MSAmba: Exploring Multimodal Sentiment Analysis with State Space Models

Xilin He[1*], Haijian Liang[1*], Boyi Peng[1], Weicheng Xie[1,2,3†], Muhammad Haris Khan[4], Siyang Song[5], Zitong Yu[6]

[1]Computer Vision Institute, School of Computer Science & Software Engineering, Shenzhen University,
[2]Shenzhen Institute of Artifcial Intelligence and Robotics for Society, [3]Guangdong Key Laboratory of Intelligent Information Processing, [4]Mohamed bin Zayed University of Artificial Intelligence,
[5]University of Exeter, [6]Great Bay University

Abstract

Multimodal sentiment analysis, which learns a model to process multiple modalities simultaneously and predict a sentiment value, is an important area of affective computing. Modeling sequential intra-modal information and enhancing cross-modal interactions are crucial to multimodal sentiment analysis. In this paper, we propose MSAmba, a novel hybrid Mamba-based architecture for multimodal sentiment analysis, consisting of two core blocks: Intra-Modal Sequential Mamba (ISM) block and Cross-Modal Hybrid Mamba (CHM) block, to comprehensively address the above-mentioned challenges with hybrid state space models. Firstly, the ISM block models the sequential information within each modality in a bi-directional manner with the assistance of global information. Subsequently, the CHM blocks explicitly model centralized cross-modal interaction with a hybrid combination of Mamba and attention mechanism to facilitate information fusion across modalities. Finally, joint learning of the intra-modal tokens and cross-modal tokens is utilized to predict the sentiment values. This paper serves as one of the pioneering works to unravel the outstanding performances and great research potential of Mamba-based methods in the task of multimodal sentiment analysis. Experiments on CMU-MOSI, CMU-MOSEI and CH-SIMS demonstrate the superior performance of the proposed MSAmba over prior Transformer-based and CNN-based methods. Code is available at here.

Introduction

Sentiment analysis is an important area of affective computing, which focuses on recognizing the sentiment values of humans from input data, and has a wide range of applications in human-computer interaction and health care services (Jiang et al. 2020; Ezzameli and Mahersia 2023). With the widespread popularity of online social media platforms such as Instagram, TikTok, and Facebook, videos containing multiple modalities have become major carriers of information. This development introduces new challenges for processing and analyzing such diverse data, thereby raising research attention on the topic of multimodal sentiment analysis (MSA), which aims to learn a model to process multiple modalities simultaneously and predict a sentiment value (Cambria et al. 2013; Kaur and Kautish 2022).

Most recent MSA methods can be categorized into two categories: representation learning methods (Hazarika, Zimmermann, and Poria 2020; Yang et al. 2022; Li, Wang, and Cui 2023; Li et al. 2023; Yang et al. 2023; Yang, Dong, and Qiang 2024) and multimodal fusion methods (Zadeh et al. 2017; Liu et al. 2018; Tsai et al. 2019a; Huang et al. 2020; Zhang et al. 2023). Representation learning methods (Hazarika, Zimmermann, and Poria 2020; Yang et al. 2022, 2023; Yang, Dong, and Qiang 2024) concentrate on extracting refined modality semantics that is rich in human sentiment clues, which enhances the efficiency of multimodal fusion, thereby improving relationship modeling. On the other hand, the multimodal fusion methods (Huang et al. 2020; Zhang et al. 2023) explore the design of sophisticated fusion mechanisms to achieve a joint representation of multimodal data. Despite the prior methods demonstrating the importance of sequential modeling and cross-modal interaction for MSA, from a model designing perspective, the above-mentioned two types of methods mainly rely on the abundant utilization of Transformers or convolution neural networks to achieve either long-range sequential modeling or cross-modal interaction, leading to an excessive number of parameters and low inference speed.

Recently, Mamba (Gu and Dao 2023), a novel architecture and low-cost operator based on state space models (SSMs) (Gu et al. 2021; Gu, Goel, and Ré 2022), has emerged as a promising alternative to the Transformer model, exhibiting better long-range dependencies modeling capacity and superior performance over Transformer models in various domains, including language modeling (Gu and Dao 2023; Grazzi et al. 2024; Lieber et al. 2024), point cloud analysis (Zhang et al. 2024; Liang et al. 2024), video understanding (Li et al. 2024; Chen et al. 2024), medical image processing (Ruan and Xiang 2024; Xing et al. 2024; Ma, Li, and Wang 2024) and graph reasoning (Behrouz and Hashemi 2024; Wang et al. 2024). However, existing research on Mamba has primarily focused on single-modality learning, leaving multimodal fusion with Mamba unexplored, which is crucial for the task of MSA. Meanwhile, a general paradigm of adapting Mamba into MSA is yet to be discussed.

To address the above-mentioned challenges and bridge the gap of Mamba research on MSA, in this paper, we

[*]These authors contributed equally.

[†]Corresponding author

propose MSAmba, a novel Mamba-based MSA architecture featuring two key components: Intra-Modal Sequential Mamba (ISM) blocks and Cross-modal Hybrid Mamba (CHM) blocks. Specifically, ISM can effectively model sequential information with the assistance of extracted global context in a bi-directional manner while CHM is designed to conduct cross-modal interaction with a centralized modality guiding the fusion process. Thereby, with ISM enhancing sequential representation learning and CHM conducting cross-modal interaction, MSAmba not only achieves superior performances on various datasets of MSA, but also obtains an efficient computational overhead.

We evaluate MSAmba on three standard datasets for MSA, namely CMU-MOSI (Zadeh et al. 2016), CMU-MOSEI (Zadeh et al. 2018) and CH-SIMS (Yu et al. 2020). Quantitative results demonstrate that MSAmba outperforms previous state-of-the-arts (SOTAs), which rely heavily on the abundant utilization of cross-attention transformers, in terms of performance, parameter size and inference speed. Further, serving as a pioneering research of adapting Mamba-based architecture in MSA, we observe that simply applying vanilla Mamba in a naive paradigm could also achieve comparative results against previous state-of-the-arts (SOTAs), demonstrating the great research potential of Mamba-based architectures in MSA. Our main contributions are summarized as three-fold:

- We propose MSAmba, a novel Mamba-based architecture for MSA, which could effectively model intra-modal sequential information and centralized cross-modal interaction.
- We introduce an effective ISM block and a novel CHM block to assist intra-modal sequential modeling with global context in a bi-directional manner and achieve centralized cross-modal interaction in MSA, respectively.
- Experimental results on three benchmark datasets demonstrate the superiority of MSAmba over prior Transformer-based methods and the great research potential of Mamba-based architectures in MSA.

