU-MOSI, CMU-MOSEI, and MELD datasets

Datasets	#Train	#Validation	#Test
CMU-MOSI	1284	229	686
CMU-MOSEI	16,326	1871	4659
MELD	1153	–	280

In addition to the main classification head, we introduce an interaction-level auxiliary classifier for each cross-modal representation:

$$\mathbf{Y}_{mn} = \text{MLP}(\mathbf{H}_{m\to n}), \tag{12}$$

where $m, n \in \{a, v, t\}$ and $m \neq n$.

Based on these predictions, the classification losses are defined as:

$$\begin{aligned} \mathcal{L}_{task1} &= \text{CE}(\mathbf{Y}, \text{label}), \\ \mathcal{L}_{task2} &= \lambda_1 \sum_{m\neq n} \text{CE}(\mathbf{Y}_{mn}, \text{label}), \end{aligned} \tag{13}$$

where $\mathcal{L}_{task2}$ serves as an interaction-level auxiliary classification loss.

The overall task loss is:

$$\mathcal{L}_{task} = \mathcal{L}_{task1} + \mathcal{L}_{task2}. \tag{14}$$

Finally, the total training objective is formulated as:

$$\mathcal{L} = \mathcal{L}_{task} + \lambda_2 \mathcal{L}_{CLT}(t, v) + \lambda_3 \mathcal{L}_{CLT}(t, a) + \lambda_4 \mathcal{L}_{CLT}(v, a), \tag{15}$$

where λ_1–λ_4 are hyperparameters controlling the contribution of each loss term.

Experiments

In this section, we provide a detailed description of the experimental setup for our work, including datasets, baselines, evaluation metric, and implementation details.

Dataset

We evaluate FMCL on CMU-MOSI [41], CMU-MOSEI [42] and MELD datasets. The experiments are conducted under the word-aligned setting. Table 2 details the split composition of the three datasets.

CMU-MOSI: This dataset includes 2199 brief monologue video clips. Acoustic features are sampled at 12.5 Hz, and visual features are sampled at 15 Hz.

CMU-MOSEI: This dataset has 22,856 movie review video samples from YouTube, about ten times larger than CMU-MOSI. Acoustic features in CMU-MOSEI are sampled at 20 Hz, while visual features are sampled at 15 Hz.

Table 3 Performance comparison on CMU-MOSI and CMU-MOSEI datasets

Model	CMU-MOSI				CMU-MOSEI			
	Acc-2 (%)	Acc-7 (%)	F1 (%)	MAE	Acc-2 (%)	Acc-7 (%)	F1 (%)	MAE
EF-LSTM	75.8	31.6	75.6	1.053	79.1	46.7	78.8	0.665
LF-LSTM	76.4	31.6	75.4	1.037	79.4	49.1	80.0	0.625
TFN	73.9	32.1	73.4	0.901	82.5	50.2	82.1	0.593
LMF	76.4	32.8	75.7	0.917	82.0	48.0	82.1	0.623
MFM	78.1	36.2	78.1	0.877	84.4	51.3	84.3	0.568
RAVEN	78.8	33.8	76.9	0.968	79.1	50.0	79.5	0.680
MCTN	79.3	40.0	79.1	0.871	79.8	49.6	80.6	0.580
MulT	83.0	40.0	82.8	0.871	82.5	51.8	82.3	0.580
MISA	83.4	42.3	83.6	0.783	85.5	52.2	85.3	0.555
DMD	85.37	45.77	85.35	0.726	85.7	52.97	85.6	0.538
Self-MI	85.37	–	85.3	0.699	85.45	–	85.37	0.527
CMHFM	81.71	37.17	81.72	0.907	84.45	52.83	83.97	0.548
TCAN	86.28	46.79	86.15	0.714	***86.27***	53.10	***86.17***	0.532
DLF	85.06	47.08	85.04	0.731	85.42	53.90	85.27	0.536
Coupled Mamba	–	–	–	–	85.60	–	85.70	0.547
FMCL (ours)	***86.43***	***48.10***	***86.37***	***0.696***	85.80	***53.98***	85.68	**0.5318**

Bold and bold-italic fonts indicate the best results in the corresponding columns

MELD: This dataset is a multi-party dataset derived from the TV series Friends. It consists of 13,708 utterances and 1433 dialogues. Each utterance is annotated with one of seven emotion categories: anger, disgust, fear, joy, neutral, sadness, and surprise.

Evaluation metric

Each sample in CMU-MOSI and CMU-MOSEI received a emotion score ranging from − 3 to 3, covering highly negative, negative, weakly negative, neutral, weakly positive, positive, and highly positive. The MER performance is evaluated using the following metrics:

- **Acc-7:** The 7-class accuracy;
- **Acc-2:** The binary accuracy with zero exclusion;
- **MAE:** The mean absolute error;
- **F1 score:** Provides a comprehensive performance evaluation of the model.

Implementation details

In the CMU-MOSI and CMU-MOSEI datasets, we used GloVe [43] to extract unimodal language features, resulting in 300-dimensional word features. To ensure fair comparison in the aligned setting, we also used a BERT-base-uncased pre-trained model [44] to get 768-dimensional hidden states as word features. For the visual modality, each video frame was encoded with Facet [45] to detect 35 facial action units. The acoustic modality was processed using COVAREP [46] to extract 74-dimensional features. Hyperparameters λ_1, λ_2, λ_3, and λ_4 were set to 0.5, 0.05, 0.03, and 0.05, respectively. For the MELD dataset, hyperparameters λ_1, λ_2, λ_3, and λ_4 were set to 0.5, 0.1, 0.1, and 0.05, respectively. All experiments were conducted using PyTorch on two RTX A4000 GPUs, each with 16GB of memory. The training batch size was set to 16.

Baseline

We compared FMCL with current state-of-the-art methods using the same dataset settings, including EF-LSTM, LF-LSTM, TFN [17], LMF [16], MFM [18], RAVEN [47], MCTN [48], MulT [27], MISA [49], DMD [19], Self-MI [50], CMHFM [51], TCAN [52], DLF [53], Coupled Mamba [36], DER-GCN [54].

Results and analysis

Quantitative analysis

Table 4 Results on the MELD dataset

Models	neutral		surprise		fear		sadness		joy		disgust		anger		ACC	w-F1
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1		
DialogueRNN	***85.11***	79.60	54.09	56.72	10.00	12.66	29.81	38.63	62.94	63.81	22.06	27.37	53.62	53.24	66.70	65.31
MMGCN	81.53	79.20	58.36	57.75	13.79	13.71	31.73	40.99	**69.90**	63.43	20.54	25.56	52.17	53.19	66.61	65.21
DialogueTRM	83.44	79.54	54.45	57.09	27.91	33.17	33.17	40.95	60.09	62.93	21.28	24.73	***58.26***	53.96	66.56	65.22
MM-DFN	83.52	79.65	***63.35***	58.17	***32.00***	26.67	26.44	35.71	63.68	64.89	19.12	24.76	49.28	52.15	66.55	65.48
SDT	83.22	80.19	61.28	***59.07***	13.80	17.88	***34.90***	***43.69***	63.24	64.29	***22.65***	***28.78***	56.93	***57.43***	67.55	66.60
DER-GCN	76.8	***80.6***	50.5	51.0	14.8	10.4	56.7	41.5	69.3	64.3	17.2	10.3	52.5	57.4	66.8	66.1
FMCL (ours)	84.40	***80.41***	59.40	58.79	18.15	23.85	32.10	42.12	67.30	64.65	15.95	23.40	54.90	55.19	***67.95***	***66.70***

Bold and bold-italic fonts indicate the best results in the corresponding columns

