

TF-Mamba: Text-enhanced Fusion Mamba with Missing Modalities for Robust Multimodal Sentiment Analysis

Xiang Li^{1*}, Xianfu Cheng^{2*}, Dezhuang Miao¹, Xiaoming Zhang^{1†}, Zhoujun Li²

¹School of Cyber Science and Technology, Beihang University, Beijing

²School of Computer Science and Engineering, Beihang University, Beijing
{xlggg, buaacf, taiyue, yolixs, lizj}@buaa.edu.cn

Abstract

Multimodal Sentiment Analysis (MSA) with missing modalities has attracted increasing attention recently. While current Transformer-based methods leverage dense text information to maintain model robustness, their quadratic complexity hinders efficient long-range modeling and multimodal fusion. To this end, we propose a novel and efficient **Text-enhanced Fusion Mamba** (TF-Mamba) framework for robust MSA with missing modalities. Specifically, a Text-aware Modality Enhancement (TME) module aligns and enriches non-text modalities, while reconstructing the missing text semantics. Moreover, we develop Text-based Context Mamba (TC-Mamba) to capture intra-modal contextual dependencies under text collaboration. Finally, Text-guided Query Mamba (TQ-Mamba) queries text-guided multimodal information and learns joint representations for sentiment prediction. Extensive experiments on three MSA datasets demonstrate the effectiveness and efficiency of the proposed method under missing modality scenarios. Code is available at <https://github.com/codemous/TF-Mamba>.

1 Introduction

Multimodal Sentiment Analysis (MSA) aims to understand and integrate sentiment cues expressed in multiple modalities (e.g., text, visual, and audio). Previous studies (Yang et al., 2022; Sun et al., 2022; Li et al., 2023; Feng et al., 2024) demonstrate that integrating complementary multimodal information gains MSA performance and enables its vital role in several applications. However, most existing methods (Han et al., 2021; Mai et al., 2022; Zhang et al., 2023; Wang et al., 2024; Zeng et al., 2024) are developed under ideal laboratory conditions where all modalities are assumed to be fully available during both training and inference. In

real-world scenarios, many inevitable factors like background noise and sensor failures often result in incomplete or corrupted modalities. These challenges severely undermine the performance and robustness of MSA models in practice.

Recent studies (Yuan et al., 2021, 2023; Li et al., 2024b; Zhang et al., 2024) attempt to tackle the challenge of missing modalities in MSA. Among them, Transformer-based fusion models achieve notable progress owing to their powerful sequence modeling capabilities. For example, TFR-Net (Yuan et al., 2021) adopts a Transformer-based feature reconstruction strategy to recover missing information in multimodal sequences. More recently, LNLN (Zhang et al., 2024) prioritizes high-quality text features and treats text as the dominant modality to improve model robustness under various noise conditions. Unfortunately, these approaches often suffer from quadratic computational complexity and fail to achieve efficient text-enhanced multimodal fusion. Mamba (Gu and Dao, 2023) emerges as a linear-time alternative to Transformers, exhibiting great promise in various domains (Yang et al., 2024; Li et al., 2024d). Nevertheless, incorporating text enhancement into Mamba-based architectures for efficient missing modality modeling and fusion remains unexplored, which could potentially contribute to enhancing MSA performance.

To address these issues, we propose a novel and efficient **Text-enhanced Fusion Mamba** (TF-Mamba) framework for robust MSA with missing modalities. TF-Mamba features three core text-dominant components: Text-aware Modality Enhancement (TME), Text-based Context Mamba (TC-Mamba), and Text-guided Query Mamba (TQ-Mamba). Specifically, the TME module first aligns and enhances audio and visual modalities with text features, and reconstructs missing textual semantics from incomplete inputs. Subsequently, TC-Mamba leverages text information as the collaborative bridge within Mamba to model contextual

*Equal contribution.

†Corresponding author.

dependencies in audio and video streams. The collaborative signals refine text representations in return for capturing shared semantics. Finally, TQ-Mamba queries informative multimodal features via text-guided cross-attention. It then learns cross-modal interactions with Mamba blocks to generate joint representations for sentiment prediction. The contributions are summarized as follows:

- To our knowledge, TF-Mamba is the first attempt incorporating text enhancement into Mamba-based fusion architecture for robust MSA under missing modality conditions.
- We design three text-dominant modules to efficiently learn intra-modal dependencies and cross-modal interactions from incomplete modalities, thereby enhancing model robustness with dense text sentiment information.
- Extensive experiments on three MSA benchmarks demonstrate the superiority of TF-Mamba under uncertain missing modality scenarios. For instance, on MOSI, TF-Mamba outperforms Transformer-based state-of-the-art with fewer parameters and FLOPs.

2 Related Work

2.1 Multimodal Sentiment Analysis

Multimodal Sentiment Analysis (MSA) integrates multimodal information to understand and analyze human sentiments. Mainstream studies in MSA (Li et al., 2024e, 2025a,b; Sun and Tian, 2025) focus on developing sophisticated fusion strategies and interaction mechanisms to improve sentiment prediction performance. For example, Han et al. (2021) hierarchically maximize mutual information between unimodal pairs to enhance multimodal fusion. Zhang et al. (2023) leverage sentiment-intensive cues to guide the representation learning of other modalities. Despite these progresses, most methods assume fully available modalities, which is rarely achievable in real-world scenarios due to random data missing issues.

Recent studies (Yuan et al., 2021, 2023; Li et al., 2024b,c; Zhang et al., 2024; Guo et al., 2024) handle missing modalities by learning joint multimodal representations from available data or reconstructing incomplete information. For instance, Li et al. (2024b) decompose modalities into sentiment-relevant and modality-specific components to reconstruct sentiment semantics through modality

translation. However, they overlook the crucial role of text in sentiment expression without ensuring the quality of dominant modality representations. Recognizing this, Zhang et al. (2024) propose a Transformer-based language-dominated noise-resistant network that prioritizes textual features to improve robustness under noisy conditions. However, Transformer-based methods suffer from considerable computational overhead due to their inherent quadratic complexity, making them inefficient for modeling long or high-dimensional sequences. Given the rich sentiment cues in text and Mamba’s efficiency in modeling long-range dependencies, our work explores text-enhanced strategies within the efficient Mamba architectures for robust MSA with random missing modalities.

2.2 State Space Models and Mamba

State Space Models (SSMs) (Gu et al., 2022a,b) recently gain considerable attention for their ability to capture long-range dependencies with linear computational complexity. Mamba (Gu and Dao, 2023) as an extension of SSMs, incorporates selection mechanisms and hardware-aware parallel algorithms to enhance sequence modeling efficiency. Its parallel scanning strategy further accelerates inference and achieves impressive performance on long-sequence tasks across various domains (Zhu et al., 2024; Yang et al., 2024; Jiang et al., 2025).

Building on Mamba’s success, several studies (Li et al., 2024d,a; Dong et al., 2025; Li et al., 2025c) explore its potential for multimodal fusion. For instance, Qiao et al. (2024) concatenate visual and textual sequences before jointly modeling them through Mamba blocks. Li et al. (2025c) learns both local and global cross-modal alignments prior to Mamba-based multimodal fusion. Recently, Ye et al. (2025) pioneer the use of Mamba for depression detection, combining hierarchical modeling with progressive fusion strategies. Overall, these Mamba-based frameworks deliver competitive performance and improved efficiency compared to Transformer-based methods. However, its application to MSA with missing modalities remains underexplored, especially in integrating text-enhanced fusion strategies into Mamba’s efficient modeling architecture. In contrast, we treat text as dominant modality during Mamba-based sequence modeling and multimodal fusion to enhance model robustness in incomplete modality settings.

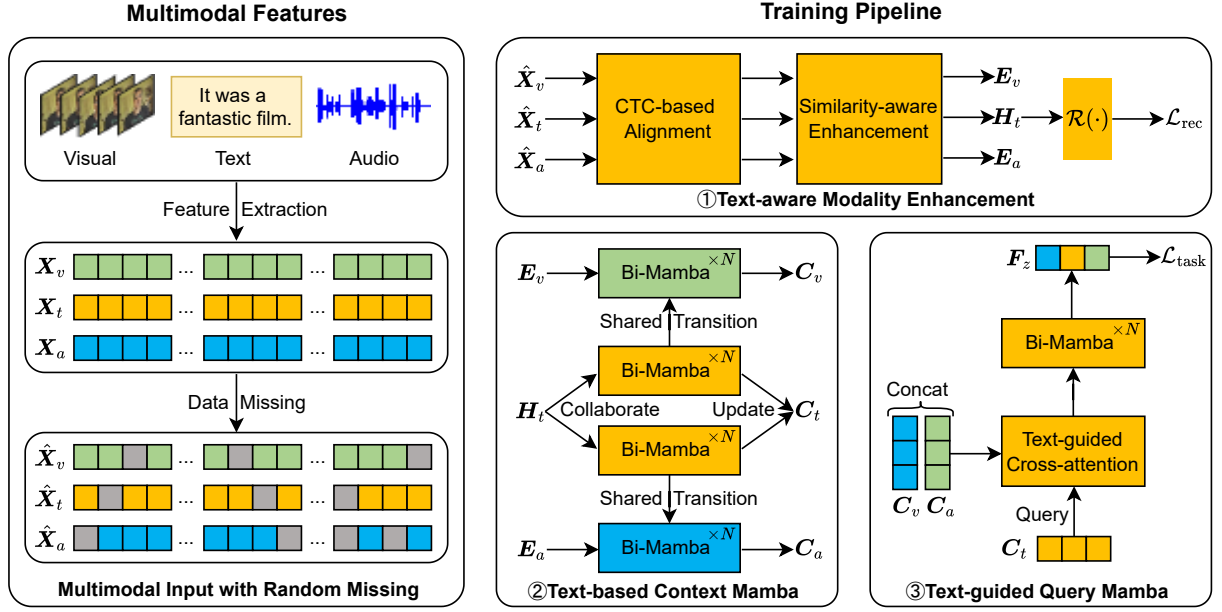


Figure 1: Overview of the TF-Mamba framework, which consists of three main components: Text-aware Modality Enhancement (TME), Text-based Context Mamba (TC-Mamba), and Text-guided Query Mamba (TQ-Mamba). Yellow blocks indicate the dominant role of the text modality in the training pipeline.

3 Method

3.1 Preliminary of Mamba

State Space Models (SSMs) (Gu et al., 2022a,b) gain increasing attention in recent years. An SSM maps a 1D sequence $\mathbf{x}(t) \in \mathbb{R}^L$ to $\mathbf{y}(t) \in \mathbb{R}^L$ via a hidden state $\mathbf{h}(t) \in \mathbb{R}^N$, where N and L denote the number of hidden states and sequence length. The above process is defined as:

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

$$\mathbf{y}(t) = \mathbf{C}\mathbf{h}(t) \quad (2)$$

where $\mathbf{A} \in \mathbb{R}^{N \times N}$ denotes the state transition matrix, and $\mathbf{B} \in \mathbb{R}^{N \times 1}$ and $\mathbf{C} \in \mathbb{R}^{1 \times N}$ are the input and output projections.