Related Work

Multimodal Sentiment Analysis

Multimodal sentiment analysis aims to predict a sentiment score from the language, visual and audio information embedded in the video clips. Mainstream MSA approaches could be divided into two main categories: representation learning-based (Hazarika, Zimmermann, and Poria 2020; Yang et al. 2022; Li, Wang, and Cui 2023; Li et al. 2023; Yang et al. 2023; Yang, Dong, and Qiang 2024) and cross-modal interaction-based (Zadeh et al. 2017; Liu et al. 2018; Tsai et al. 2019a; Huang et al. 2020; Zhang et al. 2023).

The former concentrates on extracting refined modality-specific representations. Early works (Hazarika, Zimmermann, and Poria 2020; Yang et al. 2022) argued representation learning for multiple modalities as a domain adaption task and thereby leveraging metric learning and adversarial learning to learn the modality-invariant and modality-specific information for multimodal fusion. (Han, Chen, and Poria 2021) proposed MMIM to improve representation learning with hierarchical mutual information maximization. (Yang et al. 2023) proposed ConFEDE to decompose modality features into similar and dissimilar ones with contrastive learning.

Recently, the multimodal fusion-based methods (Liu et al. 2018; Tsai et al. 2019a; Huang et al. 2020; Zhang et al. 2023), building upon cross-modal attention-based Transformers, have driven the development of MSA since they applied cross-modal interactions to obtain enhanced modality representations. For example, (Tsai et al. 2019a) proposed the multimodal transformer that consists of a cross-modal attention mechanism to learn the potential adaption and correlations from one modality to another, thereby achieving semantic alignment between modalities. (Zhang et al. 2023) proposed an adaptive hyper-modality transformer to guide the visual and audio representation learning with language modality. However, despite gaining significant performance improvement, the majority of existing methods have utilized attention-based transformers to model intra-modal dependency and conduct cross-modal interactions, resulting in an excessive amount of parameters and low inference speed, thereby limiting the model deployment.

State Space Models

State space models (SSMs) (Gu et al. 2021; Gu, Goel, and Ré 2022) are a mathematical representation of dynamic systems, which models the input-output relationship with a hidden state. Being a general architecture, SSMs have achieved great success across a wide range of applications including computational neuroscience (Friston, Harrison, and Penny 2003) and linear dynamical systems (Hespanha 2018). Recently, SSMs have been introduced as an alternative architecture to model long-range dependency with the guidance of global context. Compared with attention-based transformer models, which require the quadratic complexity of the sequence length, SSMs are more prone to be more efficient.

Various structures have been proposed recently to improve the representation ability and efficiency of SSMs. (Gu, Goel, and Ré 2022) propose structured state space models (S4) to improve computational efficiency, where the state matrix of SSMs is a sum of low-rank matrices. A line of subsequent studies attempts to further enhance the effectiveness of S4. For example, (Smith, Warrington, and Linderman 2023) introduced S5 utilizing MIMO-SSM and parallel scan technology.

Recently, (Gu and Dao 2023) proposed Mamba with an input-dependent selection mechanism based on S4, which achieves linear scaling in sequence length and demonstrates superior performance over Transformers on various benchmarks, including video understanding (Li et al. 2024; Chen et al. 2024), medical image processing (Xing et al. 2024; Ruan and Xiang 2024) and graph prediction (Behrouz and Hashemi 2024; Wang et al. 2024). However, adapting Mamba to the MSA task remains an open problem and is yet to be explored. Further, existing studies (Chen et al. 2024; Li et al. 2024) on Mamba mainly focus on single-modality modeling, while cross-modal interaction with Mamba-based

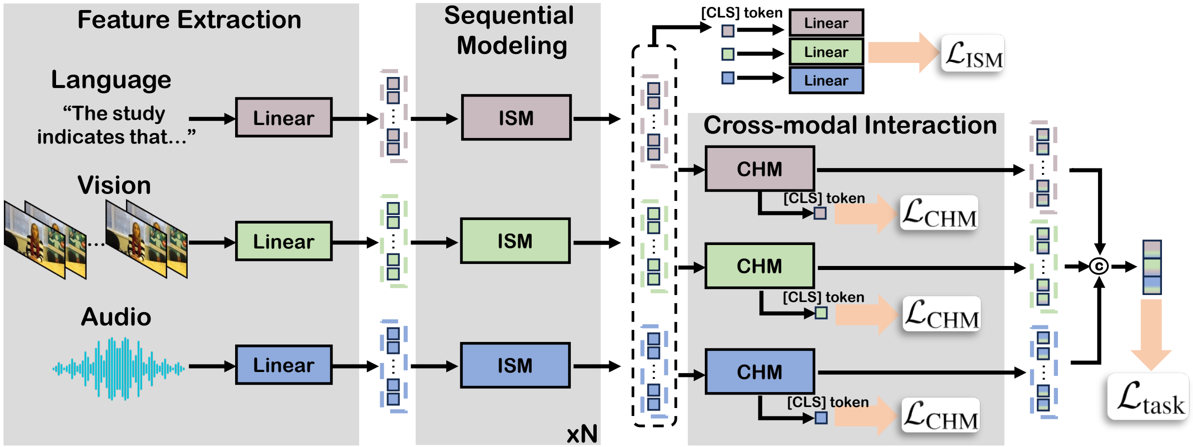

Figure 1: Overview of the MSAmba, which mainly consists of Intra-Modal Sequential Mamba (ISM) blocks and Cross-modal Hybrid Mamba (CHM) blocks.

architecture has not been widely investigated. In this paper, we thus propose MSAmba, a general paradigm to adapt Mamba for MSA with sequential modeling and cross-modal interaction. Beyond the superior performance of MSAmba compared to Transformer models, we also demonstrate the significant research potential of Mamba in MSA.