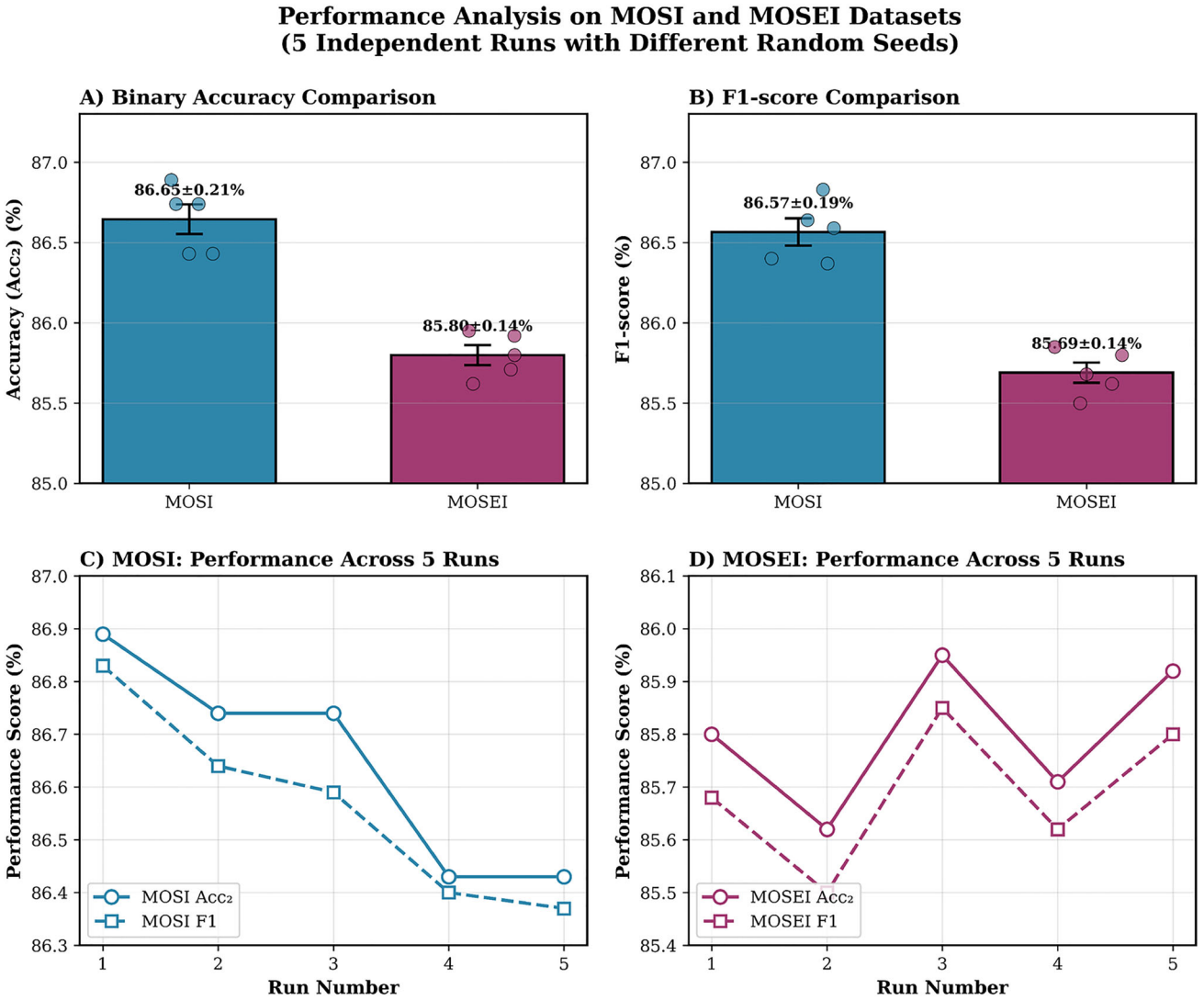


Fig. 4 Performance analysis across 5 runs with different random seeds on the MOSI and MOSEI datasets. (**a**) Binary classification accuracy comparison; (**b**) F1 score comparison; (**c**) trend of 5 runs on the MOSI dataset; (**d**) trend of 5 runs on the MOSEI dataset

We compared our method with existing state-of-the-art approaches. Table 3 presents the corresponding results on the CMU-MOSI and CMU-MOSEI datasets. Compared to previous methods, our approach shows significant improvements across all four metrics, particularly in Acc-7 and MAE, highlighting substantial progress in fine-grained emotion classification. The performance on the CMU-MOSI dataset is particularly notable. Our significant improvements over previous works can be attributed to several key enhancements:

- **Intra-modal feature extraction**: By utilizing the Mamba model to further extract intra-modal features, we enhanced the representation capability of unimodal features.
- **Multimodal interaction fusion mechanism**: Introducing a mechanism for multimodal interaction and fusion improved the collaborative effects between different modalities. **Interaction-level auxiliary classification loss**: Incorporating interaction-level auxiliary classification loss during training further optimized the model's performance.

These improvements enabled our method to perform exceptionally well across all metrics. Compared to the DMD model, which adopts a decoupling strategy, our method showed enhancements across all metrics. Particularly on the CMU-MOSI dataset, Acc-2 and F1 score improved by over 1%, strongly demonstrating the effectiveness and superiority of our proposed method.

In addition, experiments were conducted on the MELD dataset (see Table 4). This dataset exhibits an imbalanced distribution of emotion categories, with neutral and joy being overrepresented, while fear and disgust are underrepresented.

Such class imbalance may adversely affect the model's classification performance, particularly for low-frequency categories.

The experimental results demonstrate that the proposed method achieves promising performance in terms of overall classification accuracy (ACC) and weighted F1-score (W-F1), surpassing the compared methods. However, classification performance varies across different emotion categories. For instance, some baseline methods achieve better results for surprise, fear, and anger, whereas other methods perform better for disgust and sadness. This indicates that although the proposed method achieves superior overall performance, there remains room for improvement in classifying certain emotions. This phenomenon may be attributed to the dataset's class imbalance, which causes the model to focus more on high-frequency categories during training while being less effective in distinguishing low-frequency emotions. Additionally, different emotions exhibit varying feature representations across speech, text, and visual modalities, and the current multimodal fusion strategy may not fully capture key information for certain categories. To address these limitations, future work will explore class-weighted loss functions or data augmentation strategies to enhance the classification of minority emotion classes. Furthermore, improving multimodal fusion techniques could further enhance the model's performance across different emotion categories.

Notably, significant improvements were achieved in the seven-class classification task on the MOSI and MOSEI datasets. However, the absolute accuracy remains lower than that observed in the binary classification task. This phenomenon primarily stems from the decision-making mechanism of the binary classification: binary classification results are obtained by applying a threshold to the seven-class predictions (where values greater than 0 are considered positive and values less than 0 are considered negative). This hierarchical design leads to several key characteristics:

- **Decision boundary effect**: In the seven-class classification task, samples close to the decision boundary (e.g., $+1/-1$) are prone to misclassification. However, in the binary classification task, such misclassifications can often be automatically corrected—if the predicted sign is correct, the classification remains accurate.
- **Difference in classification difficulty**: The seven-class classification task requires the model to accurately distinguish between seven discrete sentiment intensity levels (e.g., differentiating between $+3$ and $+2$). In contrast, the binary classification task only involves determining the sentiment polarity. This inherently reduces the classification difficulty, resulting in a more constrained error space for binary classification compared to seven-class classification.
- **Influence of subjectivity in annotation**: Sentiment annotation inherently involves a high degree of subjectivity, as different annotators may assign varying sentiment scores to the same video sample. This annotation bias is particularly pronounced in the fine-grained seven-class classification task, making it more challenging for the model to learn clear category boundaries. For example, the distinction between $+2$ and $+3$ may depend entirely on the annotator's personal interpretation rather than objective sentiment cues from the video itself. This subjectivity increases the difficulty of the seven-class classification task, thereby complicating fine-grained sentiment intensity modeling.

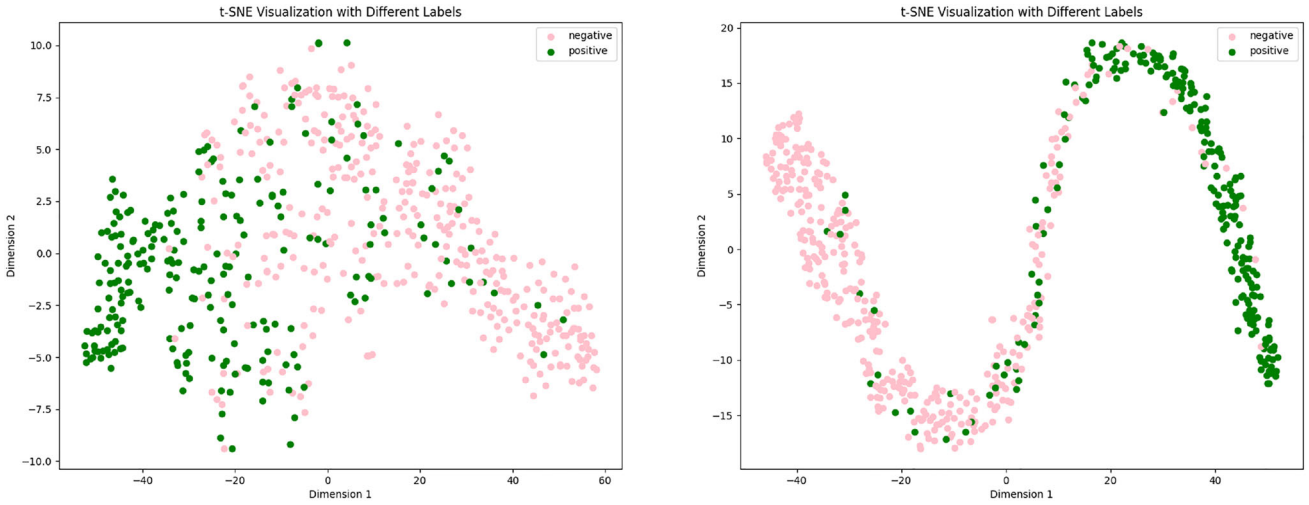
Furthermore, an overall classification accuracy (ACC) of 67.95% was achieved on the MELD dataset, significantly higher than the 48.10% on CMU-MOSI and 53.98% on CMU-MOSEI. This significant accuracy gap highlights the inherent differences between datasets in multimodal sentiment analysis tasks, demonstrating the critical impact of dataset characteristics on model performance. Additionally, this result further verifies the adaptability and robustness of the proposed method across different task scenarios. In future work, a hierarchical classification strategy will be introduced, where coarse-grained sentiment polarity (positive, neutral, and negative) is first distinguished, followed by fine-grained classification within each polarity group.

Randomness testing

The performance of multimodal sentiment recognition models can fluctuate due to random initialization. To assess the stability of our model, we conducted independent experiments on the MOSI and MOSEI datasets using five different random seeds. The experimental results demonstrate that our FMCL model exhibits good stability under different random initializations.

