Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.083112

ARTICLE

TC-DSC: Text-Centric Hierarchical Dual-Stream Interaction for Incomplete Multimodal Sentiment Analysis

Jinjin Liu[1,2,3,*], **Changchang Fan**[1,2], **Qiulu Guo**[1,2], **Yihao Xu**[1,2] **and Penghui Ma**[1,2]

[1]School of Computer Science, ZhongYuan University of Technology, Zhengzhou, China
[2]Henan International Joint Laboratory of Artificial Intelligence Interpretability Reasoning and Application, Zhengzhou, China
[3]Henan Engineering Technology Research Center of Archives Data Analysis and Security, Zhengzhou, China

*Corresponding Author: Jinjin Liu. Email: 6406@zut.edu.cn

Received: 29 March 2026; Accepted: 04 June 2026; Published: 23 July 2026

ABSTRACT: Incomplete multimodal sentiment analysis has attracted increasing research interest in recent years. Existing methods attempt to recover missing modalities through generative reconstruction and text-enhanced fusion, but these approaches may be limited in preserving sentiment-relevant information and fully leveraging complementary and hierarchical cross-modal interactions, particularly under noisy or incomplete conditions. To address these challenges, we propose TC-DSC, a text-centric hierarchical dual-stream interaction framework for incomplete multimodal sentiment analysis. Rather than reconstructing raw signals, TC-DSC performs semantic alignment and consistency modeling in the feature space through structured interactions between a text-centric stream and auxiliary audio-visual streams. A multi-scale enhanced encoder is designed to improve the robustness of non-text modalities under noisy conditions. Furthermore, a hierarchical proxy layer enables bidirectional interaction, with the text modality serving as a semantic anchor to guide cross-modal alignment. A semantic distillation strategy is also incorporated to facilitate knowledge transfer in the feature space under modality missing. Extensive experiments on MOSI, MOSEI, and SIMS demonstrate that TC-DSC achieves competitive performance and consistent improvements under both complete and incomplete settings.

KEYWORDS: Multimodal sentiment analysis; incomplete multimodal learning; text-centric modeling; cross-modal alignment

1 Introduction

Multimodal Sentiment Analysis (MSA) aims to comprehensively understand human affective states by jointly modeling heterogeneous information across modalities, including text, audio, and vision. However, in practice, multimodal data are frequently incomplete due to device malfunctions, privacy-preserving masking, or other acquisition constraints. This incompleteness makes cross-modal semantic alignment unstable and further degrades the generalization ability of multimodal sentiment analysis models.

In recent years, various solutions have been proposed to address the missing-modality problem in multimodal sentiment analysis [1–10].

As a representative method, TF-Mamba [10] introduces a text-enhanced Mamba-based fusion framework to improve robustness and efficiency under missing modality conditions. However, it still faces several limitations: (i) its reconstruction of missing textual semantics may introduce additional uncertainty and

cause the learned representations to deviate from sentiment-relevant semantic information; (ii) its text-enhanced fusion mainly focuses on aligning and enriching non-text modalities with textual cues, while the complementary relationships among modalities and hierarchical semantic interactions are not sufficiently explored, limiting its ability to capture fine-grained affective cues under noisy or incomplete conditions.

To address the above issues, we propose TC-DSC, a Text-Centric Hierarchical Dual-Stream Interaction Framework for incomplete multimodal sentiment analysis. TC-DSC models incomplete multimodal data through a hierarchical process consisting of three stages: low-level enhancement of non-text features, mid-level bidirectional proxy interaction between the text and auxiliary audio-visual streams, and high-level semantic consistency distillation. Specifically, a multi-scale enhancement encoder is employed to bolster the robustness of audio-visual representations. A hierarchical proxy-based interaction mechanism is introduced to enable bidirectional information exchange between the textual semantic stream and the auxiliary audio-visual stream. Finally, a semantic distillation strategy is applied to regularize feature-space consistency between complete and incomplete inputs, ensuring robustness in multimodal fusion.