Methods

Preliminaries

In recent years, the state space models (SSMs) have developed rapidly (Gu and Dao 2023; Gu, Goel, and Ré 2022; Gu et al. 2021). Originating from the linear system theory (Hespanha 2018), Mamba introduced a selective scanning mechanism based on S4(Gu, Goel, and Ré 2022). Based on the concept of continuous systems, Mamba map input sequence $x(t) \in \mathbb{R}$ to output sequence $y(t) \in \mathbb{R}$ though a hidden state $\mathbf{h} \in \mathbb{R}^N$ by a linear ordinary differential equation:

$$\mathbf{h}'(t) = \mathbf{A}\mathbf{h}(t) + \mathbf{B}x(t), \tag{1}$$

$$y(t) = \mathbf{C}^\top \mathbf{h}(t), \tag{2}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$ is the state matrix, $\mathbf{B} \in \mathbb{R}^N$ is the input matrix, $\mathbf{C} \in \mathbb{R}^N$ is the output matrix. Eq. (1) defines the evolution of the hidden state $\mathbf{h}(t)$, while Eq. (2) determines the output comes from a linear transformation of the hidden state $\mathbf{h}(t)$.

Since the continuous-time system is not suitable for digital computers and real-world data, Mamba uses the zero-order hold method to discretize the continuous parameters $\mathbf{A}, \mathbf{B}$ into $\overline{\mathbf{A}}, \overline{\mathbf{B}}$ with a time scale parameter Δ:

$$\overline{\mathbf{A}} = \exp\left(\mathbf{\Delta A}\right), \quad \overline{\mathbf{B}} = \left(\mathbf{\Delta A}\right)^{-1}\left(\exp\left(\mathbf{\Delta A}\right) - \mathbf{I}\right) \cdot \mathbf{\Delta B}. \tag{3}$$

Mamba introduces an input-dependent selection mechanism to allow the system to select relevant information based on the input sequence, which is achieved by making $\overline{\mathbf{B}}, \overline{\mathbf{C}}$ and $\overline{\Delta}$ as functions of the input sequence. Specifically, given an input sequence $\mathbf{x} \in \mathbb{R}^{B\times L\times D}$ where B is the batch size, L is the sequence length, and D is the feature dimension, the input-dependent parameters $\overline{\mathbf{B}}$ and $\overline{\mathbf{C}}$ are computed with learnable linear transformations on $\mathbf{x}$. The time scale parameter Δ is computed with a softplus between learnable time scale parameter $\tilde{\Delta}$ and a linear transformation on $\mathbf{x}$. After computing the discretized parameters, the output sequence $\mathbf{y} \in \mathbb{R}^{B\times L\times D}$ is computed based on Eq. (1) and Eq. (2).

Architecture Overview

An overview of the proposed MSAmba is illustrated in Fig. 1, which is a novel Mamba-based MSA framework consisting of two core components of ISM and CHM respectively designed for sequential information modeling and cross-modal interaction. Firstly, given multimodal input data, primary features are extracted by pre-trained models following previous works (Zhang et al. 2023). Subsequently, these modality-specific features are fed into stacked ISMs for sequential information modeling, resulting in a set of intra-modal class tokens. Then, features processed by ISMs are passed into CHMs for cross-modal interactions, producing cross-modal class tokens. Finally, both the intra-modal and cross-modal class tokens are integrated as the final features for sentiment value prediction.

Multimodal Input

Given the multimodal input consisting of three modalities: language, audio and video, following previous studies, precomputed sequential features $\mathbf{x}^m \in \mathbb{R}^{L^m\times D^m}$ are extracted with BERT (Devlin et al. 2019), Librosa (McFee et al. 2015) and OpenFace (Baltrušaitis, Robinson, and Morency 2016), respectively. $m = \{A, V, L\}$ denotes the three modalities of audio, video and language. L^m and D^m denote the sequence length and the feature dimension of the m modality features.

Sequential Information Modeling

Given sequential feature $\mathbf{x}^m$ from modality m, linear projections are applied for dimension alignment across various

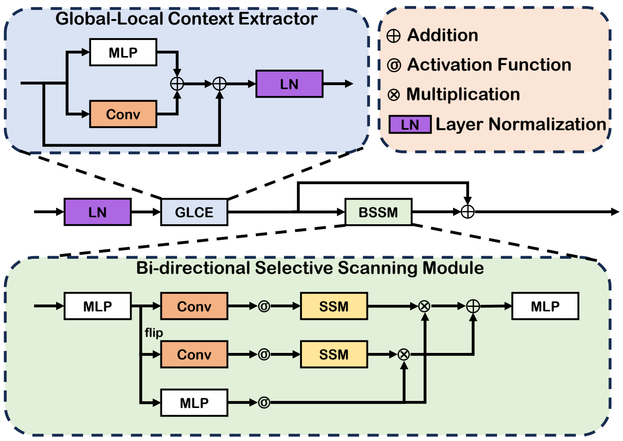

Figure 2: Illustration of the Intra-modal Sequential Mamba (ISM) block, which mainly consists of a Global-Local Context Extractor (GLCE) and a Bi-directional Selective Scanning Module (BSSM). *flip* indicates reversing the input on the temporal dimension.

modalities. Then, a learnable intra-modal class token is integrated with the modality-specific feature as:

$$\mathbf{x}^m = [\mathbf{x}^m_{\text{CLS}}, \text{Linear}(\mathbf{x}^m)] + p^m_s \tag{4}$$

where $\mathbf{x}^m_{\text{CLS}}$ and p^m_s denote the intra-modal class token and the spatial position embedding for modality m, respectively.

The modality-specific embedding is then fed into Mamba blocks for sequential information modeling. However, vanilla Mamba architecture cannot be directly applied to spatial-temporal input features. Thereby, we introduce ISM, which learns to model intra-modal sequential information with the assistance of global-local sequential context in a bi-directional manner. The overview of ISM is illustrated in Fig. 2.