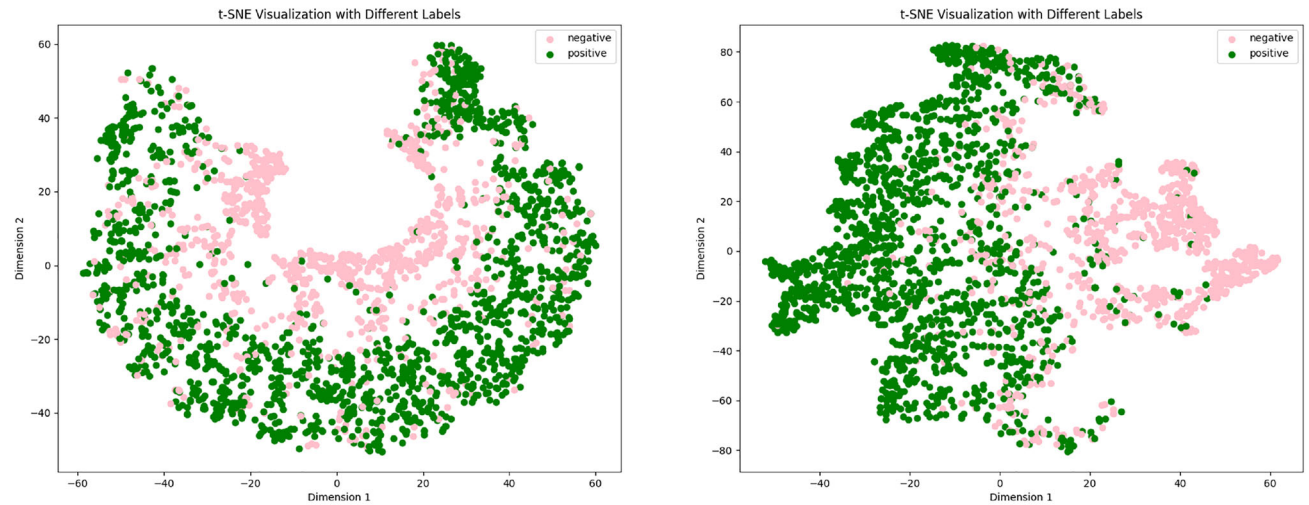
On the MOSI dataset, the binary classification accuracy (Acc-2) across five runs ranged from 86.43% to 86.89%, with a mean of 86.65% and a standard deviation of 0.19%. The F1 scores ranged from 86.37% to 86.83%, with a mean of 86.57% and a standard deviation of 0.18%. On the MOSEI dataset, the binary classification accuracy ranged from 85.62% to 85.95%, with a mean of 85.80% and a standard deviation of 0.14%. The F1 scores ranged from 85.50% to 85.85%, with a mean of 85.69% and a standard deviation of 0.14%.

As shown in Fig. 4, the model's performance on both datasets shows high consistency: figures (a) and (b) display the overall performance comparison between the MOSI and MOSEI datasets, with error bars indicating standard error; figures (c) and (d) respectively show the detailed trends of the five runs on the two datasets, where solid lines connecting



(1) Original (2) FMCL

(a) CMU – MOSI



(1) Original (2) FMCL

(b) CMU – MOSEI

Fig. 5 T-SNE visualization of binary classification on CMU-MOSI and CMU-MOSEI datasets

circular markers represent binary classification accuracy, and dashed lines connecting square markers represent F1 scores.

This stability can be attributed to several factors: (1) the selective state space mechanism of the Mamba architecture has deterministic computational properties, reducing sensitivity to initial conditions; (2) the structured multimodal fusion method provides stable learning signals; (3) contrastive learning regularization helps the model converge to similar optimization regions. These results demonstrate the reliability and reproducibility of our method under different random initialization conditions.

Visualization

To better demonstrate the effectiveness of the proposed FMCL model, we visualized the classification results for binary and seven-class tasks on two datasets using t-SNE [55], as shown in Figs. 5 and 6. It is worth noting that while the

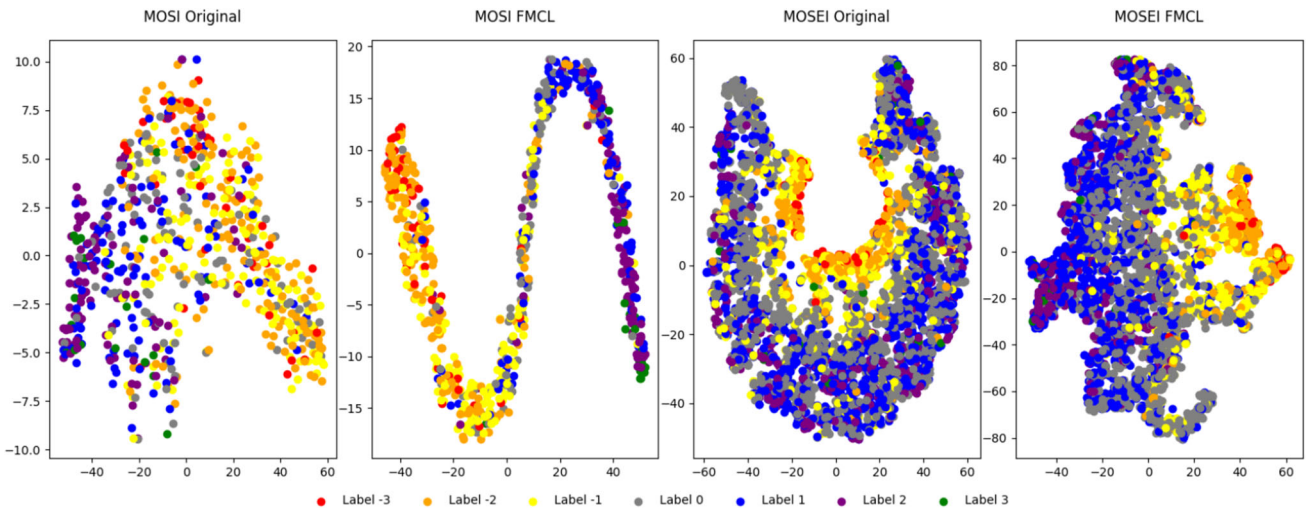


Fig. 6 T-SNE visualization of 7-class classification on CMU-MOSI and CMU-MOSEI datasets

Table 5 Quantitative cluster quality evaluation of t-SNE on MOSI and MOSEI

Dataset	Model	DBS (↓)	CHS (↑)
MOSI	Original	14.5890	63.1168
	FMCL	***3.3515***	***143.7114***
MOSEI	Original	49.4022	9.9611
	FMCL	***6.1103***	***128.0802***

Bold and bold-italic fonts indicate the best results in the corresponding columns

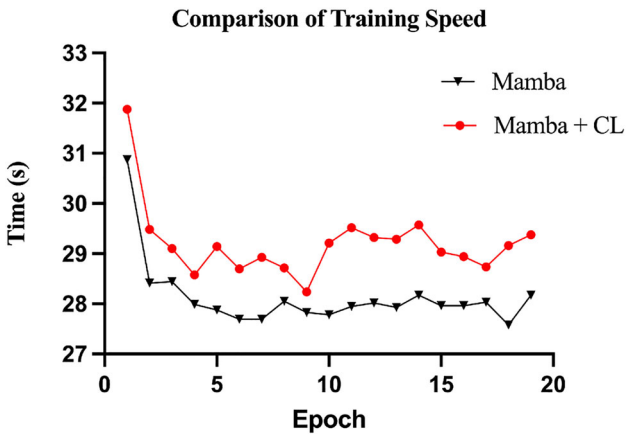


Fig. 7 Comparison of average training time per epoch before and after introducing contrastive learning (CL)

shapes of the two plots are similar, the colors differ because the binary classification results are derived from the seven-class results. Specifically, samples with a score greater than 0 in the seven-class results are labeled as positive sentiment, whereas samples with a score less than 0 are labeled as negative sentiment.

Through t-SNE visualization, the superiority of the FMCL algorithm in the binary classification task is intuitively demonstrated. In this task, samples from each category are tightly distributed and easily distinguishable. However, in the seven-class task, the results are relatively poor, with overlapping clusters and difficulty in effectively separating samples from different labels. To scientifically compare the differences in the seven-class visualization, two common clustering quality metrics were computed:

- **Calinski–Harabasz Score (CHS):** This metric reflects the compactness within clusters and the separation between clusters. A higher value indicates better clustering performance. A higher CHS suggests stronger separation between different categories and a more compact distribution of samples within the same category.
- **Davies–Bouldin Score (DBS):** This metric measures the similarity between clusters, where a lower value indicates better clustering performance. A lower DBS suggests better distinguishability between clusters and higher similarity within each cluster.

As shown in Table 5, the FMCL method shows significant improvement over the original features in these metrics. Specifically, for both the MOSI and MOSEI datasets, FMCL achieves significantly lower DBS values and significantly higher CHS values compared to the original features, indicating that our method has notably improved clustering performance, especially in the binary classification task. However, in the seven-class task, despite improvements over the original features, issues such as category overlap and insufficient separation persist. This suggests that further enhancements in feature representation or the introduction of additional strategies may be necessary to improve clustering performance in multi-class classification tasks.