To make the continuous-time SSM system suitable for digital computing and real-world data, Mamba employs the zero-order hold method to discretize the continuous parameters $\mathbf{A}$ and $\mathbf{B}$ into $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$ using a time scale parameter Δ :

$$\bar{\mathbf{A}} = \exp(\Delta\mathbf{A}) \quad (3)$$

$$\bar{\mathbf{B}} = (\Delta\mathbf{A})^{-1} (\exp(\Delta\mathbf{A}) - \mathbf{I}) \cdot \Delta\mathbf{B} \quad (4)$$

The resulting discretized system with step size Δ can then be formulated as:

$$\mathbf{h}_t = \bar{\mathbf{A}}\mathbf{h}_{t-1} + \bar{\mathbf{B}}\mathbf{x}_t \quad (5)$$

$$\mathbf{y}_t = \mathbf{C}\mathbf{h}_t \quad (6)$$

Mamba further unfolds the hidden states recursively and computes the output as:

$$\bar{\mathbf{K}} = (\mathbf{C}\bar{\mathbf{B}}, \mathbf{C}\bar{\mathbf{A}}\bar{\mathbf{B}}, \dots, \mathbf{C}\bar{\mathbf{A}}^{L-1}\bar{\mathbf{B}}) \quad (7)$$

$$\mathbf{y} = \mathbf{x} \circledast \bar{\mathbf{K}} \quad (8)$$

where $\circledast$ denotes the convolution operation, and $\bar{\mathbf{K}} \in \mathbb{R}^L$ is the global convolution kernel. Mamba introduces the content-aware selection mechanism to modulate state transitions, enhancing its capacity to model complex temporal dependencies.

3.2 Framework Overview

The objective of our robust MSA task is to predict sentiment from video clips under scenarios where one or more modality features may be partially missing. As illustrated in Figure 1, the framework first generates multimodal features with randomly missing data, establishing the foundation for TF-Mamba training. Following input preparation, the TME module performs text-centric CTC-based uni-modal alignment, conducts similarity-aware token enhancement, and reconstructs missing textual semantics. The standardized multimodal features are then fed into the TC-Mamba and TQ-Mamba modules for efficient intra-modal modeling and cross-modal fusion, generating robust joint multimodal representations for sentiment prediction. The following sections provide a detailed description of each component in TF-Mamba.

3.3 Multimodal Input with Random Missing

Following prior work (Yuan et al., 2021; Zhang et al., 2024), we obtain the initial embeddings $\mathbf{X}_m \in \mathbb{R}^{T_m \times D_m}$ using standard toolkits, where $m \in \{t, v, a\}$ represents the text, visual, and audio modalities, respectively. Here, T_m denotes the sequence length and D_m is the feature dimension.

To simulate incomplete multimodal scenarios, we apply random replacement with missing values to the input sequences, resulting in corrupted inputs $\hat{\mathbf{X}}_m$. Consistent with LNLN (Zhang et al., 2024), we randomly erase between 0% and 100% of each modality sequence. Specifically, missing values in the visual and audio modalities are replaced with zeros, while missing tokens in the text modality are substituted with the [UNK] token used by BERT (Devlin et al., 2019).

3.4 Text-aware Modality Enhancement

Motivated by the informative nature of text in sentiment expression, we introduce the Text-aware Modality Enhancement (TME) module to standardize and enrich visual and audio features based on their semantic similarity to text and reconstruct corrupted or missing text semantics.

We first unify sequence length and feature dimension using a CTC-based temporal alignment (Zhou et al., 2024) with a linear mapping:

$$\{\mathbf{H}_v, \mathbf{H}_t, \mathbf{H}_a\} = \text{CTC}(\{\hat{\mathbf{X}}_v, \hat{\mathbf{X}}_t, \hat{\mathbf{X}}_a\}) \quad (9)$$

where $\mathbf{H}_v, \mathbf{H}_t, \mathbf{H}_a \in \mathbb{R}^{L \times D}$ are the aligned features with length L (set to T_t) and dimension D . To enhance non-text modalities (i.e., visual), we compute token-wise similarities between visual and text tokens. Given L2-normalized token embeddings $\mathbf{v}_i$ and $\mathbf{t}_j$, their similarity score is calculated via temperature-scaled dot-product:

$$S_{i,j}^{vt} = \frac{\exp(\langle \mathbf{v}_i, \mathbf{t}_j \rangle / \tau)}{\sum_{k=1}^L \exp(\langle \mathbf{v}_i, \mathbf{t}_k \rangle / \tau)} \quad (10)$$

To suppress noisy or weakly correlated pairs, a hard threshold $\theta = 1/L$ is applied, generating a binary mask $\mathbf{M}_{i,j}^{vt} = \mathbb{I}[S_{i,j}^{vt} > \theta]$. The final enhanced visual representations are obtained as:

$$\mathbf{E}_v = \mathbf{H}_v + (\mathbf{M}^{vt} \odot \mathbf{S}^{vt}) \cdot \mathbf{H}_t \quad (11)$$

The same way is applied to enhance the audio modality, resulting in enriched audio representations $\mathbf{E}_a$. Given the rich sentiment semantics in text, we reconstruct missing textual information to

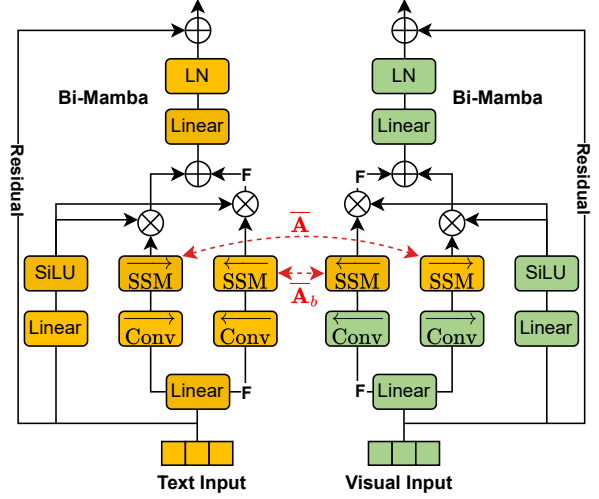


Figure 2: An illustration of TC-Mamba with text and visual inputs. Red dashed lines indicate shared state transition matrices across Bi-Mamba blocks. The symbol F denotes the temporal flip operation.

enhance model robustness. A simple MLP-based decoder $\mathcal{R}(\cdot)$, consisting of two linear layers and a ReLU activation, is adopted as the reconstructor. We employ Smooth L1 loss (Yuan et al., 2021; Sun et al., 2023) to evaluate the text reconstruction:

$$\mathcal{L}_{\text{rec}} = \text{Smooth}_{L_1}((\mathbf{X}_t - \mathcal{R}(\mathbf{H}_t)) \cdot (1 - \mathbf{P}_t)) \quad (12)$$

where $\mathbf{P}_t$ is the text temporal mask indicating missing positions. $\mathcal{L}_{\text{rec}}$ encourages the model to restore incomplete text semantics. The reconstructed semantics in turn enhances non-text modalities.

3.5 Text-based Context Mamba

To capture text collaborative intra-modal contextual dependencies, we introduce the Text-based Context Mamba (TC-Mamba) to conduct efficient long-range modeling. TC-Mamba adopts bi-directional Mamba (Bi-Mamba) blocks (Zhu et al., 2024; Yang et al., 2024) as the backbone, which learns bidirectional unimodal temporal patterns under the contextual supervision of text representations.

In Bi-Mamba, each modality stream employs forward and backward SSMs, parameterized by six matrices: $\bar{\mathbf{A}}, \bar{\mathbf{B}}, \mathbf{C}$ and $\bar{\mathbf{A}}_b, \bar{\mathbf{B}}_b, \mathbf{C}_b$. The transition matrices $\bar{\mathbf{A}}$ and $\bar{\mathbf{A}}_b$ govern temporal dynamics, while $\bar{\mathbf{B}}, \bar{\mathbf{B}}_b$ and $\mathbf{C}, \mathbf{C}_b$ manage the input and output operations. To capture temporal dependencies and common semantics, we share the bidirectional transition matrices $\bar{\mathbf{A}}$ and $\bar{\mathbf{A}}_b$ between the text modality and each non-text modality. In contrast, $\bar{\mathbf{B}}, \bar{\mathbf{B}}_b$ and $\mathbf{C}, \mathbf{C}_b$ remain independent to capture modality-specific information.

Figure 2 illustrates the modeling process using text and visual inputs as an example. Specifically, the forward dynamics can be expressed as:

$$\mathbf{h}_t^{(t)} = \overline{\mathbf{A}} \mathbf{h}_{t-1}^{(t)} + \overline{\mathbf{B}}^t \mathbf{x}_t^{(t)}, \mathbf{y}_t^{(t)} = \mathbf{C}^t \mathbf{h}_t^{(t)} \quad (13)$$

$$\mathbf{h}_t^{(v)} = \overline{\mathbf{A}} \mathbf{h}_{t-1}^{(v)} + \overline{\mathbf{B}}^v \mathbf{x}_t^{(v)}, \mathbf{y}_t^{(v)} = \mathbf{C}^v \mathbf{h}_t^{(v)} \quad (14)$$

where $\mathbf{x}_t^{(t)}$, $\mathbf{x}_t^{(v)}$, $\mathbf{h}_t^{(t)}$, $\mathbf{h}_t^{(v)}$, and $\mathbf{y}_t^{(t)}$, $\mathbf{y}_t^{(v)}$ represent the input, hidden state, and output features at t time step. The backward SSMs follows an analogous structure using the $\overline{\mathbf{A}}_b$, $\overline{\mathbf{B}}_b$, and $\mathbf{C}_b$ matrices. The overall process of Bi-Mamba is formally as:

$$\begin{cases} \mathbf{C}_t^1 = \text{Bi-Mamba}(\mathbf{H}_t) \\ \mathbf{C}_v = \text{Bi-Mamba}(\mathbf{E}_v) \end{cases} \quad (15)$$

Similarly, the text-to-audio operation is denoted as:

$$\begin{cases} \mathbf{C}_t^2 = \text{Bi-Mamba}(\mathbf{H}_t) \\ \mathbf{C}_a = \text{Bi-Mamba}(\mathbf{E}_a) \end{cases} \quad (16)$$

In each TC-Mamba block, the text features are updated twice and we average them to obtain refined text representations $\mathbf{C}_t$:

$$\mathbf{C}_t = \text{Mean}(\mathbf{C}_t^1, \mathbf{C}_t^2) \quad (17)$$

which capture the multimodal co-semantics and are used in reverse to iteratively update $\mathbf{C}_v$ and $\mathbf{C}_a$.