The ISM block consists of a layer normalization, a global-local context extractor (GLCE), a bi-directional selective scanning module (BSSM) to model sequential information and a residual connection. Formally, given an input feature $\mathbf{x}^m$, the feature first undergoes layer normalization. Subsequently, the GLCE extracts both temporal global and local representations by applying linear projection and convolution operations in parallel, as follows:

$$\mathbf{c}^m_g, \mathbf{c}^m_l = \text{Linear}(\mathbf{x}^{m\top}), \text{Conv}(\mathbf{x}^{m\top}) \tag{5}$$

where $\mathbf{c}^m_g$ and $\mathbf{c}^m_l$ denote the sequentially global and local context extracted from the m modality, respectively. $\top$ denotes the transpose operation. The global and local contexts are then added with $\mathbf{x}$:

$$\mathbf{x} = \mathbf{x} + \text{LN}(\mathbf{c}^m_g + \mathbf{c}^m_l) \tag{6}$$

After combining both the global and local contexts, the resulting feature is then fed into the Mamba-based feature extractor to model long-range sequential information. To adapt Mamba-based architecture into spatial-temporal sequential information, inspired by (Li et al. 2024), we introduce the BSSM. The BSSM employs a spatial-temporal selective scanning mechanism, which involves organizing spatial tokens based on their location and then scanning them sequentially. To further enhance sequential information modeling, a parallel backward branch for selective scanning is incorporated, enabling bi-directional modeling of sequential information. After processing by the BSSM, a residual connection is applied, resulting in the extraction of the intra-modal class tokens, which would be leveraged in the latter process for sentiment value prediction.

Centralized Cross-Modal Interaction

Prior methods (Tsai et al. 2019a; Huang et al. 2020) demonstrate the effectiveness of cross-modal interaction in MSA as cross-modal interaction facilitates the learning of more complex and meaningful representations by capturing the dependencies and correlations between modalities. With Mamba serving as an effective low-cost operator and cross-modal interaction with Mamba being unexplored, in this section, we introduce CHM, a hybrid Mamba block mixing with self-attention to achieve centralized cross-modal interaction for MSA. An overview of the CHM block is shown in Fig. 3.

Given input features from multiple modalities, CHM is designed to inject semantic information from the centralized modality into the remaining modalities. This is achieved by modifying the BSSM block to cross-modal interaction. Specifically, given modality-specific features $\mathbf{x}^m$, where $m = \{A, V, L\}$, from the three modalities of audio, video and language, with language as the centralized modality as an example, we first concatenate the resting modality-specific features with the centralized modality feature along the spatial dimension following:

$$\mathbf{x}^{\text{AL}}, \mathbf{x}^{\text{VL}} = [\mathbf{x}^{\text{A}}, \mathbf{x}^{\text{L}}], [\mathbf{x}^{\text{V}}, \mathbf{x}^{\text{L}}] \tag{7}$$

Then, the concatenated features and the centralized modality feature are passed into the corresponding ISM blocks for primarily cross-modal interaction and sequential information modeling, respectively. Subsequently, the primarily cross-modal features are projected through a linear layer for dimension alignment with the centralized modality:

$$\mathbf{x}^{\text{AL}} = \text{Linear}(\text{BSSM}^{\text{AL}}(\mathbf{x}^{\text{AL}})), \;\; \mathbf{x}^{\text{L}} = \text{BSSM}^{\text{L}}(\mathbf{x}^{\text{L}}) \tag{8}$$

We then extract the first token of the centralized modality feature as its class token to further inject modality-specific semantic information into the cross-modal features as:

$$\mathbf{x}^{\text{AL}} = \mathbf{x}^{\text{AL}} + \text{CLS}(\mathbf{x}^{\text{L}}) \tag{9}$$

$\text{CLS}(\cdot)$ denotes the process of extracting the first token of the feature as its class token. To further enhance the cross-modal representation, CHM applies a hybrid architecture, which utilizes self-attention to further model the cross-modal features. With multimodal features being processed by CHM, a set of cross-modal class tokens would be generated for sentiment value prediction. Finally, the prediction could be generated by passing the integration of both the intra-modal class tokens and the cross-modal class tokens into a linear layer.

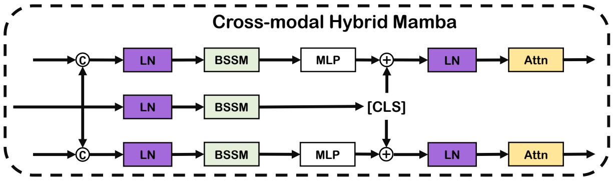

Figure 3: Illustration of the Cross-modal Hybrid Mamba block (CHM). It takes three modality-specific features as input and injects the semantic information from the class token of the centralized modality to the other two modalities. 'BSSM' and 'Attn' denote the bi-directional selective scanning module in ISM and the self-attention mechanism, respectively.

Training Loss

Given the final prediction $\hat{y}$ and the ground-truth sentiment value y, the model is trained with the L1 loss for regression tasks:

$$\mathcal{L}_{\text{task}} = |\hat{y} - y| \tag{10}$$

Further, to ensure semantic injection in ISM and CHM, the intra-modal class tokens, cross-modal class tokens and centralized modality class tokens in CHM are all leveraged to predict sentiment values, using corresponding auxiliary linear layers. These predictions are leveraged to calculate auxiliary losses as:

$$\begin{aligned}\mathcal{L}_{\text{aux}} &= \mathcal{L}_{\text{ISM}} + \mathcal{L}_{\text{CHM}} \\ &= \sum_{i=1}^{m} \left(\left|\hat{y}_i^{\text{intra}} - y\right| + \left|\hat{y}_i^{\text{cross}} - y\right| + \left|\hat{y}_i^{\text{central}} - y\right| \right)\end{aligned} \tag{11}$$

$\hat{y}_i^{\text{intra}}$, $\hat{y}_i^{\text{cross}}$ and $\hat{y}_i^{\text{central}}$ are predictions generated from the corresponding intra-modal class tokens, cross-modal tokens and centralized modality class tokens in CHM. Therefore, the overall training loss is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda \mathcal{L}_{\text{aux}} \tag{12}$$

Experiments

Datasets and Evaluation Metrics

We test the performance of MSAmba on three standard benchmark datasets: CMU-MOSI (Zadeh et al. 2016), CMU-MOSEI (Zadeh et al. 2018) and CH-SIMS (Yu et al. 2020).