Table 6 Performance comparison of different ablation modules on CMU-MOSI and CMU-MOSEI

Model	CMU-MOSI				CMU-MOSEI			
	Acc-2 (%)	Acc-7 (%)	F1 (%)	MAE	Acc-2 (%)	Acc-7 (%)	F1 (%)	MAE
(1) w/o Intra-Mamba	85.37	46.50	85.28	0.705	85.25	53.19	85.17	0.538
(2) w/o TFM	85.21	45.34	85.17	0.728	85.17	53.25	85.04	0.539
(3) w/o CL	85.52	45.63	85.41	0.715	85.00	53.36	85.03	0.551
(4) w/o $\mathcal{L}_{CLT}$	85.52	45.63	85.41	0.715	85.00	53.36	85.03	0.551
(5) w/o $\mathcal{L}_{task2}$	85.37	46.36	85.24	0.703	85.53	53.32	85.48	0.539
(6) FMCL	***86.43***	***48.10***	***86.37***	***0.696***	***85.80***	***53.98***	***85.68***	***0.5318***

Bold and bold-italic fonts indicate the best results in the corresponding columns

Table 7 Performance comparison of different ablated modalities on CMU-MOSI and CMU-MOSEI

Model	CMU-MOSI				CMU-MOSEI			
	Acc-2 (%)	Acc-7 (%)	F1 (%)	MAE	Acc-2 (%)	Acc-7 (%)	F1 (%)	MAE
(1) w/o Visual	86.13	45.19	86.04	0.720	85.11	52.46	84.83	0.546
(2) w/o Audio	85.21	46.21	85.09	0.718	85.22	52.97	85.11	0.538
(3) w/o Text	55.64	21.87	55.80	1.476	64.97	41.77	60.40	0.816
(4) V+A+T	***86.43***	***48.10***	***86.37***	***0.696***	***85.80***	***53.98***	***85.68***	***0.5318***

Bold and bold-italic fonts indicate the best results in the corresponding columns

Table 8 The best result in each column is highlighted in bolditalic and bold

Model	Acc-2 (%) ↑				F1 (%) ↑				Acc-7 (%) ↑				MAE ↓			
	30%	50%	70%	90%	30%	50%	70%	90%	30%	50%	70%	90%	30%	50%	70%	90%
MISA [49]	73.8	67.1	61.7	56.7	73.5	65.6	59.6	47.1	31.8	27.3	22.9	19.8	1.003	1.134	1.248	1.329
MAG-BERT [56]	71.0	66.3	64.0	54.8	70.5	64.5	57.9	46.4	36.3	29.5	22.6	18.5	0.994	1.128	1.253	1.416
Self-MM [57]	76.1	69.3	62.9	53.5	75.3	67.8	59.4	44.4	35.8	31.8	25.5	18.3	0.929	***1.053***	1.194	1.374
MMIM [58]	74.4	67.2	60.9	54.5	74.2	65.1	55.9	44.6	34.9	28.8	23.4	18.7	0.968	1.142	1.283	1.392
MissModal [59]	76.3	69.6	66.7	60.1	75.5	67.4	61.8	50.9	***38.7***	***32.3***	***27.1***	21.3	***0.907***	1.039	***1.164***	***1.316***
Ours (FMCL)	***78.7***	***70.1***	***68.8***	***60.2***	***78.7***	***70.3***	***68.9***	***59.9***	37.2	30.0	24.6	***21.7***	0.917	1.140	1.197	1.407

Performance comparison under different text-modality missing ratios on CMU-MOSI

Ablation study

Ablation of FMCL

To evaluate the effectiveness of our proposed model, we removed each key module and critical loss function. Specifically, we removed the Intra-Mamba module, Tri-Modal Fusion Mamba (TFM), and Contrastive Learning (CL) module from the complete FMCL model; for critical loss functions, we removed the contrastive learning loss ($\mathcal{L}_{CLT}$) and the interaction-level auxiliary classification loss ($\mathcal{L}_{\text{task2}}$), which is applied to the cross-modal interaction representations $\mathbf{H}_{m\to n}$ rather than the original unimodal features.

The ablation study results in Table 6 demonstrate that each component of the FMCL model contributes to its overall performance. Specifically, the TFM module and the Intra-Mamba module play key roles in enhancing the effectiveness of multimodal feature fusion and classification. In the CMU-MOSI dataset, removing the TFM module results in the most significant performance degradation, whereas in the CMU-MOSEI dataset, removing the Intra-Mamba module has the greatest impact on the Acc-7 metric.

In addition, removing the interaction-level auxiliary classification loss $\mathcal{L}_{\text{task2}}$ consistently leads to performance drops on both datasets, indicating that explicitly supervising intermediate cross-modal interaction representations is beneficial for learning task-discriminative features.

Although the removal of the contrastive learning module and its corresponding loss also leads to performance degradation, their impact is relatively smaller compared to the core fusion modules. Notably, the best performance is achieved when the interaction-level auxiliary supervision and contrastive learning are jointly employed, suggesting that these

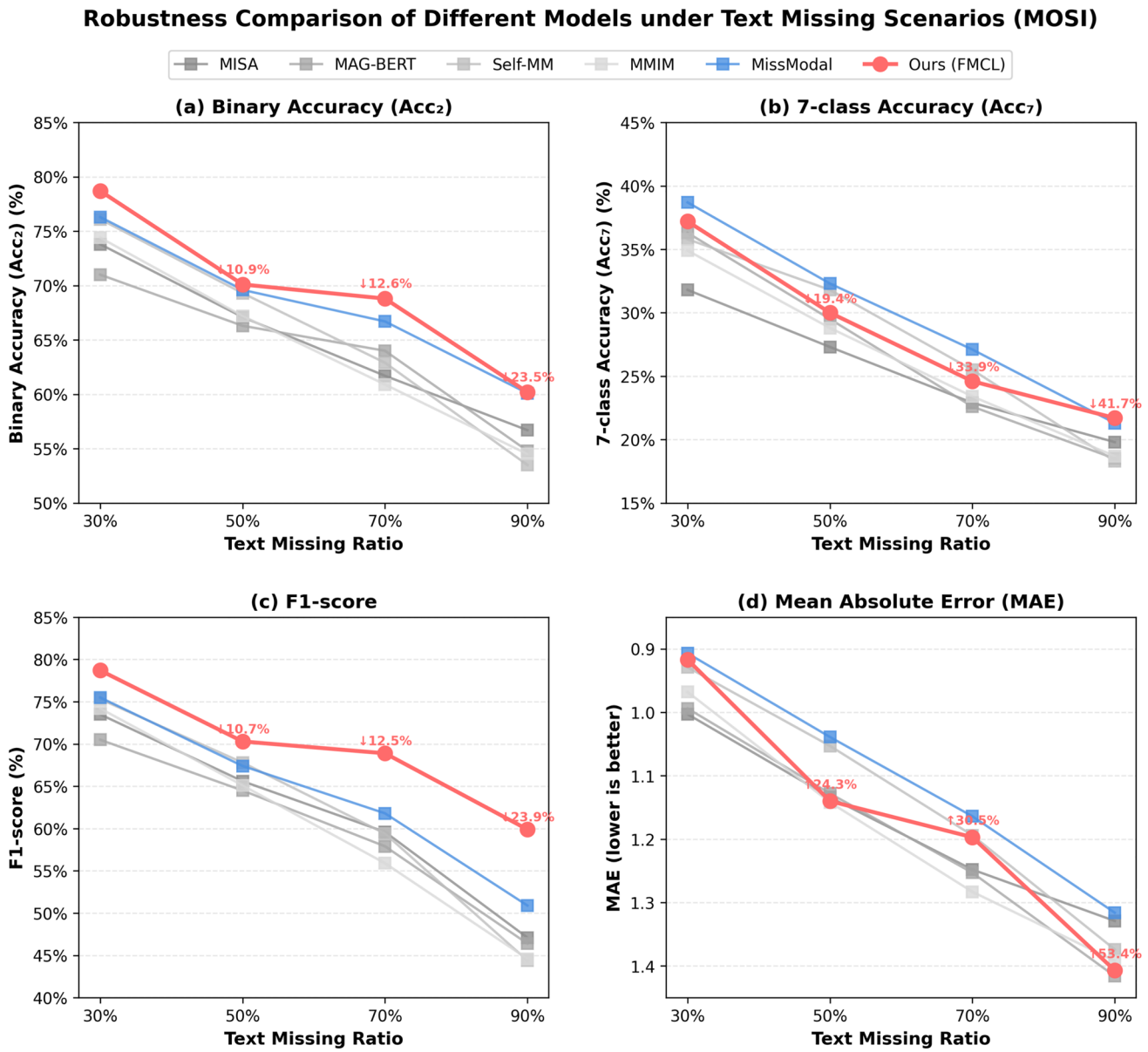


Fig. 8 Robustness comparison of different models under partial text modality missing scenarios on the MOSI dataset. The percentage drop (shown in red for FMCL) is calculated relative to the model's performance at 30% missing rate. A smaller drop indicates better robustness

two components are complementary in improving the robustness and generalization of multimodal representations.

In this study, we introduced tri-modal contrastive learning (CL) based on the Mamba architecture, which added a certain level of complexity to the model. Notably, since contrastive learning did not introduce additional training parameters, only minor changes in memory usage were observed during the experiments. To further evaluate its impact on training efficiency, we compared the training time before and after incorporating contrastive learning.

The results in Fig. 7 show that after incorporating contrastive learning, the average training time per epoch increased from 28.13 to 29.21 s, representing an increase of approximately 1.08 s (about 3.84%). Although this increase is relatively small, it reflects the additional computational burden associated with calculating similarities or differences between data from different modalities in contrastive learning.

Modality ablation and robustness analysis

To investigate the contribution of each modality and the model's robustness under realistic conditions, we conduct ablation studies with both complete modality removal and partial modality missing scenarios.