The main contributions of this work are summarized as follows:

- We propose TC-DSC, a hierarchical dual-stream interaction framework that models incomplete multimodal data by leveraging a text-centric architecture for stable semantic alignment and effective information fusion, improving cross-modal interaction.
- We introduce a hierarchical proxy layer that enables bidirectional interaction between the text and auxiliary streams, mitigating cross-modal discrepancies and ensuring consistent alignment even when one modality is absent.
- We develop a multi-scale feature enhancement strategy that uses parallel convolutional structures to improve robustness against intra-modal degradation and ensure high-quality representations for subsequent cross-modal interactions.

2 Related Work

Current multimodal sentiment analysis (MSA) methodologies can be primarily categorized into two paradigms: representation learning with full modality and robust learning with incomplete modality.

Full-Modality Representation Learning. Early research predominantly focused on designing sophisticated fusion mechanisms and interaction protocols to model intra-modal and inter-modal contextual relationships, assuming complete data. MMIM (Han et al., 2021 [11]) enhanced the discriminability of fused representations by maximizing cross-modal mutual information. ConKI (Yu et al., 2023 [12]) integrated domain-specific knowledge into multimodal representations via adapter architectures and hierarchical contrastive learning. SKESL (Qian et al., 2023 [13]) incorporated emotional knowledge and non-verbal behaviors into pre-trained representations through sentiment word masking. TETFN (Wang et al., 2023 [14]) utilized text-oriented multi-head attention to boost non-verbal features while preserving cross-modal consistency. Furthermore, MTMD (Lin and Hu, 2023 [15]) treated multimodal learning as a multi-task process, employing a teacher-student distillation framework to guide audio-visual subtasks. Recent advancements, such as KuDA (Feng et al., 2024 [16]), CFHF (Wang et al., 2024 [17]), MAMSA (Wang et al., 2025 [18]), and the work by Xu and Zhang (2025 [19]), have further refined adaptive fusion and interference reduction using dynamic attention, multi-task frameworks, and distance constraints. MAGCN (Xiao et al., 2022 [20]) integrates graph convolution and multi-head attention with textual sentiment knowledge to enhance cross-modal representation learning for multimodal sentiment analysis. However, these methods rely heavily on the integrity and stability of modal inputs, limiting their efficacy in real-world scenarios with missing data.

Incomplete-Modality Solutions. To address data incompleteness, two mainstream strategies have emerged: generative completion and robust joint learning. Generative completion aims to reconstruct missing signals or features from available modalities. For instance, EMT-DLFR (Sun et al., 2024 [1]) utilized a dual-level feature repair mechanism, while SMCMSA (Sun et al., 2024 [2]) and MTMSA (Liu et al., 2024 [3]) leveraged modality translation or similarity-based libraries to hallucinate missing information. TgRN (Shi et al., 2025 [4]) reconstructs missing modality features with textual guidance and further narrows the modality gap through progressive modality mixing and reconstruction gates. By contrast, robust joint learning avoids explicit reconstruction by imposing consistency constraints and knowledge transfer within the representation space. M2AF (Tu et al., 2025 [5]) employed meta-learning for cross-scenario generalization, while T-S-MSA (Zeng et al., 2022 [21]) and EMMR (Zeng et al., 2022 [22]) utilized pseudo-labels or residual autoencoders for semantic recovery.

Recent works like MissModal (Lin and Hu, 2023 [6]), UMDF (Li et al., 2024 [7]), and HRLF (Li et al., 2024 [8]) explore methods such as correlation decoupling and hierarchical alignment to learn robust joint representations. In addition to text-based reconstruction, text-guided approaches like TCTR (Yang et al., 2023 [9]) perform token-level reconstruction and semantic supplementation through contrastive learning, while TF-Mamba uses text-enhanced sequence modeling for efficient fusion when modalities are missing. Furthermore, distillation-based methods like CorrKD (Li et al., 2024 [23]) and HCKD (Xi et al., 2024 [24]) enhance incomplete-modality learning by transferring knowledge from complete-modality teachers to incomplete-modality students.