On evaluation metrics, following prior works (Yang et al. 2023; Yang, Dong, and Qiang 2024; Zhang et al. 2023), we adopt binary classification accuracy (Acc-2), F1, seven classification accuracy (Acc-7), mean absolute error (MAE) and the correlation of the model's prediction with human (Corr). Further, on CMU-MOSI and CMU-MOSEI, following the protocol of prior works (Yang et al. 2023; Yang, Dong, and Qiang 2024; Zhang et al. 2023), Acc-2 and F1 are calculated in two ways: negative/non-negative and negative/positive.

Implementation Details

The hidden states dimension and the expansion coefficient of each Mamba block are set as 128 and 2, respectively. The numbers of ISM blocks and CHM blocks are set as 2 and 1 across various datasets, respectively. We use AdamW to optimize the model. We train the model for 200 epochs, with a batch size of 128. λ for balancing the training loss is set as 0.5 across all datasets. All the experiments are conducted on a single NVIDIA A100-80GB GPU. More details can be found in the appendix.

Baselines

To comprehensively validate the effectiveness of the proposed MSAmba, we make a fair comparison with a range of SOTAs, including MFM (Tsai et al. 2019b), LMF (Liu et al. 2018), TFN (Zadeh et al. 2017), MuLT (Tsai et al. 2019a), MISA (Hazarika, Zimmermann, and Poria 2020), MAG-BERT (Rahman et al. 2020), HyCon (Mai et al. 2022), Self-MM (Yu et al. 2021), ConFEDE (Yang et al. 2023), DMD (Li, Wang, and Cui 2023), ALMT (Zhang et al. 2023), DBF (Wu et al. 2023), UniMSE (Hu et al. 2022), HyDiscGAN (Wu et al. 2024), MCL-MCF (Fan et al. 2024), and CLGSI (Yang, Dong, and Qiang 2024).

Comparison with the State-of-The-Art

Tab. 1 and Tab. 2 present a performance comparison of MSAmba against state-of-the-art methods on the CMU-MOSI, CMU-MOSEI, and CH-SIMS datasets, respectively.

As shown in Tab. 1 and Tab. 2, MSAmba outperforms prior SOTAs across most metrics. Compared to prior representation learning-based approaches that focus on contrastive learning (Yang, Dong, and Qiang 2024; Yang et al. 2023), MSAmba achieves consistent improvements. Notably, MSAmba surpasses the latest contrastive learning method CLGIS (Yang, Dong, and Qiang 2024) with margins of 2.01% and 1.71% on binary classification (Acc-2) and fine-grained classification (Acc-7), respectively. Compared to methods that use attention-based Transformers for cross-modal interactions, MSAmba obtains significant advantages with fewer parameters and faster inference speeds, demonstrating the success of CHM as a cross-modal interaction operator. Concretely, MSAmba surpasses previous Transformer-based SOTA ALMT (Zhang et al. 2023) by 1.00% and 0.80% on the metrics of Acc-2 and F1 on CMU-MOSEI. The latter section provides a detailed efficiency study on MSAmba and prior MSA methods.

Efficiency Study

In this section, we provide an efficiency study, comparing the proposed MSAmba with prior Transformer-based methods in terms of parameter size and performances on CMU-MOSI. As shown in Tab. 3, MSAmba surpasses the previous SOTA Transformer-based method, ALMT (Zhang et al. 2023), by a modest yet significant 0.24% on the Acc-7 metric. This achievement is notable given that MSAmba operates with a parameter count that is around half that of ALMT.

Additionally, as the self-attention module in CHM takes up a certain degree of the parameter count, we remove the self-attention in each CHM to squeeze the parameter count further and validate the efficiency of the Mamba-based architecture. As shown in Tab. 3, despite removing the self-attention module leads to a drop of performance, it still

Method	CMU-MOSI					CMU-MOSEI				
	Acc-2↑	F1↑	Acc-7↑	MAE↓	Corr↑	Acc-2↑	F1↑	Acc-7↑	MAE↓	Corr↑
LF-DNN	77.52/78.63	77.46/78.63	34.52	0.955	0.658	80.60/82.74	80.85/82.52	50.83	0.58	0.709
MFM$_{\text{ICLR'19}}$	77.4/-	77.3/-	34.1	0.965	0.632	78.94/82.86	79.55/82.85	51.34	0.573	0.718
LMF$_{\text{ACL'18}}$	-/82.5	-/82.4	33.2	0.917	0.695	80.54/83.48	80.94/83.36	51.59	0.576	0.717
TFN$_{\text{EMNLP'17}}$	-/80.8	-/80.7	34.9	0.901	0.698	78.50/81.89	78.96/81.74	51.6	0.573	0.717
MulT$_{\text{ACL'19}}$	-/83.0	-/82.8	40	0.871	0.698	81.15/84.63	81.56/84.52	52.84	0.559	0.733
MISA$_{\text{ACM MM'20}}$	81.8/83.4	81.7/83.6	42.3	0.783	0.776	83.6/85.5	83.8/85.3	52.2	0.555	0.756
MAG-BERT$_{\text{ACL'20}}$	82.13/83.54	81.12/83.58	41.43	0.79	0.766	82.51/84.82	82.77/84.71	50.41	0.583	0.741
Self-MM$_{\text{AAAI'21}}$	83.44/85.46	83.36/85.43	46.67	0.708	0.796	83.76/85.15	83.82/84.90	53.87	0.531	0.765
HyCon$_{\text{TAC'22}}$	-/85.2	-/85.1	46.6	0.713	0.79	-/85.4	-/85.6	52.8	0.601	0.776
UniMSE$_{\text{EMNLP'22}}$	85.85/86.90	85.83/86.42	48.68	0.691	**0.809**	85.36/**87.50**	85.79/87.46	54.39	0.523	0.773
ConFEDE$_{\text{ACL'23}}$	84.17/85.52	84.13/85.52	42.27	0.742	0.784	81.65/85.82	82.17/85.83	**54.86**	0.522	0.78
DMD$_{\text{CVPR'23}}$	-/86.0	-/86.0	45.6	-	-	-/86.6	-/86.6	54.6	-	-
ALMT$_{\text{EMNLP'23}}$	84.55/86.43	84.57/86.47	49.42	**0.683**	0.805	84.78/86.79	85.19/86.86	54.28	0.526	0.779
DBF$_{\text{ACL'23}}$	85.1/86.9	85.1/86.9	44.8	0.693	0.801	84.3/86.4	84.8/86.2	54.2	0.523	0.772
HyDiscGAN$_{\text{IJCAI'24}}$	84.1/86.7	83.7/86.3	43.2	0.749	0.782	81.9/86.3	82.1/86.2	54.4	0.533	0.761
MCL-MCF$_{\text{TAC'24}}$	84.9/87.3	84.7/87.2	-	0.692	0.799	84.2/86.4	84.4/86.3	-	0.536	0.767
CLGSI$_{\text{NAACL'24}}$	83.97/86.43	83.63/86.25	47.96	0.703	0.79	84.01/86.32	84.21/86.18	54.56	0.532	0.763
MSAmba	**85.99/87.43**	**85.99/87.40**	**49.67**	0.707	**0.809**	**85.78**/86.86	**85.99/86.93**	54.21	**0.507**	**0.796**