3.6 Text-guided Query Mamba

Building on the text-based context modeling in TC-Mamba, we further enhance cross-modal interactions with the proposed Text-guided Query Mamba (TQ-Mamba). TQ-Mamba explicitly performs multimodal fusion by querying informative multimodal features and capturing cross-modal interactions between text and other modalities.

Specifically, it first leverages text-guided cross-attention to identify and query the most informative segments from the visual and audio streams. The query operation is formulated as:

$$\mathbf{Q}_f = \text{Cross-Attn}(\mathbf{C}_t, [\mathbf{C}_v; \mathbf{C}_a], [\mathbf{C}_v; \mathbf{C}_a]) \quad (18)$$

where $\mathbf{Q}_f$ represents the text-guided multimodal features, and $[\mathbf{C}_v; \mathbf{C}_a]$ denotes the concatenated sequence of visual and audio features. $\mathbf{Q}_f$ is then passed through latent Bi-Mamba blocks to learn intricate multimodal interactions:

$$\mathbf{F}_z = \text{Bi-Mamba}(\mathbf{Q}_f) \quad (19)$$

Finally, the fused feature $\mathbf{F}_z$ is aggregated using max pooling and projected via a fully connected layer (FC) to infer sentiment intensity.

$$\hat{\mathbf{Y}} = \text{FC}(\text{MaxPool}(\mathbf{F}_z)), \quad (20)$$

where $\hat{\mathbf{Y}}$ represents the predicted sentiment score.

3.7 Overall Training Objective

Our model is trained end-to-end with a combined loss that integrates sentiment prediction and text reconstruction objectives. The sentiment prediction loss $\mathcal{L}_{\text{task}}$ can be described as:

$$\mathcal{L}_{\text{task}} = \frac{1}{N_b} \sum_{n=1}^{N_b} \|\mathbf{Y}^n - \hat{\mathbf{Y}}^n\|_2^2 \quad (21)$$

Therefore, the overall loss $\mathcal{L}$ can be written as:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda \mathcal{L}_{\text{rec}}, \quad (22)$$

where $\mathcal{L}_{\text{rec}}$ is the reconstruct loss mentioned above and hyperparameter λ balances the two terms.

Dataset	#Train	#Valid	#Test	#Total	Language
MOSI	1284	229	686	2199	English
MOSEI	16326	1871	4659	22856	English
SIMS	1368	456	457	2281	Chinese

Table 1: The statistics of MOSI, MOSEI, and SIMS.

Descriptions	MOSI	MOSEI	SIMS
Length L	50	50	39
Mamba State	12	12	16
Mamba Expansion	4	4	2
Mamba Depth	{1,1}	{2,2}	{1,2}
Attention Head	8	8	8
Loss Weight λ	0.7	0.3	1.0
Warm Up	✓	✓	✓

Table 2: Hyper-parameters settings on different datasets.

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate our method on three widely used MSA datasets: MOSI (Zadeh et al., 2016), MOSEI (Zadeh et al., 2018), and SIMS (Yu et al., 2020). All experiments are conducted under the unaligned data setting. We adopt the publicly released features provided by each benchmark. Table 1 shows the statistic details. The details of feature extraction procedures are reported in Appendix A.

Method	MOSI						MOSEI					
	Acc-7	Acc-5	Acc-2	F1	MAE	Corr	Acc-7	Acc-5	Acc-2	F1	MAE	Corr
MISA	29.85	33.08	71.49 / 70.33	71.28 / 70.00	1.085	0.524	40.84	39.39	71.27 / 75.82	63.85 / 68.73	0.780	0.503
Self-MM	29.55	34.67	70.51 / 69.26	66.60 / 67.54	1.070	0.512	44.70	45.38	73.89 / 77.42	68.92 / 72.31	0.695	0.498
MMIM	31.30	33.77	69.14 / 67.06	66.65 / 64.04	1.077	0.507	40.75	41.74	73.32 / 75.89	68.72 / 70.32	0.739	0.489
TFR-Net	29.54	34.67	68.15 / 66.35	61.73 / 60.06	1.200	0.459	46.83	34.67	73.62 / 77.23	68.80 / 71.99	0.697	0.489
CENET	30.38	33.62	71.46 / 67.73	68.41 / 64.85	1.080	0.504	47.18	47.83	74.67 / 77.34	70.68 / 74.08	0.685	0.535
ALMT	30.30	33.42	70.40 / 68.39	72.57 / 71.80	1.083	0.498	40.92	41.64	76.64 / 77.54	77.14 / 78.03	0.674	0.481
BI-Mamba	31.20	34.02	71.74 / 71.12	71.83 / 71.11	1.087	0.498	45.12	45.76	76.82 / 76.72	76.35 / 76.38	0.701	0.545
LNLN	32.53	36.25	71.91 / 70.11	71.71 / 70.02	1.062	0.503	45.42	46.17	76.30 / 78.19	77.77 / 79.95	0.692	0.530
TF-Mamba	33.95	37.74	73.46 / 72.54	73.59 / 72.57	1.035	0.548	45.66	<u>46.64</u>	77.34 / <u>77.61</u>	<u>77.18</u> / 77.43	0.673	0.578

Table 3: Overall performance comparison on the MOSI and MOSEI datasets under missing modality settings.

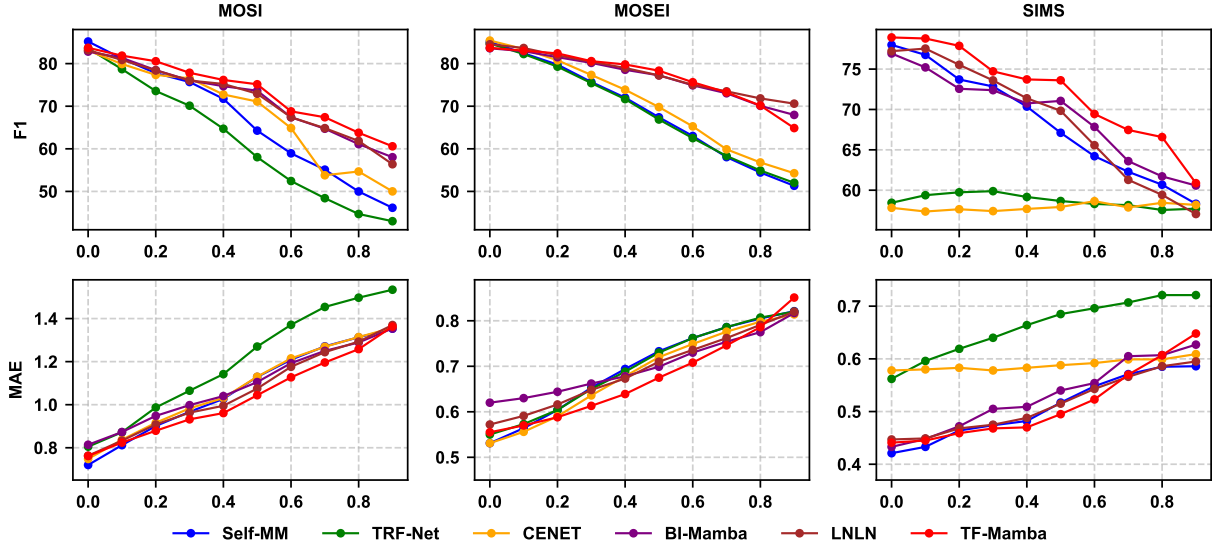


Figure 3: Performance trends of models under varying missing rates on MOSI, MOSEI, and SIMS datasets.

For evaluation, following [Zhang et al. \(2024\)](#), we report 5-class (Acc-5) and 7-class (Acc-7) Accuracy on MOSI and MOSEI, and 3-class (Acc-3) and 5-class (Acc-5) Accuracy on SIMS. Additionally, 2-class Accuracy (Acc-2), mean absolute error (MAE), Pearson correlation (Corr), and F1 score (F1) are reported on all datasets. For Acc-2 and F1 on MOSI and MOSEI, results are provided under two settings: negative vs. positive (left of "/") and negative vs. non-negative (right of "/"). Except for MAE, higher values indicate better performance.

4.2 Implementation Details

To ensure fair comparison and evaluation, we follow the same missing modality setting from LNLN ([Zhang et al., 2024](#)), where training data is randomly dropped with an uncertain probability. During testing, the missing rate r is varied from 0 to 0.9 in increments of 0.1. Evaluation at $r = 1.0$ is excluded as it removes all modalities, making the task ill-posed. Final results are averaged across all missing rates to assess model robustness.

All experiments are conducted using PyTorch 2.1.0. Models are optimized with AdamW and a cosine annealing learning rate schedule, starting at 0.0001. The feature dimension for each modality is unified to 128. Training is performed for 200 epochs with a batch size of 64 on a single NVIDIA Tesla V100-DGXS GPU (32 GB memory). More implementation details are provided in Table 2.

4.3 Baseline Models

To evaluate the robustness of TF-Mamba, we conduct fair comparisons with a range of state-of-the-art (SOTA) methods under the same missing modality conditions. The compared baselines include MISA ([Hazarika et al., 2020](#)), Self-MM ([Yu et al., 2021](#)), MMIM ([Han et al., 2021](#)), CENET ([Wang et al., 2022](#)), TETFN ([Wang et al., 2023](#)), TFR-Net ([Yuan et al., 2021](#)), ALMT ([Zhang et al., 2023](#)), BI-Mamba ([Yang et al., 2024](#)), and LNLN ([Zhang et al., 2024](#)). Detailed baseline settings and descriptions are provided in Appendix B.

Method	Acc-5	Acc-3	Acc-2	F1	MAE	Corr
MISA	31.53	56.87	72.71	66.30	0.539	0.348
Self-MM	32.28	56.75	72.81	68.43	0.508	0.376
MMIM	31.81	52.76	69.86	66.21	0.544	0.339
TFR-Net	26.52	52.89	68.13	58.70	0.661	0.169
CENET	22.29	53.17	68.13	57.90	0.589	0.107
ALMT	20.00	45.36	69.66	72.76	0.561	0.364
BI-Mamba	31.90	54.95	70.79	69.26	0.529	0.345
LNLN	33.08	56.01	73.62	68.84	0.514	0.389
TF-Mamba	34.46	55.51	74.68	<u>72.20</u>	<u>0.512</u>	<u>0.386</u>

Table 4: Overall performance comparison on the SIMS dataset under missing modality settings.

Model	MOSI			SIMS		
	MAE	F1	Acc-7	MAE	F1	Acc-5
w/o Enhancement	1.104	72.78	29.17	0.522	68.83	29.67
w/o Reconstruction	1.085	73.09	31.73	0.520	70.21	32.89
w/o TME	1.071	73.06	32.81	0.517	69.07	33.24
TME + BI-Mamba	1.059	74.17	32.52	0.512	70.20	32.58
w/o TC-Mamba	1.064	73.23	32.39	0.517	68.33	34.35
w/o TQ-Mamba	1.055	72.68	32.55	0.517	69.44	33.79
TF-Mamba	1.035	73.46	33.95	0.512	72.20	34.46

Table 5: Ablation study of different modules and strategies on the MOSI and SIMS datasets.