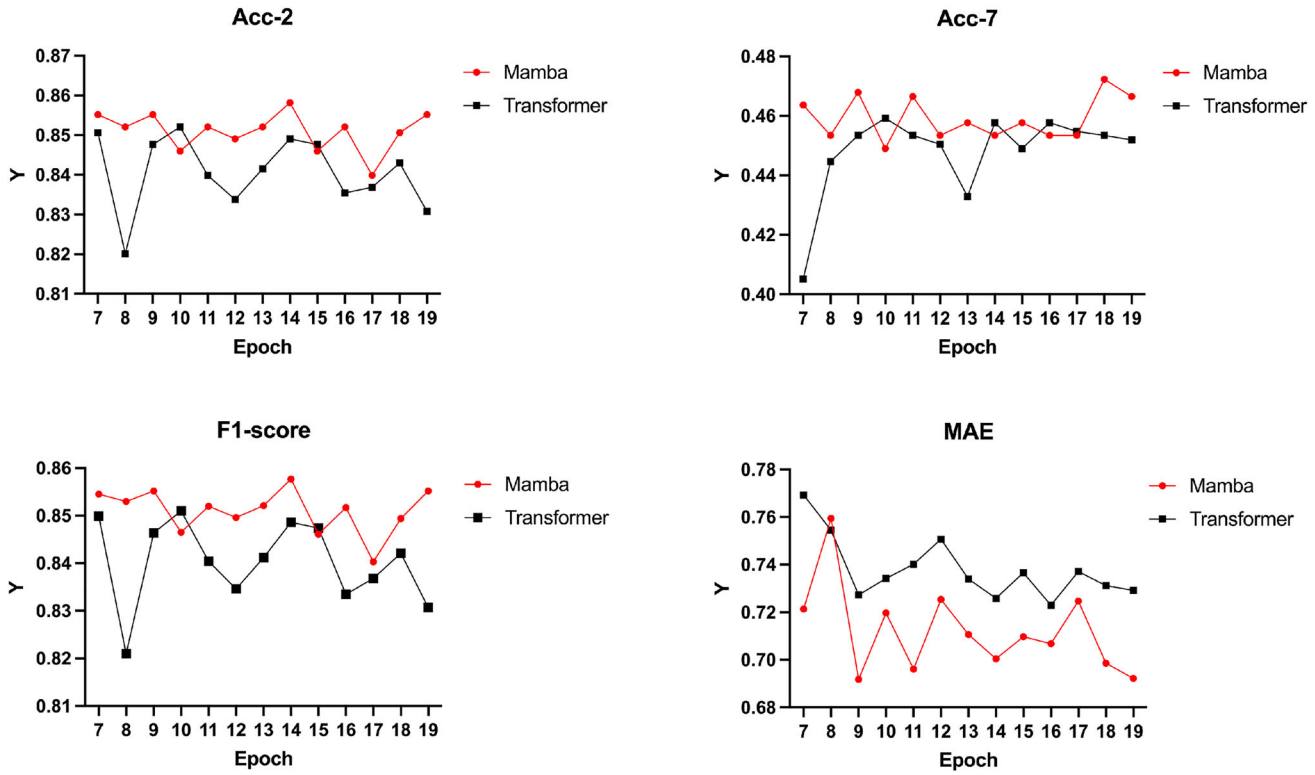


Fig. 9 Performance comparison of Mamba and Transformer based on the DMD model on the CMU-MOSI dataset

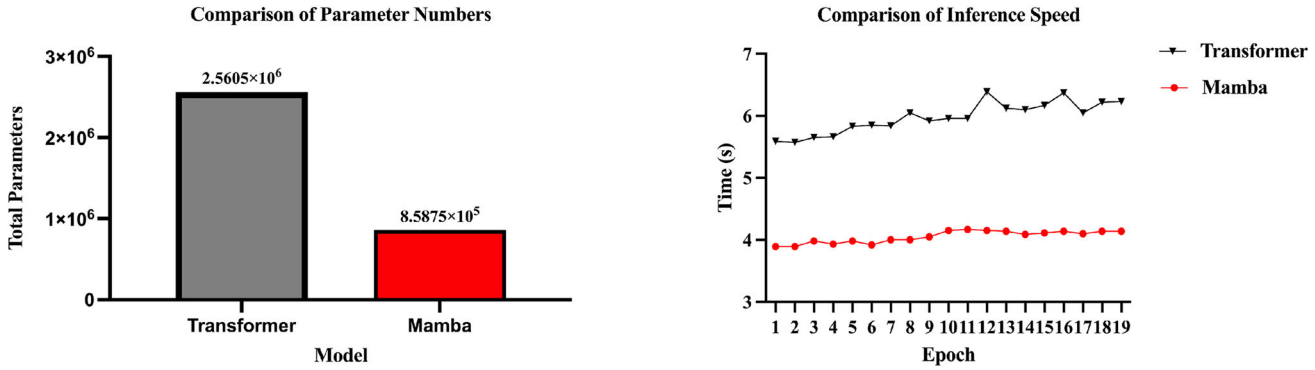


Fig. 10 Comparison of parameter numbers and inference speed between Mamba and Transformer

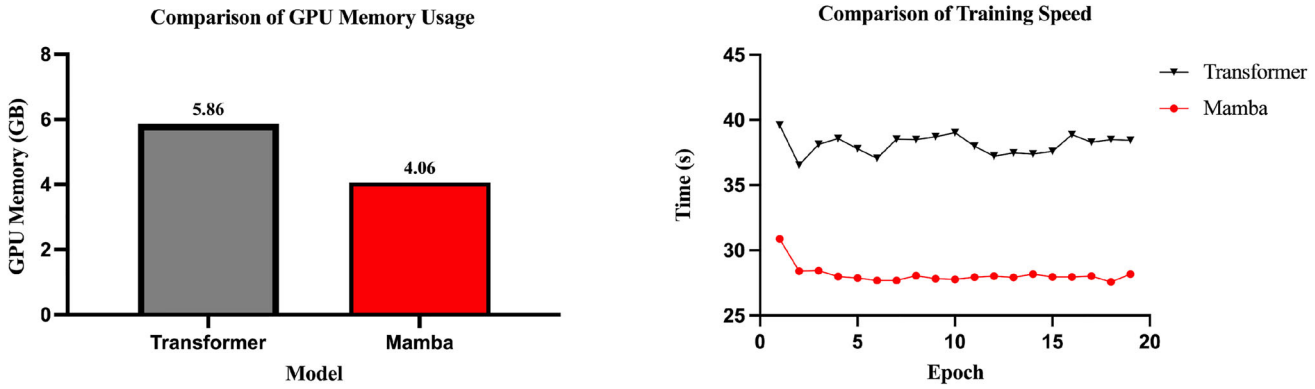


Fig. 11 Comparison of GPU memory usage and training speed between Mamba and Transformer