Table 1: Comparison with the SOTAs on CMU-MOSI and CMU-MOSEI. The best results are labeled in bold.

Method	Acc-5↑	Acc-3↑	Acc-2↑	F1↑	MAE↓	Corr↑
TFN	39.30	65.12	78.38	78.62	0.432	0.591
LMF	40.53	64.68	77.77	77.88	0.441	0.576
MFM	-	-	75.06	75.58	0.477	0.525
MuLT	37.94	64.77	78.56	79.66	0.453	0.564
MISA	-	-	76.54	76.59	0.447	0.563
MAG-BERT	-	-	74.44	71.75	0.492	0.399
Self-MM	41.53	65.47	80.04	80.44	0.425	0.595
ALMT	45.73	**68.93**	81.19	81.57	0.404	0.619
ConFEDE	46.34	-	81.05	81.13	**0.377**	**0.655**
CLGSI	45.95	-	81.18	80.59	0.408	0.634
MSAmba	**47.17**	68.83	**82.30**	**81.75**	0.403	0.646

Table 2: Comparison with the SOTAs on CH-SIMS. The best results are labeled in bold.

achieves comparative performances against previous SOTAs, further validating the superior efficiency of the proposed MSAmba.

Method	Parameters↓	Acc-7↑
MuLT	2.57M	40.00
MISA	3.10M	42.3
MAG-BERT	1.22M	43.62
ALMT	2.49M	49.42
MSAmba (w/o self-attention)	**1.10M**	47.83
MSAmba	1.41M	**49.67**

Table 3: Efficiency comparison with SOTA Transformer-based methods on CMU-MOSI. Parameter sizes are computed with open-source code and default configs from the corresponding papers.

Ablation Study

In this section, we provide a detailed ablation study on each component of the proposed MSAmba to validate its effectiveness with both quantitative and qualitative results.

Module Design of ISM The core of the ISM block is the global-local context extract (GLCE) and the bi-directional selective scanning mechanism (BSSM). To validate the effectiveness of the ISM module design, we provide an ablation study on each module within ISM. The experiments are conducted on CMU-MOSI, where we remove the GLCE and replace the BSSM with a unidirectional state space model. Experiment results are shown in Tab. 4. As can be seen, both the absence of GLCE and BSSM results in performance degradation, validating the effectiveness of the module design in ISM. Notably, when BSSM is replaced with a unidirectional state space model, the performance drops significantly with a metric of 3.10% in Acc-7. This further demonstrates the importance of applying bi-directional scanning for sequential modeling.

Method	Acc-2↑	F1↑	Acc-7↑
w/o GLCE	85.26/86.80	85.17/87.21	48.92
BSSM → SSM	84.30/85.76	83.88/85.78	46.57
ISM	**85.99/87.43**	**85.99/87.40**	**49.67**

Table 4: Ablation study of ISM's module design on CMU-MOSI. w/o GLCE and BSSM → SSM denote removing the global-local context extractor and replacing the bi-directional selective scanning mechanism with a unidirectional state space model, respectively.

Sequential Modeling Capacity of ISM To demonstrate the superior sequential modeling capacity of ISM, we conduct experiments on CMU-MOSI, comparing against existing Mamba blocks (Vision Mamba (Zhu et al. 2024) and Video Mamba (Li et al. 2024)) and Transformer, where we replace the proposed ISM with corresponding blocks.

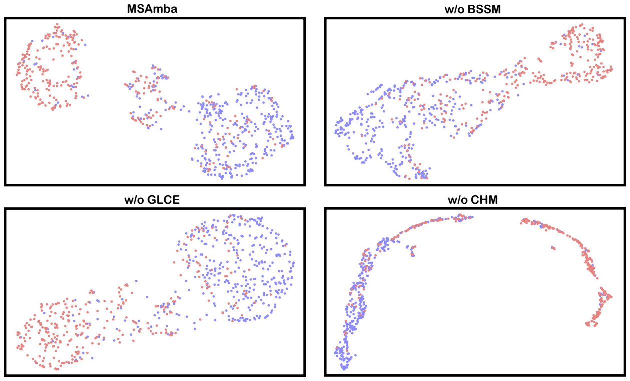

Figure 4: t-SNE visualizations on the integrated final tokens on the test-set of CMU-MOSI. Red and purple dots denote positive and negative samples, respectively.

Experiment results are shown in Tab. 5. As can be seen, compared with prior Mamba blocks (Vision Mamba and Video Mamba), MSAmba demonstrates clear performance advantages over them with similar computational overhead and parameter size. Conversely, replacing ISM with a Transformer leads to performance degradation despite having a larger parameter size and computational overhead. This further validates the exceptional sequential modeling capacity of ISM.

Method	Parameters↓	FLOPS↓	Acc-2↑	Acc-7↑
Vision Mamba	**1.40M**	**0.12G**	84.97/85.49	47.01
Video Mamba	1.41M	0.13G	85.74/86.89	47.18
Transformer	4.45M	0.38G	84.74/86.20	47.67
MSAmba	1.41M	0.13G	**85.99/87.43**	**49.67**

Table 5: Sequential modeling capacity study of ISM against previous Mamba blocks (Zhu et al. 2024; Li et al. 2024) and Transformer.