We observe that textual modality provides more stable and semantically rich information for sentiment analysis. Therefore, we propose a text-centric hierarchical dual-stream interaction framework to enable effective cross-modal alignment and robust representation learning under incomplete settings. Unlike reconstruction-based or distillation-based methods, TC-DSC explicitly models bidirectional hierarchical interactions between textual and audio-visual streams, enabling direct feature-space alignment under incomplete modalities rather than relying solely on pseudo-label transfer or token-level reconstruction.

3 Methodology

3.1 Model Framework

Fig. 1 shows the key components and workflow of the proposed Text-Centric Hierarchical Dual-Stream Interaction (TC-DSC) framework. TC-DSC consists of three main modules: Multimodal Data Processing, Text-Centric Bidirectional Interaction (TCBI), and Semantic Compensation Distillation (SCD). During training, both complete inputs and randomly masked incomplete inputs are constructed to simulate modality-incomplete scenarios. In TCBI, the Unimodal Feature Enhancement Encoder (UFEE) first extracts textual representations using a pre-trained language model and enhances audio-visual representations through multi-scale temporal convolution. Then, Bidirectional Attention Fusion (BAF) performs hierarchical text-guided interactions between the textual stream and the auxiliary audio-visual stream, allowing text to absorb complementary audio-visual cues while guiding auxiliary modalities toward a shared semantic space. SCD refines the learned representations by compensating for semantic discrepancies across modalities, thereby enhancing overall robustness.

3.2 Multimodal Data Processing

Multimodal Input and Random Missing. A multimodal sample comprises three modalities: text, visual, and audio, denoted as X_L, X_V, and X_A, respectively. The text modality is encoded with BERT (Devlin et al., 2019 [25]) to obtain token-level representations H_L. Visual and audio features are extracted using OpenFace (Baltrusaitis et al., 2018 [26]) and Librosa (McFee et al., 2015 [27]), respectively.

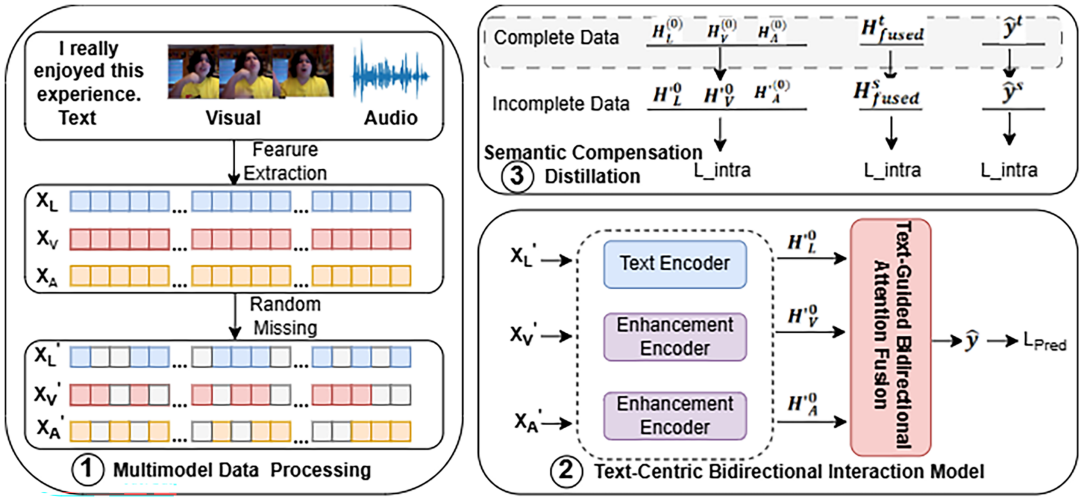

Figure 1: Overview of the proposed framework.