4.4 Robustness Comparison

Tables 3 and 4 present the robustness results on the MOSI, MOSEI, and SIMS datasets. The best results are shown in bold, while the second-best results (ours) are underlined. TF-Mamba achieves superior performance across most metrics. Compared with Transformer-based models (CENET, TETFN, TFR-Net, and ALMT), Mamba-based approaches gain performance, underscoring their potential as competitive alternatives. Notably, TF-Mamba outperforms BI-Mamba, owing to its text enhancement strategy. On MOSI, TF-Mamba outperforms the previous SOTA model LNLN by 4.36% in Acc-7 and 2.62% in F1 score, showing its strong resilience to varying missing rates. On MOSEI, LNLN achieves better binary classification results, likely benefiting from the larger data scale. TF-Mamba also achieves strong performance on SIMS, obtaining the best five-class score of 34.46% and binary classification Accuracy of 74.68%.

Additionally, as shown in Figure 3, all models degrade as the missing rate increases, reflecting their sensitivity to incomplete data (see Appendix C for more details). TF-Mamba still maintains relatively stable performance under varying missing rates. These results validate the robustness of TF-Mamba and its promise for robust MSA.

4.5 Ablation Study

We conduct ablation experiments to assess the effectiveness of the strategies and modules within TF-Mamba. The average results on the MOSI and SIMS datasets are summarized in Table 5.

Effect of Text Modality When the text enhancement strategy is removed, model performance degrades notably, which underscores the effectiveness of text-enhanced fusion for sentiment prediction. Similarly, without the text reconstruction loss, the model struggles to recover missing textual semantics, leading to poorer multimodal representations. This highlights the importance of reconstructing incomplete text features for robust MSA. We further evaluate model performance under complete modality missing conditions ($r = 1.0$), with results and discussions summarized in Appendix D. The complete absence of the text modality causes a marked performance drop, even when combined with missing audio or visual inputs. In contrast, missing only audio or visual modality has a limited impact. These findings further underscore the significance of text enhancement in robust MSA.

Module Design of TME Removing the TME module leads to consistent performance drops on both datasets. These results show that TME effectively facilitates representation learning and multimodal fusion through text-aware alignment and enhancement. In addition, incorporating the TME module into the baseline BI-Mamba improves Acc-7 by 4.23% on MOSI and Acc-5 by 2.13% on SIMS. Such outcomes substantiate both the effectiveness and rationality of the TME module.

Role of TC-Mamba and TQ-Mamba Eliminating either TC-Mamba or TQ-Mamba leads to performance drops. When TC-Mamba is removed, the model struggles to learn text-related context within non-text modalities, which reduces overall effectiveness and underscores the role of TC-Mamba in collaborative intra-modal modeling. The impact is even greater when TQ-Mamba is excluded, indicating that guiding multimodal fusion with text queries is vital for producing more robust and expressive joint representations. Together, these modules strengthen both intra-modal representation learning and cross-modal fusion, resulting in improved robustness and predictive accuracy.

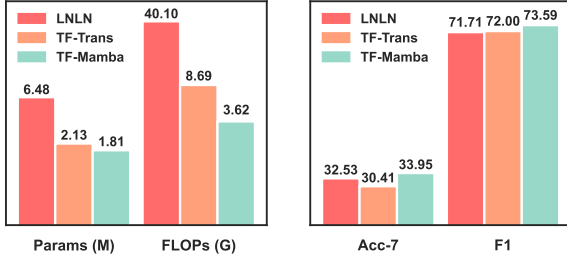


Figure 4: Model performance and complexity comparison during inference on MOSI dataset.

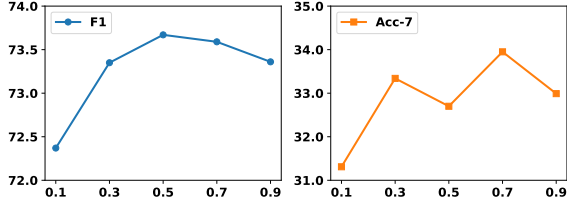


Figure 5: Effect of the regularization parameter λ on F1 Score and Acc-7 on the MOSI dataset.

4.6 Efficiency Study

We conduct an efficiency analysis of TF-Mamba, focusing on its intra-modal modeling and cross-modal interaction components. To ensure fair comparisons, we exclude the computational cost of pre-trained encoders. All experiments are performed under identical conditions, evaluating model parameters (Params), floating-point operations (FLOPs), and performance (Acc-7 and F1) on MOSI dataset. Comparisons are made against the Transformer-based SOTA baseline LNLN and a TF-Mamba variant, TF-Trans, where Mamba blocks are replaced with Transformer blocks.

As shown in Figure 4, TF-Mamba achieves superior performance while significantly improving efficiency over previous Transformer-based fusion models. Benefiting from the linear complexity of Mamba, TF-Mamba requires fewer parameters and reduces inference costs by 36.48G FLOPs compared to LNLN. Although replacing Mamba with Transformers in TF-Trans increases computational overhead, it retains competitive robustness and Accuracy. These results demonstrate that TF-Mamba offers substantial computational advantages while maintaining strong performance, validating the effectiveness of its efficient modeling and fusion strategy for robust MSA with missing modalities.

4.7 Sensitivity Analysis

In this section, we analyze the sensitivity of key hyperparameters on model performance. We focus

TC-Mamba	TQ-Mamba	MAE	F1	Acc-5
1	1	0.521	71.18	33.83
1	2	0.512	72.20	34.46
1	3	0.523	69.36	34.31
2	1	0.522	70.46	33.04
2	2	0.508	70.97	34.09
2	3	0.518	70.15	33.85

Table 6: Different Mamba layer settings and model performance on SIMS dataset.

on two factors: (1) the reconstruction loss weight, controlled by λ , and (2) the number of layers in the TC-Mamba and TQ-Mamba modules, which govern model capacity and computational cost.

Effect of Loss Weight λ We first examine the influence of the reconstruction loss weight λ on model performance. This hyperparameter controls the relative importance of the reconstruction loss in guiding the model to recover missing semantic information. We evaluated the model across $\lambda \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ on the MOSI dataset, and the results are visualized in Figure 5. The analysis reveals a clear trend: excessively small values (e.g., 0.1) provide insufficient supervision for reconstructing missing semantics, leading to suboptimal sentiment prediction. Conversely, excessively large values (e.g., 0.9) cause the reconstruction loss to dominate, which interferes with the model’s primary sentiment inference. A gradual increase in λ leads to consistent gains in F1 and Acc-7, with the overall best performance observed at $\lambda = 0.7$. This choice strikes an effective balance between reconstruction and sentiment prediction, underscoring the necessity of tuning λ appropriately.

Effect of Mamba Layers We further investigate the impact of varying the number of TC-Mamba and TQ-Mamba layers on model performance. The evaluation on the SIMS dataset is summarized in Table 6. The results show that using too few layers fails to capture the complex cross-modal interactions, while excessive layers introduce higher computational cost without consistent performance gains and may even risk overfitting. Overall, performance improves as the number of layers increases, with the (1,2) configuration of TC-Mamba and TQ-Mamba achieving the best trade-off between efficiency and Accuracy, delivering strong results at a moderate computational cost.

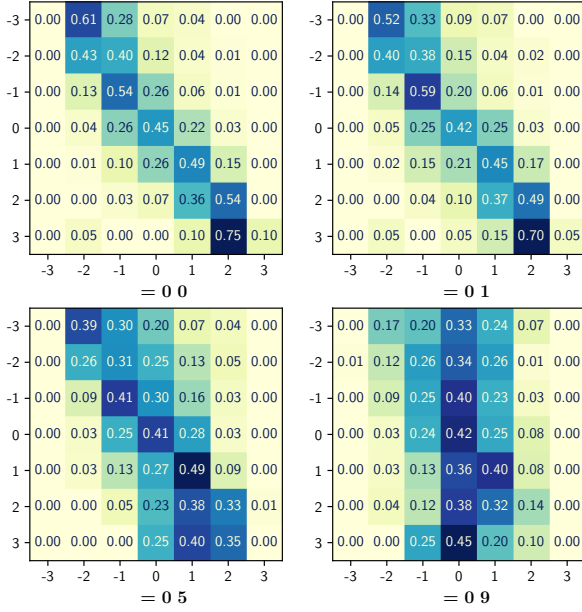


Figure 6: Seven-class confusion matrices of TF-Mamba on the MOSI dataset. Labels -3 to 3 represent sentiment levels from strongly negative to strongly positive.

4.8 Further Analysis

To further assess the effectiveness and robustness of our approach, we visualize the confusion matrices on the MOSI dataset under different missing rates in Figure 6. As expected, the model’s robustness and performance progressively decline as the missing rate r increases. At $r = 0$, the model achieves strong classification performance, with high diagonal values indicating effective category discrimination. When the missing rate increases to $r = 0.5$, classification Accuracy drops due to partial data loss, though the diagonal values remain relatively high, suggesting the model still correctly classifies most samples. At a severe missing rate of $r = 0.9$, excessive data missing leads the model to favor certain categories, exhibiting the typical “lazy” prediction behavior described in LNLN. Nevertheless, our model does not degenerate into random guessing, retaining a learned bias towards sentiment-relevant categories. These findings demonstrate the robustness of TF-Mamba in handling varying levels of incomplete data and underscore the importance of designing tailored fusion strategies for modeling incomplete data.

5 Conclusion

In this paper, a novel and efficient text-enhanced fusion Mamba (TF-Mamba) framework is developed to tackle random missing modality issues in the

MSA task. By integrating text enhancement strategies into the Mamba-based fusion architecture, our model effectively captures cross-modal interactions and informative multimodal representations with low computational overhead. Extensive experiments on three public benchmarks demonstrate that TF-Mamba outperforms current leading baselines across varying levels of data incompleteness. Further analysis reveals that TF-Mamba achieves notable reductions in parameters and computational costs relative to Transformer-based methods, underscoring the advantages and potential of Mamba-based fusion techniques in achieving robust MSA.

Limitations

Although TF-Mamba demonstrates strong robustness and efficiency under various missing modality scenarios, it has yet to be evaluated under more complex real-world missing patterns. The proposed method may experience minor performance degradation when applied to real-world scenarios. In addition, our work relies on pre-extracted unimodal features, which may limit its end-to-end optimization potential. In the future, we will explore more complex modality missing cases and develop suitable methods to overcome these limitations.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 62272025 and U22B2021), and by the Fund of the State Key Laboratory of Software Development Environment.

References

- Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. [Openface 2.0: Facial behavior analysis toolkit](#). In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66.
- Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. Covarep—a collaborative voice analysis repository for speech technologies. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 960–964. IEEE.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics*, pages 4171–4186.