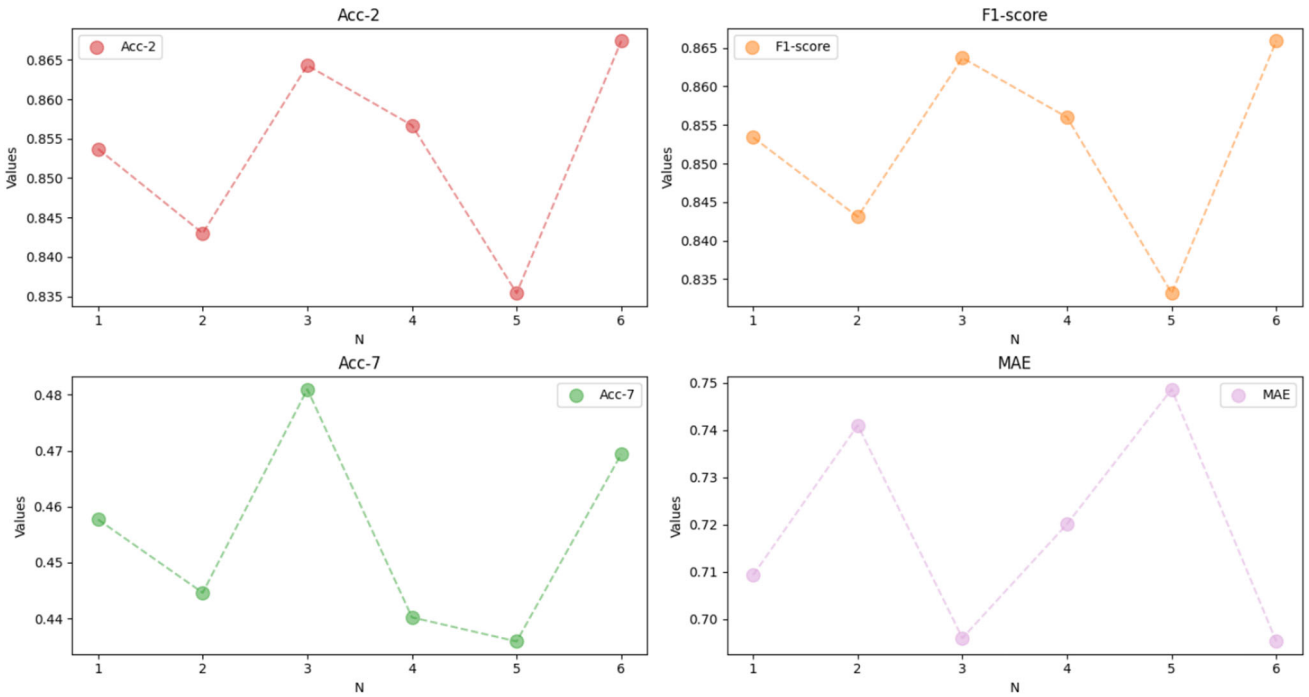


Fig. 12 Performance of different numbers of Mamba blocks on CMU-MOSI

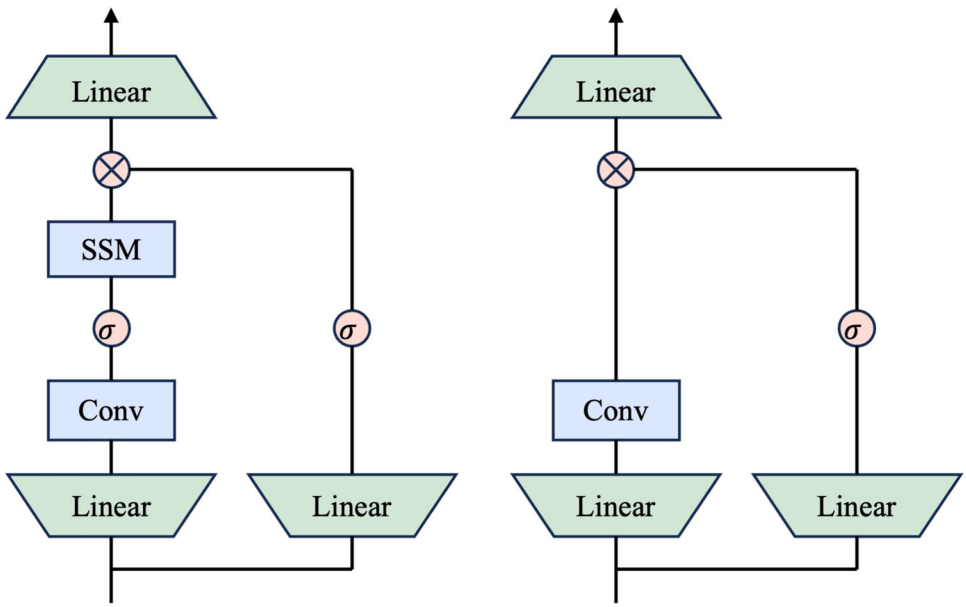


Fig. 13 Comparison of removing the SSM module

Complete modality removal. In Table 7, we individually remove each modality to explore the performance of the dual-modality FMCL. Compared to the tri-modal FMCL, the dual-modality FMCL consistently performs worse, indicating that each modality provides indispensable information. Notably, when the text modality is completely removed, the model's performance degrades significantly, while removing the visual or auditory modalities results in only a minor performance drop. This confirms that the text modality is the most crucial for sentiment analysis, likely because it contains richer and more explicit semantic information compared to the other two modalities.

Partial text modality missing (robustness analysis). While the complete removal experiments show the *presence*

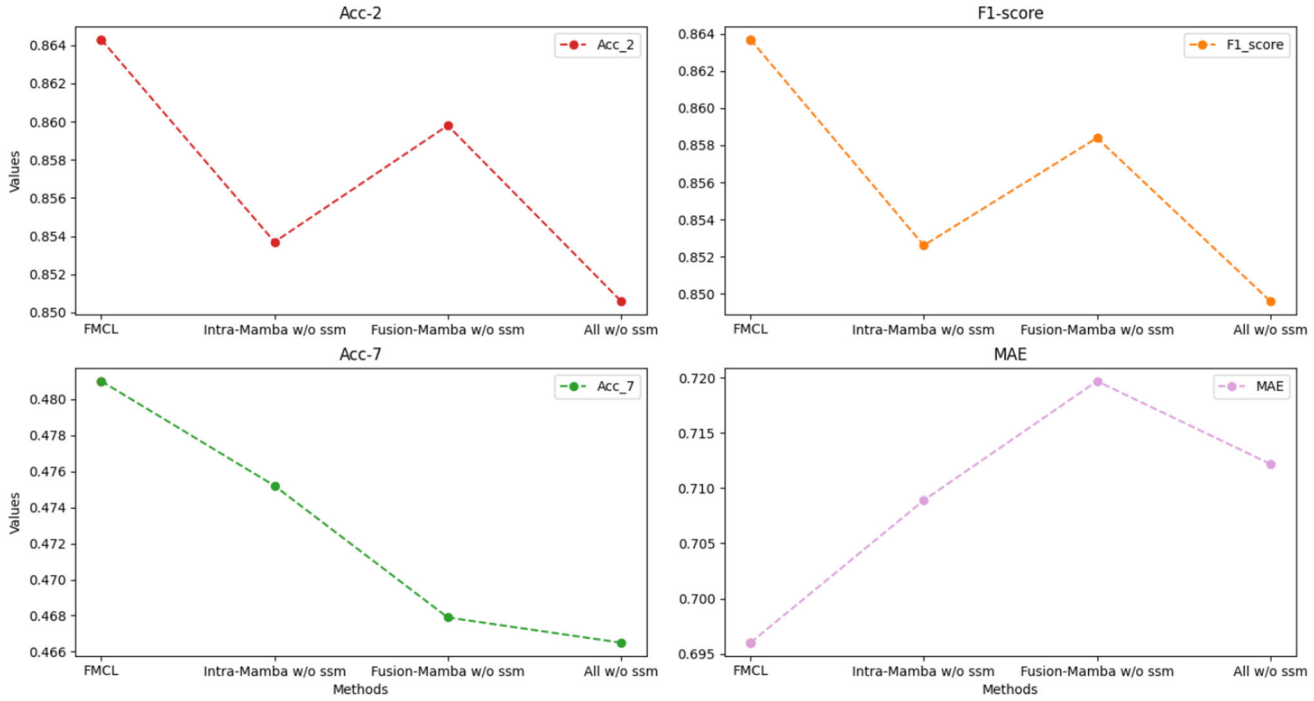


Fig. 14 Visualization of different configurations without the SSM module on CMU-MOSI

Table 9 Performance of different numbers of Mamba blocks on CMU-MOSI

N	Acc-2 (%)	F1 score (%)	Acc-7 (%)	MAE
1	85.37	85.34	45.77	0.709
2	84.30	84.31	44.46	0.741
3	86.43	86.37	48.10	0.696
4	85.67	85.60	44.02	0.720
5	83.54	83.32	43.59	0.748
6	86.74	86.59	46.94	0.695

of each modality is essential, real-world applications often face scenarios where modalities are partially available or corrupted. Since text is the dominant modality, its degradation most severely impacts performance. To assess model robustness under such realistic conditions, we simulate partial text modality missing on the MOSI dataset and compare against several strong baselines.

The results are shown in Fig. 8 and Table 8. As the text missing ratio increases from 30% to 90%, all models' performance declines. However, our FMCL model demonstrates superior robustness, particularly in maintaining a high F1 score. For instance, with 90% of the text missing, our model achieves an F1 score of 59.9%, significantly higher than other baselines (50.9% or lower). In terms of absolute change, the F1 score of our model decreases by only 18.8 points (from 78.7% at 30% missing to 59.9% at 90% missing), which is the smallest among all compared methods. This indicates that our fusion strategy and contrastive learning objective enable the model to better utilize the remaining modality information and maintain a good precision-recall balance even when the key modality is severely compromised.

Analysis and practical implications. The complete ablation highlights the *necessity* of all modalities, especially text. The partial missing experiments further demonstrate the *robustness* of our model's architecture. This robustness stems from: (1) the selective state space mechanism of Mamba, which dynamically focuses on relevant features from available data; (2) the multimodal contrastive learning, which strengthens cross-modal alignment and representation qual-

Table 10 Performance of different configurations without the SSM module on CMU-MOSI

Configuration	Acc-2 (%)	F1 score (%)	Acc-7 (%)	MAE
FMCL	86.43	86.37	48.10	0.696
Intra-Mamba w/o SSM	85.37	85.26	47.52	0.7089
Fusion-Mamba w/o SSM	85.98	85.84	46.79	0.7197
All w/o SSM	85.06	84.96	46.65	0.7122

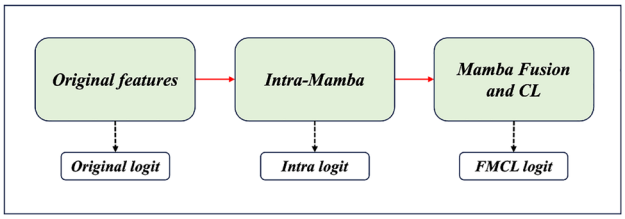


Fig. 15 Case of model

ity; and (3) our structured fusion strategy, which does not overly rely on any single modality being complete.

These findings are crucial for practical deployment. In real applications, textual information can be partially lost or corrupted due to reasons such as **automatic speech recognition (ASR) errors, data transmission issues, or incomplete user input**. Our model's ability to maintain reasonable performance under such conditions makes it more reliable and applicable for real-world multimodal sentiment analysis systems.

Ablation of Mamba

To further verify the effectiveness of the Mamba module, we compared our proposed FMCL with the state-of-the-art DMD model. Specifically, we used DMD as the baseline, where DMD employed Transformer. For comparison, we replaced the Transformer in DMD with Mamba. For single-modality Transformers, we used Intra-Mamba as a replacement, and for Multimodal Transformers, we used Fusion Mamba as a replacement. The results are shown in Fig. 9.

Figure 6 shows the performance metrics (Acc-2, Acc-7, F1, and MAE) over training epochs, comparing Mamba and Transformer implementations. The visualized results indicate consistent improvements when using Mamba, thus affirming its effectiveness in enhancing MER.

As shown in Fig. 10, we compared the number of parameters and inference speed of Mamba and Transformer. The results indicate that Mamba has significantly fewer parameters than Transformer while providing higher performance. Mamba's parameter count is 858,750, whereas Transformer's parameter count is 2,560,500, a reduction of 66.5%. This demonstrates a significant advantage in model size for Mamba. In terms of inference speed, Mamba also outperforms Transformer. Over 19 epochs, Mamba's average inference time is 4.05 s, compared to Transformer's average inference time of 5.98 s, an improvement of 32.2%. This highlights Mamba's significantly enhanced efficiency in practical applications.