To simulate realistic scenarios with modality incompleteness, we randomly apply temporal masking to each modality during training. For the visual and audio modalities, masked positions are replaced with zero vectors, while for text, masked tokens are replaced with `[UNK]` while preserving special tokens. The masked inputs are denoted as:

$$X'_L, \quad X'_V, \quad X'_A, \tag{1}$$

where the missing patterns are independently sampled for each modality at every training epoch.

3.3 Text-Centric Bidirectional Interaction Module

3.3.1 Unimodal Feature Enhancement Encoder

Audio and visual modalities are often affected by noise and redundancy. To improve robustness, we introduce UFEE.

Given the input X_m (or its masked version X'_m), the modality representation is obtained as:

$$H_m^{(0)} = \mathrm{Enc}_m(X_m), \tag{2}$$

where $\mathrm{Enc}_m(\cdot)$ denotes a modality-enhancement encoder that integrates projection, multi-scale temporal convolution (with kernel sizes $k \in \{3,5,7\}$), and Transformer layers.

Where $\mathrm{Enc}_m(\cdot)$ denotes a lightweight enhancement encoder comprising a two-layer projection module, a learnable positional embedding, parallel multi-scale temporal convolutions with kernel sizes $\{3,5,7\}$, residual normalization, a two-layer Transformer encoder, and adaptive average pooling for sequence-length alignment. The same encoder is applied to both complete and incomplete modalities.

3.3.2 Bidirectional Attention Fusion

In multimodal sentiment analysis, textual features provide stable semantic information, while audio and visual modalities offer complementary affective cues. However, when a modality is missing or degraded, direct fusion may introduce noise and compromise semantic consistency.

To address this issue, we introduce Bidirectional Attention Fusion (BAF) as the core interaction stage of the TCBI module, as illustrated in Fig. 2. BAF performs hierarchical text-guided bidirectional attention

between the textual stream and the auxiliary audio-visual stream. Specifically, let $H_L'^i$, $H_A'^i$, and $H_V'^i$ denote the textual, audio, and visual representations at the i-th layer, respectively. Each representation satisfies $H_m'^i \in \mathbb{R}^{T_m \times d}$, where $m \in \{L, A, V\}$, T_m denotes the sequence length of modality m, and d is the feature dimension.

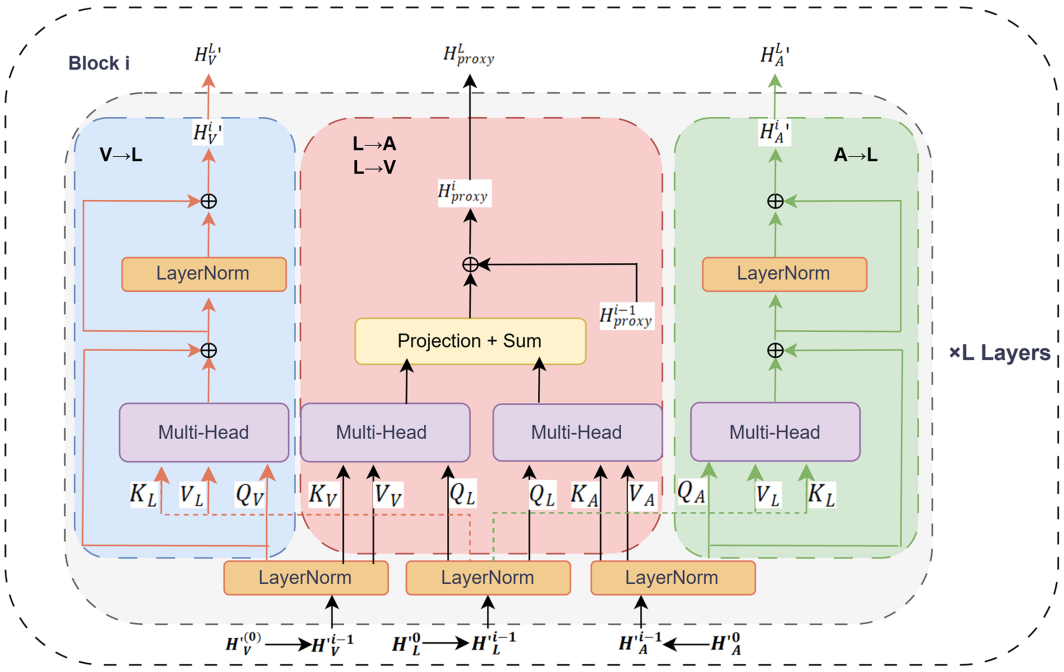

Figure 2: Text-centric bidirectional interaction module with proxy-based aggregation.