- Wenhao Dong, Haodong Zhu, Shaohui Lin, Xiaoyan Luo, Yunhang Shen, Guodong Guo, and Baochang Zhang. 2025. Fusion-mamba for cross-modality object detection. *IEEE Transactions on Multimedia*.
- Xinyu Feng, Yuming Lin, Lihua He, You Li, Liang Chang, and Ya Zhou. 2024. Knowledge-guided dynamic modality attention fusion framework for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14755–14766.
- Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.
- Albert Gu, Karan Goel, Ankit Gupta, and Christopher Ré. 2022a. On the parameterization and initialization of diagonal state space models. *Advances in Neural Information Processing Systems*, 35:35971–35983.
- Albert Gu, Karan Goel, and Christopher Ré. 2022b. [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.
- Zirun Guo, Tao Jin, and Zhou Zhao. 2024. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1726–1736.
- Wei Han, Hui Chen, and Soujanya Poria. 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*, pages 9180–9192.
- Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1122–1131.
- iMotions. 2017. [Facial expression analysis](#).
- Xilin Jiang, Cong Han, and Nima Mesgarani. 2025. Dual-path mamba: Short and long-term bidirectional selective structured state space models for speech separation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Kai Li, Guo Chen, Runxuan Yang, and Xiaolin Hu. 2024a. Spmamba: State-space model is all you need in speech separation. *arXiv preprint arXiv:2404.02063*.
- Mingcheng Li, Dingkan Yang, Yuxuan Lei, Shunli Wang, Shuaibing Wang, Liuzhen Su, Kun Yang, Yuzheng Wang, Mingyang Sun, and Lihua Zhang. 2024b. A unified self-distillation framework for multimodal sentiment analysis with uncertain missing modalities. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 10074–10082.
- Mingcheng Li, Dingkan Yang, Yang Liu, Shunli Wang, Jiawei Chen, Shuaibing Wang, Jinjie Wei, Yue Jiang, Qingyao Xu, Xiaolu Hou, and 1 others. 2024c. Toward robust incomplete multimodal sentiment analysis via hierarchical representation learning. *Advances in Neural Information Processing Systems*, 37:28515–28536.
- Wenbing Li, Hang Zhou, Junqing Yu, Zikai Song, and Wei Yang. 2024d. Coupled mamba: Enhanced multimodal fusion with coupled state space model. *Advances in Neural Information Processing Systems*, 37:59808–59832.
- Xiang Li, Zhiqiang Dong, Xianfu Cheng, Dezhuang Miao, Haijun Zhang, Tianbo Wang, Xiaoming Zhang, and Zhoujun Li. 2025a. A multi-scale representation and multi-level decision learning network for multimodal sentiment analysis. *Expert Systems with Applications*, page 129341.
- Xiang Li, Ming Lu, Ziming Guo, and Xiaoming Zhang. 2024e. Adaptive token selection and fusion network for multimodal sentiment analysis. In *International Conference on Multimedia Modeling*, pages 228–241. Springer.
- Xiang Li, Haijun Zhang, Zhiqiang Dong, Xianfu Cheng, Yun Liu, and Xiaoming Zhang. 2025b. Learning fine-grained representation with token-level alignment for multimodal sentiment analysis. *Expert Systems with Applications*, 269:126274.
- Yan Li, Yifei Xing, Xiangyuan Lan, Xin Li, Haifeng Chen, and Dongmei Jiang. 2025c. Alignmamba: Enhancing multimodal mamba with local and global cross-modal alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24774–24784.
- Yong Li, Yuanzhi Wang, and Zhen Cui. 2023. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6631–6640.
- Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. 2022. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 14(3):2276–2289.
- Huisheng Mao, Ziqi Yuan, Hua Xu, Wenmeng Yu, Yihe Liu, and Kai Gao. 2022. M-sena: An integrated platform for multimodal sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 204–213.
- Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. *SciPy*, 2015:18–24.

- Yanyuan Qiao, Zheng Yu, Longteng Guo, Sihan Chen, Zijia Zhao, Mingzhen Sun, Qi Wu, and Jing Liu. 2024. VI-mamba: Exploring state space models for multimodal learning. *arXiv preprint arXiv:2403.13600*.
- Hao Sun, Hongyi Wang, Jiaqing Liu, Yen-Wei Chen, and Lanfen Lin. 2022. Cubemlp: An mlp-based model for multimodal sentiment analysis and depression estimation. In *Proceedings of the 30th ACM international conference on multimedia*, pages 3722–3729.
- Kaiwei Sun and Mi Tian. 2025. Sequential fusion of text-close and text-far representations for multimodal sentiment analysis. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 40–49.
- Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. 2023. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 15(1):309–325.
- Di Wang, Xutong Guo, Yumin Tian, Jinhui Liu, LiHuo He, and Xuemei Luo. 2023. Tetfn: A text enhanced transformer fusion network for multimodal sentiment analysis. *Pattern Recognition*, 136:109259.
- Di Wang, Shuai Liu, Quan Wang, Yumin Tian, Lihuo He, and Xinbo Gao. 2022. Cross-modal enhancement network for multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 25:4909–4921.
- Lan Wang, Junjie Peng, Cangzhi Zheng, Tong Zhao, and Li'an Zhu. 2024. A cross modal hierarchical fusion multimodal sentiment analysis method based on multi-task learning. *Information Processing & Management*, 61(3):103675.
- Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. 2022. Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the 30th ACM international conference on multimedia*, pages 1642–1651.
- Zefan Yang, Jiajin Zhang, Ge Wang, Mannudeep K Kalra, and Pingkun Yan. 2024. Cardiovascular disease detection from multi-view chest x-rays with bi-mamba. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 134–144. Springer.
- Jiaxin Ye, Junping Zhang, and Hongming Shan. 2025. Depmamba: Progressive fusion mamba for multimodal depression detection. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3718–3727.
- Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 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, pages 10790–10797.
- Ziqi Yuan, Wei Li, Hua Xu, and Wenmeng Yu. 2021. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In *Proceedings of the 29th ACM international conference on multimedia*, pages 4400–4407.
- Ziqi Yuan, Yihe Liu, Hua Xu, and Kai Gao. 2023. Noise imitation based adversarial training for robust multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 26:529–539.
- Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88.
- AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 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)*, pages 2236–2246.
- Ying Zeng, Wenjun Yan, Sijie Mai, and Haifeng Hu. 2024. Disentanglement translation network for multimodal sentiment analysis. *Information Fusion*, 102:102031.
- Haoyu Zhang, Wenbin Wang, and Tianshu Yu. 2024. Towards robust multimodal sentiment analysis with incomplete data. *Advances in Neural Information Processing Systems*, 37:55943–55974.
- Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yanyuan Liu, and Tianshu Yu. 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*, pages 756–767.
- Qianrui Zhou, Hua Xu, Hao Li, Hanlei Zhang, Xiaohan Zhang, Yifan Wang, and Kai Gao. 2024. Token-level contrastive learning with modality-aware prompting for multimodal intent recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 17114–17122.
- Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. 2024. [Vision mamba: Efficient visual representation learning with bidirectional state space model](#). In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*.

A Feature Extraction

To ensure fair and consistent comparisons with existing SOTA methods, we adopt the officially released features provided by the corresponding benchmark datasets, following the unified MMSA framework (Mao et al., 2022). All features are pre-extracted for text, audio, and visual modalities using standard toolkits, as detailed below.

Text Modality For MOSI and MOSEI, the bert-base-uncased (Devlin et al., 2019) model is used to encode raw utterances. For SIMS, the bert-base-chinese model is employed. The extracted text feature dimension is 768 for all datasets, with input sequence lengths of 50, 50, and 39 for MOSI, MOSEI, and SIMS, respectively.

Audio Modality For MOSI and MOSEI, COVAREP (Degottex et al., 2014) is applied to extract 5 and 74 low-level acoustic features, including pitch, glottal source parameters, and MFCCs, with input sequence lengths of 375 and 500, respectively. For SIMS, Librosa (McFee et al., 2015) is used to extract 33-dimensional audio features with a sequence length of 400.

Visual Modality For MOSI and MOSEI, Facet (iMotions, 2017) is utilized to extract 20 and 35 facial-related features, such as facial action units and head pose, with sequence lengths of 500. For SIMS, OpenFace 2.0 (Baltrusaitis et al., 2018) is used to extract 709-dimensional visual features with a sequence length of 55.

Sentiment Labels For MOSI and MOSEI, sentiment labels range from -3 to 3 , representing sentiment intensity from strongly negative to strongly positive. For SIMS, labels are scaled within -1 to 1 with the same polarity interpretation.

B Baseline Settings

We compare TF-Mamba against several MSA baselines. All experiments are conducted under the same dataset settings for fairness. We reproduce BI-Mamba and LNLN in our environment, while the results for other baselines are reported from (Zhang et al., 2024). Below are brief descriptions of them.

MISA (Hazarika et al., 2020) models both shared and private features across modalities to improve robustness in sentiment prediction.

Self-MM (Yu et al., 2021) generates unimodal labels and conducts multi-task training to capture both consistent and differential representations.

MMIM (Han et al., 2021) introduces mutual information maximization techniques for multimodal fusion to better capture correlated representations between modalities.

CENET (Wang et al., 2022) is a cross-modal enhancement network that adaptively aligns and fuses multimodal signals through cross-attentive interaction mechanisms.

TFR-Net (Yuan et al., 2021) is a Transformer-based model incorporating a fusion-reconstruction strategy to enhance sentiment prediction performance under incomplete modality conditions.

ALMT (Zhang et al., 2023) utilizes language features to suppress irrelevant or conflicting information from visual and audio inputs and learn complementary multimodal representations.

BI-Mamba (Yang et al., 2024) integrates bidirectional Mamba networks for multi-view medical image fusion. In our adaptation, we treat multimodal features as multi-view inputs to suit the model.

LNLN (Zhang et al., 2024) treats the language modality as dominant and introduces a dominant modality correction and dominant modality-based multimodal learning to enhance robustness against noisy and missing modality scenarios.

C Details of Robust Comparison

Tables 10, 11, and 12 present detailed robustness comparisons on the MOSI, MOSEI, and SIMS datasets, respectively. We observe that when the missing rate r is low, Self-MM achieves notable advantages across several evaluation metrics. However, as r increases, TF-Mamba consistently outperforms other methods on most metrics, demonstrating its ability to learn robust multimodal representations under varying levels of data incompleteness. Additionally, while models such as LNLN and TF-Mamba perform well under high missing modality rates, they often struggle to maintain optimal performance when modality missingness is low. Balancing robustness and Accuracy across different missing conditions remains challenging.

D Analysis of Complete Modality Missing

We conduct comprehensive experiments under complete modality missing conditions to further assess model robustness and examine the impact of discarding different modalities. The detailed results on the MOSI, MOSEI, and SIMS datasets are presented in Tables 7, 8, and 9, respectively.