As shown in Fig. 11, we compared the GPU memory usage and training speed of Mamba and Transformer during training. The experimental results indicate that Mamba achieves faster training while consuming less GPU memory. Specifically, Mamba's GPU memory usage is 4.06 GB, whereas Transformer requires 5.86 GB, representing a reduction of approximately 30.7%, highlighting Mamba's significant advantage in memory efficiency. Moreover, Mamba also outperforms Transformer in terms of training efficiency. After 19 training epochs, the average training time for Mamba is 28.13 s, compared to 38.12 s for Transformer, reflecting a 26.21% improvement. These findings demonstrate the advantages of Mamba in efficient training and resource utilization, providing valuable insights for optimiz ing DL models.

Ablation of Mamba layers

To investigate the impact of Mamba blocks, we explored how different numbers of Mamba blocks behave in terms of model performance.

Figure 12 and Table 9 illustrate the performance variations of different numbers of Mamba blocks across various metrics on the validation set. As the number of Mamba blocks increases, Acc-2 and F1 scores show a significant improvement. However, after reaching 3 Mamba blocks, the performance gain starts to diminish, and in some cases, the improvement even levels off. Acc-7 follows a similar trend, and MAE achieves the best results with 6 Mamba blocks.

Based on the above results, we identified that the optimal number of Mamba blocks lies between 3 and 6. With this range, we can achieve the best overall performance. However, employing 6 Mamba blocks, while providing a slight advantage over 3 Mamba blocks, requires significantly more computational memory. Therefore, for practical applications, balancing computational efficiency and performance, selecting an appropriate number of Mamba blocks is crucial.

Ablation of SSM

To investigate the impact of the SSM module, we conducted ablation experiments, as shown in Fig. 13, by removing the SSM module and comparing the performance before and after its removal. This analysis aims to explore the role of the SSM module in MER. Previous work [60] has examined whether the SSM module significantly contributes to visual tasks; therefore, we extend this investigation to MER.

By comparing the data in Table 10 and Fig. 14, it is evident that the inclusion of the SSM module significantly improves the model's performance across various metrics. Specifically, Acc-2 and F1-score both increase by approximately 1%, Acc-7 increases by 1.45%, and MAE decreases by about 0.0169. This indicates that the SSM module plays a crucial role in enhancing model accuracy and reducing error.

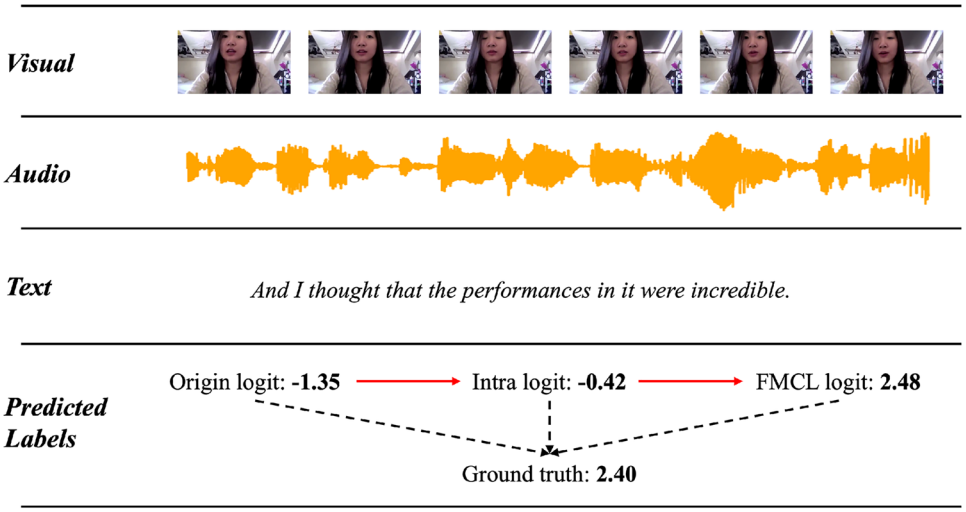


Fig. 16 Case of positive sentiment

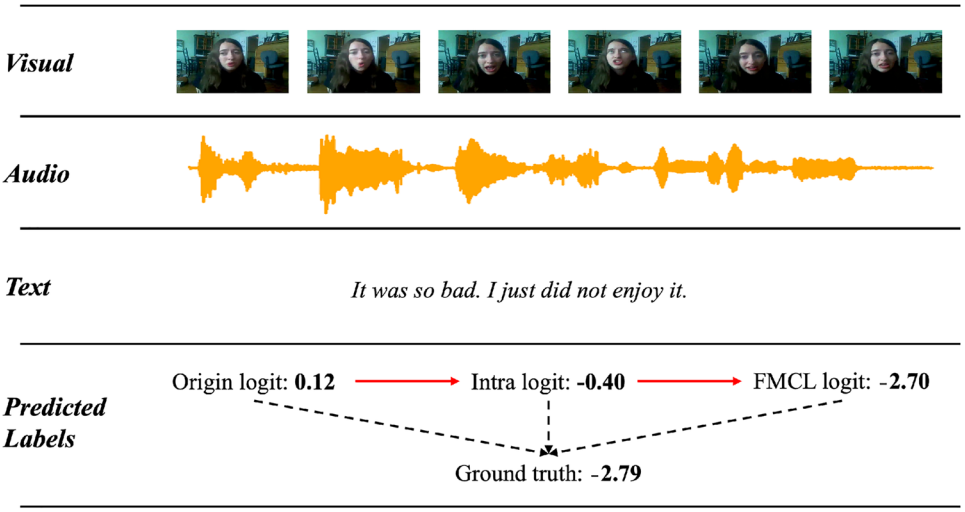


Fig. 17 Case of negative sentiment

Further analysis reveals that the SSM module effectively captures the similarity between multimodal data, improving feature extraction accuracy, and thus yielding better performance in emotion recognition tasks. Therefore, retaining the SSM module is essential to ensure the model's efficiency and accuracy in practical applications.

Case study

This section validates the FMCL at the example level. We selected two examples from CMU-MOSI for positive and negative emotions. As shown in Fig. 15, we perform emotion prediction in different parts of the whole framework. Specifically, the original logit is obtained by simply fusing the original features first, then the Intra logit is obtained by fusing the three modalities from the Intra-Mamba posterior, and finally the FMCL logit is obtained from the full model.We compare with the real labels in the examples.

As shown in Figs. 16 and 17, these two emotion recognition examples demonstrate the performance of the model in recognising positive and negative emotions. The text in Fig. 16 is a sentence with positive sentiment, indicating that the person holds a positive evaluation of the performance; the text in Fig. 17 is a sentence with negative sentiment, indicating that the person holds a negative evaluation of something. By comparing the Origin logit, Intra logit and FMCL logit with Ground truth, it can be seen that the FMCL logit is closer to Ground truth, which shows that it is more accurate in emotion recognition, and more importantly, it shows the effectiveness of the various modules of FMCL.

Conclusion

This paper proposes a novel emotion recognition algorithm (FMCL) based on multimodal fusion Mamba and contrastive learning. To effectively extract different features, we use

Intra-Mamba to enhance unimodal representations. Simultaneously, to achieve comprehensive information fusion, we extend the Mamba block to accommodate dual inputs, creating a new module called the Two-Fusion Mamba block and design a tri-modal, tri-branch fusion architecture on this basis. Additionally, we introduce contrastive learning and interaction-level auxiliary classification to further enhance the performance. Experimental results on three public datasets demonstrate the superiority of our method. In future work, we will explore more feasible schemes for Mamba.

Discussion on real-world applicability

In practical applications, the FMCL model, by integrating information from video, audio, and text modalities, demonstrates outstanding potential for applicability across multiple fields. For instance, in health monitoring, the model can capture patients' facial expressions from video data, extract emotional fluctuations from audio signals, and combine text data (e.g., medical records or self-reported symptoms) to provide robust support for comprehensive health evaluations. In intelligent customer service, the FMCL model can identify customers' identities and emotional states through video analysis, quickly interpret customer needs via speech recognition, and integrate text-based information to deliver more precise solutions, thereby enhancing the overall customer experience. In educational settings, video data can be used to capture students' classroom performance, audio data can record teacher-student interactions, and text data can facilitate the analysis of assignments and feedback, providing a basis for personalized teaching and real-time instructional adjustments. Furthermore, on social media platforms, by comprehensively analyzing multimodal information from posts related to emerging events, the FMCL model can integrate various data types to promptly reflect public sentiment toward an incident, offering valuable insights for public opinion monitoring and the formulation of response strategies. Overall, the FMCL model's strengths in multimodal information fusion not only enhance task efficiency across diverse domains but also lay a solid theoretical and practical foundation for future cross-domain applications and system integration.

Acknowledgements This study was supported by the GPU cluster provided by the CUC public computing cloud.

Author Contributions Qianjun Shuai conducted the experiments and analyzed the data. Xiaohao Chen designed the research and wrote the manuscript. Feng Hu participated in data collection and interpretation. Yongqiang Cheng provided guidance and revised the manuscript.

Funding This work was supported by the National Key R&D Program of China (Grant No. 2024YFF0907401).

Data availability All data are from open-source datasets and can be obtained upon request.

Declarations

Conflict of interest The authors declare no potential conflict of interest.