At layer i, the model takes as input the representations from the previous layer:

$$H_L'^{i-1}, \quad H_A'^{i-1}, \quad H_V'^{i-1}, \tag{3}$$

which serve as inputs to the subsequent cross-modal attention. The interaction module consists of three hierarchical layers.

Text-guided interaction. The textual modality serves as a semantic anchor to guide information extraction from auxiliary modalities. Specifically, we use the textual modality as the query and the auxiliary modality as the key and value, enabling the text representation to incorporate complementary cues from auxiliary modalities. Let $H_{L\leftarrow m}'^i$ denote the cross-modal context extracted from modality m via text-guided attention, serving as input to the proxy aggregation. It is computed as:

$$H_{L\leftarrow m}'^i = \text{MultiHead}(H_L'^{i-1}, H_m'^{i-1}, H_m'^{i-1}), \quad m \in \{V, A\}, \tag{4}$$

where $\text{MultiHead}(\cdot)$ denotes the standard multi-head attention mechanism with 4 attention heads and a head dimension of 32.

Auxiliary-guided enhancement. To enable bidirectional interaction, we further allow auxiliary modalities to be updated under textual guidance. Specifically, we use the auxiliary modality as the query and the textual modality as the key and value, enabling auxiliary features to absorb high-level semantic information from text. Let $H_{m\leftarrow L}'^i$ denote the auxiliary representation enhanced by textual guidance. Let W_p denote a learnable linear projection. It is defined as:

$$H_{m\leftarrow L}'^i = H_m'^{i-1} + W_m \,\text{MultiHead}(H_m'^{i-1}, H_L'^{i-1}, H_L'^{i-1}), \quad m \in \{V, A\}. \tag{5}$$

Proxy-based aggregation. To stabilize cross-modal interactions and suppress noise propagation, we introduce a learnable proxy representation. Let $H_{\text{proxy}}^{\prime i} \in \mathbb{R}^{T_p \times d}$ denote the proxy representation at layer i, with the proxy sequence length fixed at $T_p = 8$ and $d = 128$. Therefore, $H_{\text{proxy}}^{\prime i} \in \mathbb{R}^{8\times 128}$.

The proxy aggregates cross-modal information as:

$$H_{\text{proxy}}^{\prime i} = H_{\text{proxy}}^{\prime i-1} + W_p\Big(\sum_{m\in\{V,A\}} H_{L\leftarrow m}^{\prime i}\Big), \tag{6}$$

where W_p denotes a learnable linear projection for feature alignment.

Fusion with proxy guidance. Let $\tilde{H}_L^{\prime i}$ denote the final refined textual representation after incorporating proxy-guided information. It is computed as:

$$\tilde{H}_L^{\prime i} = H_L^{\prime i} + \text{CrossAttn}(H_L^{\prime i}, H_{\text{proxy}}^{\prime i}, H_{\text{proxy}}^{\prime i}), \tag{7}$$

where $\text{CrossAttn}(\cdot)$ denotes cross-attention with the proxy as the key and value.

The final refined textual representation is obtained via a feed-forward network:

$$\tilde{H}_L^{\prime i} = \hat{H}_L^{\prime i} + \text{FFN}(\hat{H}_L^{\prime i}) \tag{8}$$

where $\text{FFN}(\cdot)$ denotes a two-layer feed-forward network with non-linear activation.