These results consistently show that missing the text modality (T) leads to the largest performance drop, whether alone or combined with the audio (A) or visual (V) modality. This reveals the central role of text modality in MSA, as it typically provides the most direct and sentiment-rich information. In contrast, audio and visual modalities offer complementary cues, with their absence causing relatively minor degradation. This can be attributed to our text enhancement strategy, which provides effective task and sentiment information when other modalities are missing, thereby maintaining robust performance. Importantly, models trained under random missing assumptions struggle with complete modality loss, highlighting the need for customized fusion strategies to better handle structured or extreme missing scenarios.

Missing Condition	MOSI					
	Acc-7	Acc-5	Acc-2	F1	MAE	Corr
Missing T	19.39	19.97	52.59 / 53.79	51.33 / 52.31	1.505	0.108
Missing A	42.71	49.71	83.84 / 81.92	83.88 / 81.90	0.782	0.765
Missing V	42.27	48.40	82.32 / 81.05	82.42 / 81.10	0.778	0.772
Missing T & A	16.03	16.76	57.01 / 55.83	54.41 / 53.10	1.446	0.038
Missing T & V	16.91	19.10	55.49 / 56.27	54.86 / 55.45	1.590	0.183
Missing A & V	41.40	47.52	84.60 / 82.22	84.54 / 82.07	0.783	0.767
Average	29.78	33.58	69.31 / 68.51	68.57 / 67.65	1.147	0.439
TF-Mamba	33.95	37.74	73.46 / 72.54	73.59 / 72.57	1.035	0.548

Table 7: Performance of TF-Mamba under complete modality missing settings on MOSI dataset.

Method	MOSEI					
	Acc-7	Acc-5	Acc-2	F1	MAE	Corr
Missing T	36.40	36.40	60.68 / 66.52	61.16 / 63.94	0.943	0.152
Missing A	49.58	50.78	83.32 / 81.99	83.16 / 82.02	0.569	0.742
Missing V	50.44	52.11	83.13 / 83.30	82.71 / 83.15	0.565	0.738
Missing T & A	37.39	37.39	61.81 / 69.14	60.27 / 62.93	0.886	0.142
Missing T & V	41.36	41.36	62.85 / 71.02	48.51 / 58.99	0.855	0.124
Missing A & V	50.01	51.43	83.68 / 78.06	83.61 / 78.87	0.608	0.737
Average	44.20	44.91	72.59 / 75.00	69.90 / 71.65	0.738	0.439
TF-Mamba	45.66	46.64	77.34 / 77.61	77.18 / 77.43	0.673	0.578

Table 8: Performance of TF-Mamba under complete modality missing settings on MOSEI dataset.

Method	SIMS					
	Acc-5	Acc-3	Acc-2	F1	MAE	Corr
Missing T	26.91	49.67	69.37	56.82	0.738	0.043
Missing A	35.01	61.71	73.96	74.18	0.456	0.535
Missing V	36.76	62.58	77.68	77.10	0.447	0.539
Missing T & A	20.79	34.79	68.71	56.50	0.759	0.040
Missing T & V	18.60	30.63	36.76	32.93	0.907	0.014
Missing A & V	35.23	61.27	78.56	77.33	0.459	0.532
Average	28.88	50.11	67.51	62.48	0.628	0.284
TF-Mamba	34.46	55.51	74.68	72.20	0.512	0.386

Table 9: Performance of TF-Mamba under complete modality missing settings on SIMS dataset.

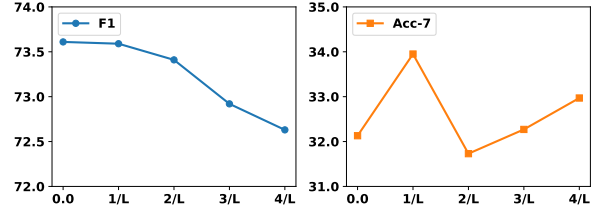


Figure 7: Model performance: F1 and Acc-7 on the MOSI dataset under different threshold settings.

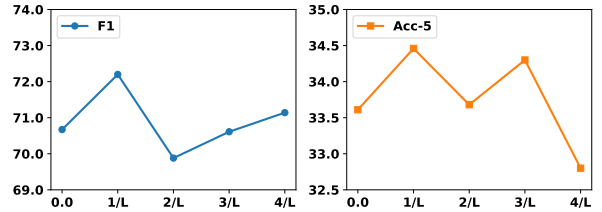


Figure 8: Model performance: F1 and Acc-5 on the SIMS dataset under different threshold settings.

E Analysis of Threshold Selection in TME

In the TME module, a threshold θ is introduced to generate binary masks over the similarity matrix between text tokens and other modalities. This mechanism suppresses irrelevant information by filtering out weakly correlated token pairs. The threshold is set as the average similarity value $1/L$, where L denotes the number of text tokens. Token pairs with similarity scores below the threshold are masked, whereas those exceeding the threshold are retained, thereby mitigating the influence of weak correlations while ensuring the preservation of informative semantic relationships.

To assess robustness, we conduct a sensitivity analysis by varying the threshold within $0, 1/L, 2/L, 3/L, 4/L$ on the MOSI and SIMS datasets. The corresponding model performance is shown in Figure 7 and Figure 8. From Figure 7 (MOSI), a threshold of $1/L$ provides the overall best performance across F1 and Acc-7, whereas $\theta = 0$ retains excessive irrelevant information and higher thresholds gradually remove important cues. Figure 8 (SIMS) exhibits a similar trend: $1/L$ consistently yields stable and optimal performance across both F1 and Acc-5, while $\theta = 0$ degrades performance and overly high thresholds suppress too much semantic information. Overall, this analysis confirms $1/L$ as an effective threshold choice, striking a balance between noise suppression and semantic retention, and demonstrating consistent advantages across both datasets.

Method	Acc-7	Acc-5	Acc-2	F1	MAE	Corr	Method	Acc-7	Acc-5	Acc-2	F1	MAE	Corr
Random Missing Rate $r = 0$							Random Missing Rate $r = 0.5$						
MISA	43.05	48.30	82.78 / 81.24	82.83 / 81.23	0.771	0.777	MISA	28.14	30.61	70.53 / 69.34	70.50 / 69.20	1.124	0.519
Self-MM	42.81	52.38	85.22 / 83.24	85.19 / 83.26	0.720	0.790	Self-MM	26.97	31.39	67.43 / 67.54	64.27 / 66.81	1.129	0.503
MMIM	45.92	49.85	83.43 / 81.97	83.43 / 81.94	0.744	0.778	MMIM	28.23	29.89	68.09 / 66.52	66.15 / 64.59	1.128	0.501
TFR-Net	40.82	47.91	83.64 / 81.68	83.57 / 81.61	0.805	0.760	TFR-Net	25.85	30.71	64.83 / 63.02	58.04 / 56.64	1.270	0.443
CENET	43.20	50.39	83.08 / 81.49	83.06 / 81.48	0.748	0.785	CENET	28.33	30.90	72.46 / 66.08	71.10 / 63.50	1.130	0.496
ALMT	42.37	48.49	84.91 / 82.75	85.01 / 82.94	0.752	0.768	ALMT	28.42	31.25	68.24 / 65.94	69.74 / 68.54	1.138	0.485
BI-Mamba	39.21	43.15	82.77 / 81.20	82.82 / 81.18	0.814	0.740	BI-Mamba	31.78	34.40	73.48 / 73.32	73.64 / 73.39	1.105	0.497
LNLN	40.77	46.41	83.28 / 79.78	83.17 / 79.73	0.764	0.772	LNLN	33.82	37.36	73.12 / 72.06	72.94 / 72.02	1.075	0.509
TF-Mamba	44.31	50.58	83.69 / 81.63	83.71 / 81.58	0.762	0.774	TF-Mamba	33.67	37.46	75.00 / 74.20	75.15 / 74.27	1.044	0.557
Random Missing Rate $r = 0.1$							Random Missing Rate $r = 0.6$						
MISA	40.28	46.21	80.18 / 79.01	80.21 / 78.97	0.847	0.721	MISA	24.68	27.12	66.97 / 65.84	66.94 / 65.69	1.200	0.441
Self-MM	40.33	49.03	81.40 / 80.03	81.19 / 80.03	0.812	0.728	Self-MM	24.34	27.31	63.47 / 63.36	58.94 / 62.07	1.209	0.425
MMIM	42.61	46.65	79.98 / 78.13	79.83 / 77.99	0.825	0.718	MMIM	25.41	27.11	63.67 / 62.49	60.87 / 59.48	1.208	0.418
TFR-Net	38.63	45.82	79.27 / 77.99	78.70 / 77.61	0.872	0.705	TFR-Net	24.05	28.33	61.64 / 59.47	52.44 / 50.53	1.371	0.363
CENET	40.13	46.60	80.08 / 78.38	79.91 / 78.20	0.837	0.719	CENET	24.54	26.53	67.58 / 61.47	64.87 / 57.86	1.215	0.415
ALMT	39.84	45.48	80.90 / 78.67	81.15 / 79.08	0.843	0.703	ALMT	25.41	27.36	64.53 / 62.15	66.81 / 65.87	1.214	0.407
BI-Mamba	38.78	42.71	81.25 / 80.03	81.34 / 80.06	0.873	0.699	BI-Mamba	29.88	29.01	67.38 / 67.35	67.41 / 66.75	1.193	0.409
LNLN	39.26	44.70	80.99 / 78.96	80.86 / 78.92	0.834	0.714	LNLN	29.98	33.23	67.58 / 66.71	67.41 / 67.53	1.177	0.421
TF-Mamba	42.86	48.40	81.86 / 80.03	81.87 / 79.97	0.824	0.732	TF-Mamba	30.76	33.53	68.60 / 68.37	68.79 / 68.45	1.127	0.487
Random Missing Rate $r = 0.2$							Random Missing Rate $r = 0.7$						
MISA	36.25	41.55	77.54 / 76.34	77.58 / 76.30	0.939	0.654	MISA	21.14	23.27	65.09 / 63.89	65.07 / 63.74	1.257	0.381
Self-MM	36.64	43.98	78.15 / 76.48	77.76 / 76.51	0.901	0.660	Self-MM	20.70	23.81	61.74 / 61.46	55.11 / 58.97	1.271	0.339
MMIM	39.07	42.66	76.42 / 74.54	76.12 / 74.22	0.918	0.651	MMIM	22.35	24.00	61.23 / 59.18	57.15 / 54.36	1.267	0.342
TFR-Net	34.70	40.13	74.70 / 73.52	73.57 / 72.70	0.987	0.622	TFR-Net	23.71	26.92	59.91 / 57.34	48.41 / 45.48	1.454	0.276
CENET	38.00	42.32	77.49 / 74.64	77.35 / 74.28	0.916	0.654	CENET	22.35	23.57	63.82 / 59.43	53.79 / 54.22	1.269	0.335
ALMT	35.33	40.33	77.64 / 75.70	77.94 / 76.24	0.927	0.645	ALMT	23.71	24.97	61.84 / 59.67	65.30 / 65.19	1.266	0.336
BI-Mamba	33.24	39.36	78.35 / 77.26	78.47 / 77.31	0.948	0.648	BI-Mamba	27.41	27.26	64.79 / 65.01	64.72 / 64.81	1.250	0.333
LNLN	37.22	42.13	78.51 / 75.95	78.42 / 75.95	0.908	0.653	LNLN	27.26	30.52	64.94 / 63.95	64.85 / 63.98	1.244	0.341
TF-Mamba	39.21	44.75	80.49 / 79.15	80.56 / 79.17	0.879	0.693	TF-Mamba	27.26	29.30	67.23 / 66.91	67.41 / 66.98	1.196	0.411
Random Missing Rate $r = 0.3$							Random Missing Rate $r = 0.8$						
MISA	34.60	38.97	75.76 / 74.54	75.82 / 74.51	0.989	0.618	MISA	19.92	20.99	63.56 / 62.24	63.16 / 61.67	1.311	0.321
Self-MM	34.89	40.67	76.37 / 74.98	75.68 / 74.94	0.967	0.614	Self-MM	19.29	22.11	59.55 / 58.26	49.98 / 53.56	1.313	0.282
MMIM	36.83	40.43	74.08 / 71.91	73.47 / 71.28	0.974	0.612	MMIM	20.26	21.77	58.33 / 55.30	52.46 / 47.89	1.312	0.287
TFR-Net	32.55	38.34	72.36 / 71.28	70.12 / 69.58	1.065	0.572	TFR-Net	23.23	27.70	58.49 / 55.98	44.70 / 41.88	1.497	0.155
CENET	34.74	38.97	76.83 / 72.01	76.56 / 71.30	0.983	0.605	CENET	21.14	21.67	60.93 / 57.53	54.68 / 50.80	1.314	0.274
ALMT	33.04	37.17	75.15 / 72.94	75.51 / 73.66	0.992	0.596	ALMT	23.13	23.66	60.37 / 58.31	65.45 / 66.14	1.310	0.273
BI-Mamba	33.38	37.17	75.91 / 75.07	76.06 / 75.14	0.998	0.597	BI-Mamba	24.20	26.38	61.13 / 60.93	61.10 / 60.78	1.289	0.300
LNLN	36.44	40.62	76.12 / 74.68	76.02 / 74.69	0.963	0.600	LNLN	23.13	25.70	62.04 / 60.40	61.85 / 60.32	1.294	0.288
TF-Mamba	37.76	42.27	77.74 / 76.53	77.85 / 76.56	0.932	0.645	TF-Mamba	24.93	26.38	63.57 / 63.12	63.77 / 63.20	1.258	0.353
Random Missing Rate $r = 0.4$							Random Missing Rate $r = 0.9$						
MISA	32.65	35.37	73.88 / 72.59	73.88 / 72.49	1.041	0.585	MISA	17.78	18.41	58.64 / 58.21	56.84 / 56.19	1.369	0.226
Self-MM	31.20	36.30	73.17 / 71.96	71.74 / 71.75	1.027	0.579	Self-MM	18.32	19.78	58.59 / 55.25	46.16 / 47.46	1.353	0.197
MMIM	33.38	35.76	70.84 / 68.90	69.69 / 67.80	1.034	0.576	MMIM	18.95	19.53	55.29 / 51.65	47.33 / 40.89	1.357	0.186
TFR-Net	30.17	35.76	68.75 / 67.74	64.71 / 64.41	1.142	0.537	TFR-Net	21.67	25.12	57.93 / 55.44	43.01 / 40.18	1.534	0.155
CENET	32.26	36.15	73.38 / 71.53	72.75 / 70.26	1.031	0.574	CENET	19.15	19.10	58.99 / 54.76	50.01 / 46.58	1.357	0.181
ALMT	31.44	35.03	73.12 / 71.14	73.85 / 72.47	1.045	0.560	ALMT	20.31	20.50	57.32 / 56.66	64.92 / 67.82	1.349	0.205
BI-Mamba	32.22	35.86	74.54 / 73.91	74.70 / 73.97	1.040	0.555	BI-Mamba	21.87	24.93	57.77 / 57.14	58.03 / 57.25	1.354	0.203
LNLN	35.42	38.87	75.30 / 73.67	75.23 / 73.66	0.995	0.575	LNLN	22.01	22.93	57.17 / 54.91	56.38 / 54.13	1.371	0.164
TF-Mamba	35.86	40.23	76.07 / 75.22	76.16 / 75.25	0.961	0.617	TF-Mamba	22.89	24.49	60.37 / 60.20	60.59 / 60.30	1.363	0.215