Hybrid Architecture of CHM To validate the effectiveness of the hybrid architecture of CHM, we conduct experiments on CMU-MOSI to study the performances of three configurations: pure Mamba architecture, CHM hybrid architecture and pure cross-attention Transformer-based architecture. Pure Mamba architecture replaces the self-attention module in CHM as an SSM, while the pure cross-attention Transformer-based method employs a cross-attention Transformer with a similar parameter size to replace CHM.

As shown in Tab. 6, compared with applying pure cross-attention Transformers for cross-modal interaction, CHM achieves a smaller computational overhead and better performances, demonstrating the effectiveness of Mamba-based architecture in cross-modal interaction. On the other hand, the hybrid CHM outperforms pure SSM, further demonstrating the superiority of the hybrid architecture.

Visualization of Representation We provide t-SNE visualizations on the integrated tokens for final sentiment prediction as qualitative results for the ablation study. As shown in Fig. 4, both the module design of BSSM and GLCE in ISM help achieve feature separation between negative and positive samples. Specifically, the cross-modal fusion features learned by MSAmba exhibit stronger discrimination compared with the simple integration of modality-specific tokens when without CHM, demonstrating the effectiveness of CHM in cross-modal interaction.

Method	Parameters↓	FLOPS↓	Acc-2↑	Acc-7↑
MSAmba(Self-Attn→SSM)	1.45M	0.13G	85.24/86.45	46.29
MSAmba(CHM→Transformer)	1.45M	0.26G	84.37/85.04	47.81
MSAmba	1.41M	0.13G	85.99/87.43	49.67

Table 6: Efficiency study on the hybrid architecture of CHM on CMU-MOSI.

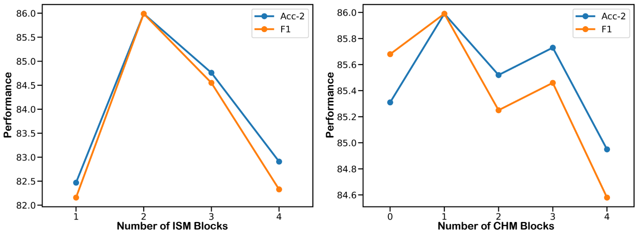

Figure 5: Hyperparameter study on the number of ISM blocks and CHM blocks on CMU-MOSI. In the study of ISM blocks, the number of CHM blocks is set as 1. In the study of CHM blocks, the number of ISM blocks is set as 2.

Hyperparameter Study

In this section, we conduct hyperparameter studies on the effect of the layer numbers on models' performances on CMU-MOSI. Experiment results are shown in Fig. 5. It can be seen that, in CMU-MOSI, the optimal configuration of the ISM blocks is set as 2. Initially, increasing the number of ISM blocks achieves stronger performances due to the improved sequential modeling capacity brought by duplicated ISM blocks. However, given the limited dataset size, we reckon that the excessive parameters brought by more duplicated blocks would overfit the datasets, and thus undergo performance degradation.

Regarding the hyperparameter study on the number of CHM blocks, it can be seen that the implementation of the CHM blocks would indeed enhance the performances. However, similar to the increase of the ISM blocks, the model would overfit the dataset and thus undergo a performance drop. This overfitting tendency is further amplified by the presence of the self-attention module within the CHM block.

Conclusion

In this paper, we propose MSAmba, a novel Mamba-based architecture for multimodal sentiment analysis. With two core components (ICM and CHM blocks), MSAmba can effectively model intra-modal sequential representation and achieve efficient cross-modal interaction. Experiments on three benchmark datasets demonstrate the superior performance of MSAmba over previous attention-based Transformer methods and the great research potential of Mamba architecture in MSA.

Acknowledgment

The work was supported by the Natural Science Foundation of China under grants no. 62276170, 82261138629, the Science and Technology Project of Guangdong Province under grants no. 2023A1515011549, 2023A1515010688, the Science and Technology Innovation Commission of Shenzhen under grant no. JCYJ20220531101412030, the Open Research Fund from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), under Grant No. GML-KF-24-11, and Guangdong Provincial Key Laboratory under grant no. 2023B1212060076.

References

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

Behrouz, A.; and Hashemi, F. 2024. Graph Mamba: Towards Learning on Graphs with State Space Models. *CoRR*, abs/2402.08678.

Cambria, E.; Rajagopal, D.; Olsher, D.; and Das, D. 2013. Big social data analysis. *Big data computing*, 13: 401–414.

Chen, G.; Huang, Y.; Xu, J.; Pei, B.; Chen, Z.; Li, Z.; Wang, J.; Li, K.; Lu, T.; and Wang, L. 2024. Video mamba suite: State space model as a versatile alternative for video understanding. *arXiv preprint arXiv:2403.09626*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Ezzameli, K.; and Mahersia, H. 2023. Emotion recognition from unimodal to multimodal analysis: A review. *Information Fusion*, 99: 101847.

Fan, C.; Zhu, K.; Tao, J.; Yi, G.; Xue, J.; and Lv, Z. 2024. Multi-level Contrastive Learning: Hierarchical Alleviation of Heterogeneity in Multimodal Sentiment Analysis. *IEEE Transactions on Affective Computing*.

Friston, K. J.; Harrison, L.; and Penny, W. 2003. Dynamic causal modelling. *Neuroimage*, 19(4): 1273–1302.

Grazzi, R.; Siems, J.; Schrodi, S.; Brox, T.; and Hutter, F. 2024. Is mamba capable of in-context learning? *arXiv preprint arXiv:2402.03170*.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Gu, A.; Goel, K.; and Ré, C. 2022. Efficiently Modeling Long Sequences with Structured State Spaces. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Gu, A.; Johnson, I.; Goel, K.; Saab, K.; Dao, T.; Rudra, A.; and Ré, C. 2021. Combining Recurrent, Convolutional, and Continuous-time Models with Linear State Space Layers. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y. N.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 572–585.