References

1. Singh U, Abhishek K, Azad HK (2024) A survey of cutting-edge multimodal sentiment analysis. ACM Comput Surv 56(9):1–38
2. Ma H, Wang J, Lin H, Zhang B, Zhang Y, Xu B (2023) A transformer-based model with self-distillation for multimodal emotion recognition in conversations. IEEE Trans Multimed 33:5467–5479
3. He L, Wang Z, Wang L, Li F (2023) Multimodal mutual attention-based sentiment analysis framework adapted to complicated contexts. IEEE Trans Circuits Syst Video Technol 15:2096–2109
4. Mai S, Xing S, Hu H (2021) Analyzing multimodal sentiment via acoustic-and visual-lstm with channel-aware temporal convolution network. IEEE/ACM Trans Audio Speech Lang Process 29:1424–1437
5. He J, Shi X, Li X, Toda T (2024) MF-AED-AEC: speech emotion recognition by leveraging multimodal fusion, asr error detection, and asr error correction. arXiv:2401.13260
6. Han W, Chen H, Gelbukh A, Zadeh A, Morency L-p, Poria S (2021) Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In: Proceedings of the 2021 international conference on multimodal interaction, pp 6–15
7. Shao R, Wu T, Liu Z (2023) Detecting and grounding multi-modal media manipulation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6904–6913
8. Peng S, Zhu X, Deng H, Lei Z, Deng L-J (2024) Fusionmamba: efficient image fusion with state space model. arXiv:2404.07932
9. Gu A, Dao T (2023) Mamba: linear-time sequence modeling with selective state spaces. arXiv:2312.00752
10. Patro BN, Agneeswaran VS (2024) Mamba-360: survey of state space models as transformer alternative for long sequence modelling: methods, applications, and challenges. arXiv:2404.16112
11. Yue Y, Li Z (2024) Medmamba: vision mamba for medical image classification. arXiv:2403.03849
12. Yang Z, Zhang J, Wang G, Kalra MK, Yan P (2024) Cardiovascular disease detection from multi-view chest X-rays with bi-mamba. arXiv:2405.18533

13. Li Z, Pan H, Zhang K, Wang Y, Yu F (2024) Mambadfuse: a mamba-based dual-phase model for multi-modality image fusion. arXiv:2404.08406
14. Shams S, Dindar SS, Jiang X, Mesgarani N (2024) SSAMBA: self-supervised audio representation learning with mamba state space model. arXiv:2405.11831
15. Wang X, Wang S, Ding Y, Li Y, Wu W, Rong Y, Kong W, Huang J, Li S, Yang H et al (2024) State space model for new-generation network alternative to transformers: a survey. arXiv:2404.09516
16. Liu Z, Shen Y, Lakshminarasimhan VB, Liang PP, Zadeh A, Morency L-P (2018) Efficient low-rank multimodal fusion with modality-specific factors. arXiv:1806.00064
17. Zadeh A, Chen M, Poria S, Cambria E, Morency L-P (2017) Tensor fusion network for multimodal sentiment analysis. arXiv:1707.07250
18. Zadeh A, Liang PP, Mazumder N, Poria S, Cambria E, Morency L-P (2018) Memory fusion network for multi-view sequential learning. In: Proceedings of the AAAI conference on artificial intelligence, vol 32
19. Li Y, Wang Y, Cui Z (2023) Decoupled multimodal distilling for emotion recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6631–6640
20. Wang H, Luo S, Hu G, Zhang J (2024) Gradient-guided modality decoupling for missing-modality robustness. arXiv:2402.16318
21. Zhou Y, Liang X, Chen H, Zhao Y (2024) Triple disentangled representation learning for multimodal affective analysis. arXiv:2401.16119
22. Dai W, Li X, Hu P, Wang Z, Qi J, Peng J, Zhou Y (2024) MInD: improving multimodal sentiment analysis via multimodal information disentanglement. arXiv:2401.11818
23. Liang T, Lin G, Feng L, Zhang Y, Lv F (2021) Attention is not enough: mitigating the distribution discrepancy in asynchronous multimodal sequence fusion. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 8148–8156
24. Sun H, Liu J, Chen Y-W, Lin L (2023) Modality-invariant temporal representation learning for multimodal sentiment classification. Inf Fusion 91:504–514
25. Shi H, Pu Y, Zhao Z, Huang J, Zhou D, Xu D, Cao J (2024) Co-space representation interaction network for multimodal sentiment analysis. Knowl Based Syst 283:111149
26. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30:5434–5446
27. Tsai Y-HH, Bai S, Liang PP, Kolter JZ, Morency L-P, Salakhutdinov R (2019) Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the conference. Association for computational linguistics. Meeting, NIH Public Access, vol 2019, p 6558
28. Gu A, Goel K, Ré C (2021) Efficiently modeling long sequences with structured state spaces. arXiv:2111.00396
29. Smith JT, Warrington A, Linderman SW (2022) Simplified state space layers for sequence modeling. arXiv:2208.04933
30. Fu DY, Dao T, Saab KK, Thomas AW, Rudra A, Ré C (2022) Hungry hungry hippos: towards language modeling with state space models. arXiv:2212.14052
31. Zhu L, Liao B, Zhang Q, Wang X, Liu W, Wang X (2024) Vision mamba: efficient visual representation learning with bidirectional state space model. arXiv:2401.09417
32. Liu Y, Tian Y, Zhao Y, Yu H, Xie L, Wang Y, Ye Q, Liu Y (2024) Vmamba: visual state space model. arXiv:2401.10166
33. He X, Cao K, Yan K, Li R, Xie C, Zhang J, Zhou M (2024) Pan-mamba: effective pan-sharpening with state space model. arXiv:2402.12192
34. Li H, Hu Q, Yao Y, Yang K, Chen P (2024) CFMW: cross-modality fusion mamba for multispectral object detection under adverse weather conditions. arXiv:2404.16302
35. Erol MH, Senocak A, Feng J, Chung JS (2024) Audio mamba: bidirectional state space model for audio representation learning. arXiv:2406.03344
36. Li W, Zhou H, Yu J, Song Z, Yang W (2024) Coupled Mamba: enhanced multi-modal fusion with coupled state space model. arXiv:2405.18014
37. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J et al (2021) Learning transferable visual models from natural language supervision. In: International conference on machine learning. PMLR, pp 8748–8763
38. Zhu B, Lin B, Ning M, Yan Y, Cui J, Wang H, Pang Y, Jiang W, Zhang J, Li Z et al (2023) Languagebind: extending video-language pretraining to n-modality by language-based semantic alignment. arXiv:2310.01852
39. Akbari H, Yuan L, Qian R, Chuang W-H, Chang S-F, Cui Y, Gong B (2021) Vatt: transformers for multimodal self-supervised learning from raw video, audio and text. Adv Neural Inf Process Syst 34:24206–24221
40. Liu R, Zuo H, Lian Z, Schuller BW, Li H (2024) Contrastive learning based modality-invariant feature acquisition for robust multimodal emotion recognition with missing modalities. IEEE Trans Affect Comput 30:5998–6008
41. Zadeh A, Zellers R, Pincus E, Morency L-P (2016) Multimodal sentiment intensity analysis in videos: facial gestures and verbal messages. IEEE Intell Syst 31(6):82–88
42. Zadeh AB, Liang PP, Poria S, Cambria E, Morency L-P (2018) Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers), pp 2236–2246
43. Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543
44. Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805
45. Baltrušaitis T, Robinson P, Morency L-P (2016) Openface: an open source facial behavior analysis toolkit. In: 2016 IEEE winter conference on applications of computer vision (WACV). IEEE, pp 1–10
46. Degottex G, Kane J, Drugman T, Raitio T, Scherer S (2014) Covarep—a collaborative voice analysis repository for speech technologies. In: 2014 IEEE international conference on acoustics, speech and signal processing (icassp). IEEE, pp 960–964
47. Wang Y, Shen Y, Liu Z, Liang PP, Zadeh A, Morency L-P (2019) Words can shift: dynamically adjusting word representations using nonverbal behaviors. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, pp 7216–7223
48. Pham H, Liang PP, Manzini T, Morency L-P, Póczos B (2019) Found in translation: learning robust joint representations by cyclic translations between modalities. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, pp 6892–6899
49. Hazarika D, Zimmermann R, Poria S (2020) Misa: modality-invariant and-specific representations for multimodal sentiment analysis. In: Proceedings of the 28th ACM international conference on multimedia, pp 1122–1131
50. Thi N-HN, Le D-T, Ha QT et al (2023) Self-MI: efficient multimodal fusion via self-supervised multi-task learning with auxiliary mutual information maximization. In: Proceedings of the 37th Pacific Asia conference on language, information and computation, pp 582–590
51. Wang L, Peng J, Zheng C, Zhao T et al (2024) A cross modal hierarchical fusion multimodal sentiment analysis method based on multi-task learning. Inf Process Manag 61(3):103675

52. Zhou M, Quan W, Zhou Z, Wang K, Wang T, Yan D-M (2024) TCAN: text-oriented cross attention network for multimodal sentiment analysis. arXiv:2404.04545
53. Wang P, Zhou Q, Wu Y, Chen T, Hu J (2024) DLF: disentangled-language-focused multimodal sentiment analysis. arXiv:2412.12225
54. Ai W, Shou Y, Meng T, Li K (2024) DER-GCN: dialog and event relation-aware graph convolutional neural network for multimodal dialog emotion recognition. IEEE Trans Neural Netw Learn Syst
55. Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9(11):5467–5479
56. Rahman W, Hasan MK, Lee S, Zadeh AB, Mao C, Morency L-P, Hoque E (2020) Integrating multimodal information in large pretrained transformers. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 2359–2369
57. Yu W, Xu H, Yuan Z, Wu J (2021) Learning modality-specific representations with self-supervised multitask learning for multimodal sentiment analysis. In: Proceedings of the AAAI conference on artificial intelligence, vol 35, pp 10790–10797
58. Han W, Chen H, Poria S (2021) Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In: Proceedings of the 2021 conference on empirical methods in natural language processing, pp 9180–9192
59. Lin R, Hu H (2023) Missmodal: increasing robustness to missing modality in multimodal sentiment analysis. Trans Assoc Comput Linguist 11:1686–1702
60. Yu W, Wang X (2024) MambaOut: do we really need mamba for vision? arXiv:2405.07992