Table 10: Details of robust comparison on MOSI with different random missing rates.

Method	Acc-7	Acc-5	Acc-2	F1	MAE	Corr	Method	Acc-7	Acc-5	Acc-2	F1	MAE	Corr
Random Missing Rate $r = 0$							Random Missing Rate $r = 0.5$						
MISA	51.79	53.85	85.28 / 84.10	85.10 / 83.75	0.552	0.759	MISA	38.12	36.05	67.38 / 73.21	58.38 / 64.14	0.834	0.492
Self-MM	53.89	55.72	85.34 / 84.68	85.11 / 84.66	0.531	0.764	Self-MM	42.70	43.14	71.97 / 75.81	67.40 / 70.38	0.733	0.477
MMIM	50.76	53.04	83.53 / 81.65	83.39 / 81.41	0.576	0.724	MMIM	38.68	39.21	71.75 / 74.45	67.70 / 67.96	0.775	0.470
TFR-Net	53.71	47.91	84.96 / 84.65	84.71 / 84.34	0.550	0.745	TFR-Net	45.00	30.71	71.53 / 75.69	66.88 / 70.07	0.730	0.471
CENET	54.39	56.12	85.49 / 82.30	85.41 / 82.60	0.531	0.770	CENET	45.12	45.52	73.33 / 77.16	69.80 / 74.14	0.720	0.515
ALMT	52.18	53.89	85.62 / 83.99	85.69 / 84.53	0.542	0.752	ALMT	37.82	38.34	77.40 / 77.48	77.73 / 77.80	0.683	0.461
BI-Mamba	48.40	49.71	83.65 / 81.84	83.60 / 81.77	0.620	0.677	BI-Mamba	45.44	46.21	77.41 / 77.48	77.25 / 77.17	0.699	0.550
LNLN	50.66	51.94	84.14 / 83.61	84.53 / 84.02	0.572	0.735	LNLN	44.90	45.59	76.44 / 78.10	77.23 / 79.30	0.710	0.529
TF-Mamba	52.26	53.83	83.82 / 82.89	83.71 / 82.92	0.556	0.748	TF-Mamba	45.68	46.60	78.59 / 77.94	78.34 / 77.78	0.676	0.583
Random Missing Rate $r = 0.1$							Random Missing Rate $r = 0.6$						
MISA	50.13	51.34	82.21 / 82.28	81.28 / 80.79	0.598	0.722	MISA	36.16	33.30	65.55 / 72.30	54.64 / 62.12	0.875	0.415
Self-MM	51.80	53.18	83.03 / 83.79	82.43 / 83.23	0.564	0.725	Self-MM	41.47	41.75	69.33 / 73.93	63.01 / 66.76	0.762	0.401
MMIM	49.09	51.19	82.00 / 81.09	81.57 / 80.15	0.602	0.696	MMIM	37.13	37.48	68.83 / 73.16	63.09 / 65.43	0.808	0.402
TFR-Net	52.29	45.82	82.92 / 83.31	82.25 / 82.40	0.573	0.715	TFR-Net	43.88	28.33	68.80 / 74.05	62.51 / 67.07	0.762	0.397
CENET	52.83	54.23	83.75 / 82.41	83.42 / 82.34	0.556	0.739	CENET	44.45	44.64	70.50 / 75.39	65.27 / 70.86	0.749	0.446
ALMT	49.98	51.38	84.14 / 82.84	84.23 / 83.04	0.583	0.718	ALMT	35.99	36.30	74.98 / 76.26	75.44 / 76.71	0.710	0.395
BI-Mamba	48.36	49.43	82.97 / 80.79	82.92 / 80.61	0.630	0.662	BI-Mamba	43.16	43.92	75.21 / 75.17	74.90 / 74.94	0.730	0.512
LNLN	49.96	51.25	83.32 / 82.73	83.66 / 82.91	0.591	0.712	LNLN	43.52	44.00	73.82 / 76.50	75.03 / 78.33	0.736	0.471
TF-Mamba	50.53	52.07	83.16 / 82.68	83.03 / 82.69	0.570	0.730	TF-Mamba	43.96	44.73	75.89 / 75.77	75.63 / 75.59	0.709	0.534
Random Missing Rate $r = 0.2$							Random Missing Rate $r = 0.7$						
MISA	47.24	47.66	77.84 / 79.93	75.56 / 76.88	0.659	0.674	MISA	34.54	31.21	64.28 / 71.71	51.82 / 60.65	0.906	0.344
Self-MM	49.44	50.51	80.84 / 82.33	79.76 / 81.17	0.604	0.678	Self-MM	39.93	40.12	66.79 / 72.55	58.05 / 63.45	0.786	0.329
MMIM	46.27	47.99	79.93 / 79.66	79.08 / 77.68	0.642	0.653	MMIM	35.25	35.47	66.89 / 72.26	58.90 / 63.26	0.834	0.341
TFR-Net	51.04	40.13	80.47 / 81.61	79.29 / 79.99	0.604	0.672	TFR-Net	42.91	26.92	66.64 / 72.77	58.32 / 64.02	0.786	0.322
CENET	50.72	51.85	81.46 / 81.62	80.78 / 81.17	0.590	0.698	CENET	43.93	44.03	67.50 / 73.39	59.88 / 67.02	0.776	0.384
ALMT	46.61	47.82	82.71 / 81.65	82.82 / 81.83	0.607	0.669	ALMT	34.78	34.95	71.62 / 73.98	72.24 / 74.54	0.743	0.315
BI-Mamba	47.91	49.26	81.56 / 80.30	81.51 / 80.03	0.644	0.641	BI-Mamba	42.20	43.16	72.04 / 73.45	71.44 / 73.04	0.754	0.466
LNLN	48.75	49.95	81.70 / 81.68	81.95 / 81.89	0.616	0.677	LNLN	42.22	42.56	71.55 / 74.74	73.49 / 77.40	0.762	0.408
TF-Mamba	49.17	50.50	82.55 / 81.84	82.40 / 81.83	0.588	0.710	TF-Mamba	42.78	43.40	73.50 / 73.49	73.26 / 73.19	0.747	0.469
Random Missing Rate $r = 0.3$							Random Missing Rate $r = 0.8$						
MISA	43.99	43.40	73.32 / 77.28	68.91 / 72.25	0.724	0.615	MISA	33.29	29.51	63.43 / 71.30	49.95 / 59.69	0.927	0.267
Self-MM	47.23	48.07	77.63 / 79.99	75.69 / 77.74	0.653	0.610	Self-MM	38.69	38.78	65.07 / 71.83	54.44 / 61.49	0.805	0.259
MMIM	43.25	44.73	77.08 / 77.79	75.46 / 74.49	0.690	0.597	MMIM	33.64	33.71	64.97 / 71.57	54.76 / 61.45	0.858	0.269
TFR-Net	48.75	38.34	77.48 / 79.29	75.43 / 76.52	0.650	0.604	TFR-Net	42.23	27.70	65.05 / 71.95	54.91 / 61.82	0.807	0.241
CENET	48.49	49.37	78.65 / 80.02	77.34 / 78.94	0.636	0.640	CENET	42.71	42.74	65.88 / 72.16	56.80 / 64.67	0.798	0.316
ALMT	43.04	44.05	80.94 / 79.94	81.15 / 80.20	0.632	0.598	ALMT	34.01	34.09	68.15 / 71.48	69.12 / 72.28	0.774	0.231
BI-Mamba	46.98	48.38	80.27 / 79.48	80.17 / 79.13	0.662	0.615	BI-Mamba	42.05	41.40	70.12 / 71.35	69.08 / 70.90	0.775	0.421
LNLN	47.36	48.40	80.11 / 80.45	80.44 / 80.91	0.648	0.629	LNLN	40.76	40.97	68.62 / 72.86	71.83 / 76.80	0.791	0.325
TF-Mamba	47.89	49.00	80.82 / 81.22	80.58 / 81.10	0.613	0.675	TF-Mamba	40.37	40.93	70.34 / 71.60	70.16 / 71.27	0.786	0.408
Random Missing Rate $r = 0.4$							Random Missing Rate $r = 0.9$						
MISA	40.87	39.53	70.46 / 75.04	64.02 / 67.93	0.780	0.561	MISA	32.29	28.03	62.95 / 71.07	48.80 / 59.12	0.941	0.180
Self-MM	44.40	45.04	75.02 / 78.09	72.01 / 74.48	0.694	0.554	Self-MM	37.46	37.50	63.85 / 71.24	51.32 / 59.72	0.821	0.188
MMIM	40.84	41.86	74.56 / 76.15	71.98 / 71.40	0.732	0.542	MMIM	32.61	32.67	63.69 / 71.10	51.26 / 59.99	0.877	0.197
TFR-Net	46.70	35.76	74.74 / 77.65	71.67 / 73.71	0.688	0.548	TFR-Net	41.73	25.12	63.64 / 71.34	52.02 / 59.99	0.820	0.175
CENET	47.12	47.74	76.03 / 78.57	73.87 / 76.75	0.678	0.587	CENET	42.08	42.08	64.14 / 70.42	54.27 / 62.33	0.814	0.254
ALMT	40.40	41.21	79.40 / 79.16	79.68 / 79.50	0.651	0.536	ALMT	34.40	34.40	61.41 / 68.65	63.32 / 69.83	0.810	0.138
BI-Mamba	45.98	46.53	78.62 / 78.58	78.55 / 78.26	0.678	0.588	BI-Mamba	40.74	39.56	66.32 / 68.77	64.11 / 67.96	0.817	0.321
LNLN	45.99	46.88	78.49 / 79.70	78.98 / 80.46	0.673	0.592	LNLN	40.10	40.19	64.83 / 71.51	70.60 / 77.52	0.820	0.221
TF-Mamba	46.73	47.80	80.02 / 80.02	79.80 / 79.80	0.639	0.638	TF-Mamba	37.24	37.56	64.75 / 68.68	64.87 / 68.18	0.851	0.291