Han, W.; Chen, H.; and 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*, 9180–9192.

Hazarika, D.; Zimmermann, R.; and Poria, S. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, 1122–1131.

Hespanha, J. P. 2018. *Linear systems theory*. Princeton university press.

Hu, G.; Lin, T.-E.; Zhao, Y.; Lu, G.; Wu, Y.; and Li, Y. 2022. UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 7837–7851. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Huang, J.; Tao, J.; Liu, B.; Lian, Z.; and Niu, M. 2020. Multimodal transformer fusion for continuous emotion recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3507–3511. IEEE.

Jiang, Y.; Li, W.; Hossain, M. S.; Chen, M.; Alelaiwi, A.; and Al-Hammadi, M. 2020. A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition. *Information Fusion*, 53: 209–221.

Kaur, R.; and Kautish, S. 2022. Multimodal sentiment analysis: A survey and comparison. *Research anthology on implementing sentiment analysis across multiple disciplines*, 1846–1870.

Li, B.; Fei, H.; Liao, L.; Zhao, Y.; Teng, C.; Chua, T.-S.; Ji, D.; and Li, F. 2023. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, 5923–5934.

Li, K.; Li, X.; Wang, Y.; He, Y.; Wang, Y.; Wang, L.; and Qiao, Y. 2024. Videomamba: State space model for efficient video understanding. *arXiv preprint arXiv:2403.06977*.

Li, Y.; Wang, Y.; and Cui, Z. 2023. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6631–6640.

Liang, D.; Zhou, X.; Wang, X.; Zhu, X.; Xu, W.; Zou, Z.; Ye, X.; and Bai, X. 2024. Pointmamba: A simple state space model for point cloud analysis. *arXiv preprint arXiv:2402.10739*.

Lieber, O.; Lenz, B.; Bata, H.; Cohen, G.; Osin, J.; Dalmedigos, I.; Safahi, E.; Meirom, S.; Belinkov, Y.; Shalev-Shwartz, S.; et al. 2024. Jamba: A hybrid transformer-mamba language model. *arXiv preprint arXiv:2403.19887*.

Liu, Z.; Shen, Y.; Lakshminarasimhan, V. B.; Liang, P. P.; Zadeh, A. B.; and Morency, L.-P. 2018. 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)*, 2247–2256.

Ma, J.; Li, F.; and Wang, B. 2024. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*.

Mai, S.; Zeng, Y.; Zheng, S.; and Hu, H. 2022. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 14(3): 2276–2289.

McFee, B.; Raffel, C.; Liang, D.; Ellis, D. P.; McVicar, M.; Battenberg, E.; and Nieto, O. 2015. librosa: Audio and music signal analysis in python. In *SciPy*, 18–24.

Rahman, W.; Hasan, M. K.; Lee, S.; Zadeh, A.; Mao, C.; Morency, L.-P.; and Hoque, E. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, 2359. NIH Public Access.

Ruan, J.; and Xiang, S. 2024. Vm-unet: Vision mamba unet for medical image segmentation. *arXiv preprint arXiv:2402.02491*.

Smith, J. T. H.; Warrington, A.; and Linderman, S. W. 2023. Simplified State Space Layers for Sequence Modeling. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Tsai, Y.-H. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.-P.; and Salakhutdinov, R. 2019a. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, 6558. NIH Public Access.

Tsai, Y.-H. H.; Liang, P. P.; Zadeh, A.; Morency, L.-P.; and Salakhutdinov, R. 2019b. Learning Factorized Multimodal Representations. In *International Conference on Representation Learning*.

Wang, C.; Tsepa, O.; Ma, J.; and Wang, B. 2024. Graph-Mamba: Towards Long-Range Graph Sequence Modeling with Selective State Spaces. *CoRR*, abs/2402.00789.

Wu, S.; Dai, D.; Qin, Z.; Liu, T.; Lin, B.; Cao, Y.; and Sui, Z. 2023. Denoising Bottleneck with Mutual Information Maximization for Video Multimodal Fusion. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2231–2243. Toronto, Canada: Association for Computational Linguistics.

Wu, Z.; Zhang, Q.; Miao, D.; Yi, K.; Fan, W.; and Hu, L. 2024. HyDiscGAN: A Hybrid Distributed cGAN for Audio-Visual Privacy Preservation in Multimodal Sentiment Analysis. In Larson, K., ed., *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, 6550–6558. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Xing, Z.; Ye, T.; Yang, Y.; Liu, G.; and Zhu, L. 2024. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. *arXiv preprint arXiv:2401.13560*.

Yang, D.; Huang, S.; Kuang, H.; Du, Y.; and Zhang, L. 2022. Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1642–1651.

Yang, J.; Yu, Y.; Niu, D.; Guo, W.; and Xu, Y. 2023. Confede: Contrastive feature decomposition for multimodal sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7617–7630.

Yang, Y.; Dong, X.; and Qiang, Y. 2024. CLGSI: A Multimodal Sentiment Analysis Framework based on Contrastive Learning Guided by Sentiment Intensity. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 2099–2110.

Yu, W.; Xu, H.; Meng, F.; Zhu, Y.; Ma, Y.; Wu, J.; Zou, J.; and Yang, K. 2020. CH-SIMS: A Chinese Multimodal Sentiment Analysis Dataset with Fine-grained Annotation of Modality. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3718–3727. Online: Association for Computational Linguistics.

Yu, W.; Xu, H.; Yuan, Z.; and Wu, J. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 10790–10797.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L.-P. 2016. Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages. *IEEE Intelligent Systems*, 31(6): 82–88.

Zadeh, A. B.; Liang, P. P.; Poria, S.; Cambria, E.; and 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)*, 2236–2246.

Zhang, H.; Wang, Y.; Yin, G.; Liu, K.; Liu, Y.; and Yu, T. 2023. Learning Language-guided Adaptive Hyper-modality Representation for Multimodal Sentiment Analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 756–767. Association for Computational Linguistics.

Zhang, T.; Li, X.; Yuan, H.; Ji, S.; and Yan, S. 2024. Point could mamba: Point cloud learning via state space model. *arXiv preprint arXiv:2403.00762*.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. In *Forty-first International Conference on Machine Learning*.