Table 11: Details of robust comparison on MOSEI with different random missing rates.

Method	Acc-5	Acc-3	Acc-2	F1	MAE	Corr	Method	Acc-5	Acc-3	Acc-2	F1	MAE	Corr
Random Missing Rate $r = 0$							Random Missing Rate $r = 0.5$						
MISA	40.55	63.38	78.19	77.22	0.449	0.576	MISA	30.56	54.78	71.26	64.16	0.552	0.367
Self-MM	40.77	64.92	78.26	78.00	0.421	0.584	Self-MM	32.02	53.90	71.41	67.11	0.517	0.390
MMIM	37.42	60.69	75.42	73.10	0.475	0.528	MMIM	33.41	52.37	68.49	64.81	0.553	0.336
TFR-Net	33.85	54.12	69.15	58.44	0.562	0.254	TFR-Net	24.65	52.37	67.47	58.66	0.685	0.171
CENET	23.85	54.05	68.71	57.82	0.578	0.137	CENET	23.12	54.05	68.71	57.92	0.588	0.107
ALMT	23.41	54.78	75.64	76.27	0.527	0.536	ALMT	18.38	47.12	68.27	71.22	0.563	0.395
BI-Mamba	41.58	63.02	76.59	76.93	0.433	0.574	BI-Mamba	30.42	56.46	71.33	71.06	0.540	0.334
LNLN	38.51	61.78	77.68	77.22	0.448	0.561	LNLN	35.81	57.70	73.74	69.84	0.515	0.416
TF-Mamba	37.86	61.93	79.65	78.92	0.441	0.548	TF-Mamba	37.20	58.64	75.71	73.61	0.495	0.424
Random Missing Rate $r = 0.1$							Random Missing Rate $r = 0.6$						
MISA	38.88	63.02	77.39	75.82	0.461	0.561	MISA	27.72	53.97	70.46	61.81	0.578	0.286
Self-MM	40.26	63.53	77.32	76.76	0.433	0.563	Self-MM	29.10	51.86	70.02	64.21	0.548	0.313
MMIM	37.27	60.90	74.25	72.08	0.473	0.529	MMIM	29.18	49.31	67.91	63.86	0.578	0.270
TFR-Net	30.12	53.25	68.85	59.38	0.596	0.203	TFR-Net	24.80	52.59	67.03	58.30	0.696	0.157
CENET	22.83	53.98	68.57	57.36	0.580	0.136	CENET	22.46	53.69	69.00	58.64	0.592	0.102
ALMT	22.10	55.14	74.40	75.19	0.530	0.537	ALMT	18.67	43.69	66.81	70.69	0.574	0.322
BI-Mamba	40.70	62.80	74.84	75.21	0.448	0.561	BI-Mamba	26.04	50.98	68.49	67.84	0.554	0.281
LNLN	37.27	62.80	78.12	77.54	0.450	0.554	LNLN	32.31	54.85	71.77	65.57	0.543	0.345
TF-Mamba	36.98	62.36	79.65	78.79	0.445	0.550	TF-Mamba	32.82	54.92	73.09	69.44	0.523	0.370
Random Missing Rate $r = 0.2$							Random Missing Rate $r = 0.7$						
MISA	38.15	59.23	74.33	71.70	0.489	0.490	MISA	24.87	52.52	69.95	59.54	0.601	0.167
Self-MM	38.37	61.71	74.98	73.71	0.464	0.500	Self-MM	25.53	50.62	69.58	62.28	0.571	0.198
MMIM	37.27	57.33	72.36	69.80	0.504	0.460	MMIM	28.59	46.53	66.89	62.23	0.595	0.190
TFR-Net	29.03	53.61	68.64	59.74	0.619	0.191	TFR-Net	23.78	52.30	67.18	58.15	0.707	0.163
CENET	22.25	54.20	68.57	57.64	0.583	0.132	CENET	21.81	53.32	67.69	57.87	0.599	0.070
ALMT	21.08	53.17	72.65	73.90	0.541	0.485	ALMT	18.02	38.66	65.57	70.27	0.586	0.218
BI-Mamba	37.20	59.74	72.65	72.56	0.472	0.506	BI-Mamba	24.95	45.08	66.30	63.60	0.605	0.148
LNLN	35.59	60.69	76.29	75.53	0.468	0.509	LNLN	29.91	50.47	70.17	61.28	0.566	0.236
TF-Mamba	38.29	61.05	78.77	77.88	0.459	0.507	TF-Mamba	28.45	48.80	71.77	67.46	0.570	0.248
Random Missing Rate $r = 0.3$							Random Missing Rate $r = 0.8$						
MISA	36.40	59.30	74.11	70.40	0.505	0.464	MISA	22.69	52.22	69.37	57.82	0.610	0.092
Self-MM	37.93	59.81	74.76	72.85	0.474	0.487	Self-MM	22.03	50.77	69.51	60.68	0.585	0.138
MMIM	37.71	58.06	72.36	69.52	0.512	0.436	MMIM	22.32	44.35	65.28	60.53	0.607	0.145
TFR-Net	27.64	52.30	68.42	59.88	0.640	0.182	TFR-Net	22.97	52.74	67.54	57.55	0.721	0.100
CENET	21.44	54.05	68.42	57.41	0.578	0.175	CENET	21.73	52.15	67.47	58.44	0.599	0.074
ALMT	20.35	50.62	72.06	73.64	0.546	0.469	ALMT	18.60	34.06	64.19	69.64	0.597	0.133
BI-Mamba	33.04	58.21	72.21	72.39	0.505	0.451	BI-Mamba	26.04	50.11	66.96	61.71	0.607	0.129
LNLN	36.32	58.94	75.35	73.60	0.475	0.502	LNLN	25.82	49.16	69.66	59.42	0.587	0.164
TF-Mamba	39.82	58.42	75.93	74.72	0.468	0.485	TF-Mamba	26.48	46.39	71.55	66.58	0.607	0.139
Random Missing Rate $r = 0.4$							Random Missing Rate $r = 0.9$						
MISA	34.86	57.33	72.87	67.52	0.523	0.436	MISA	20.64	52.95	69.22	57.01	0.617	0.041
Self-MM	34.57	58.28	73.30	70.36	0.482	0.479	Self-MM	22.17	52.15	68.92	58.32	0.586	0.111
MMIM	34.57	55.36	69.95	66.49	0.533	0.399	MMIM	20.35	42.67	65.72	59.64	0.610	0.096
TFR-Net	25.31	51.86	67.91	59.16	0.664	0.176	TFR-Net	23.05	53.76	69.08	57.71	0.721	0.088
CENET	22.54	54.12	68.49	57.68	0.583	0.141	CENET	20.86	48.07	65.72	58.18	0.609	-0.002
ALMT	19.91	49.45	70.75	72.97	0.549	0.470	ALMT	19.47	26.91	66.23	73.76	0.596	0.076
BI-Mamba	32.17	55.58	70.90	70.75	0.509	0.419	BI-Mamba	26.91	47.48	67.61	60.58	0.627	0.049
LNLN	35.74	58.86	74.18	71.38	0.488	0.478	LNLN	23.49	44.86	69.29	57.05	0.595	0.128
TF-Mamba	40.04	60.39	75.49	73.72	0.470	0.477	TF-Mamba	26.70	42.23	65.21	60.87	0.648	0.114

Table 12: Details of robust comparison on SIMS with different random missing rates.