Finally, let h denote the aggregated sentence-level representation obtained by mean pooling:

$$h = \text{MeanPooling}(\tilde{H}_L^{\prime L}), \tag{9}$$

where L is the total number of layers.

Let $\hat{y}$ denote the predicted sentiment distribution. It is computed as:

$$\hat{y} = f_{\text{reg}}(h), \tag{10}$$

where $f_{\text{reg}}(\cdot)$ denotes the regression head, consisting of a two-layer MLP with GELU activation, followed by a linear projection to a scalar output.

3.4 Semantic Compensation Distillation

To compensate for semantic information loss caused by modality incompleteness and noise, we adopt a teacher–student distillation framework in which the teacher processes complete inputs and the student operates on incomplete data. The teacher network is frozen with respect to gradient updates and provides supervision at both the representation and prediction levels.

We employ a hierarchical distillation strategy with three levels: intra-modal, inter-modal, and prediction-level alignment.

Intra-modal distillation. Let H_m^s and H_m^t denote the modality-specific representations of the student and teacher networks for modality $m \in \{L, A, V\}$. To align these representations, we minimize the discrepancy between their pooled features:

$$\mathcal{L}_{\text{intra}} = \frac{1}{|\mathcal{M}|} \sum_{m\in\{L,A,V\}} \text{SmoothL1}(u_m^s, u_m^t), \tag{11}$$

where $u_m = \text{MeanPooling}(H_m)$ denotes the modality-level representation obtained via temporal mean pooling, and $\text{SmoothL1}(\cdot)$ is the smooth L1 loss.

Inter-modal distillation. Let H^s_{fused} and H^t_{fused} denote the fused multimodal representations of the student and teacher networks, respectively. To align cross-modal representations, we introduce a projection head $g(\cdot)$:

$$\mathcal{L}_{\text{inter}} = (1-\alpha)\,\text{SmoothL1}(g(H^s_{\text{fused}}), g(H^t_{\text{fused}})) + \alpha\left(1 - \text{CosSim}(g(H^s_{\text{fused}}), g(H^t_{\text{fused}}))\right), \tag{12}$$

where $g(\cdot)$ denotes a two-layer MLP with Layer Normalization, $\text{CosSim}(\cdot)$ represents cosine similarity, and $\alpha \in [0,1]$ is a balancing coefficient.

Prediction-level distillation. Let $\hat{y}_s$ and $\hat{y}_t$ denote the predicted sentiment intensities for the student and teacher networks, respectively. The final prediction alignment is defined as:

$$\mathcal{L}_{\text{pred}} = \text{SmoothL1}(\hat{y}_s, \hat{y}_t). \tag{13}$$

The overall distillation enforces consistency across modality-specific, fused, and prediction spaces, thereby improving robustness to modality uncertainty.

3.5 Overall Learning Objectives

The overall training objective integrates sentiment prediction and knowledge distillation. The fused multimodal representation H^s_{fused} from the student network is fed into a multi-layer perceptron (MLP) to produce the final prediction:

$$\hat{y}_s = f_{\text{reg}}(h_s), \tag{14}$$

where $\hat{y}_s = f_{\text{reg}}(h_s)$ as defined in Eq. (10), with $h_s = \text{MeanPool}(\tilde{H}'^L_L)$.

The regression loss for sentiment intensity prediction is defined as:

$$\mathcal{L}_{\text{reg}} = \text{SmoothL1}(\hat{y}_s, y), \tag{15}$$

where y denotes the ground-truth sentiment label.

The overall distillation loss is:

$$\mathcal{L}_{\text{distill}} = \lambda_{\text{intra}}\mathcal{L}_{\text{intra}} + \lambda_{\text{inter}}\mathcal{L}_{\text{inter}} + \lambda_{\text{pred}}\mathcal{L}_{\text{pred}}. \tag{16}$$

Finally, the total loss function is defined as:

$$\mathcal{L} = \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{distill}}, \tag{17}$$

where λ_{intra}, λ_{inter}, λ_{pred}, and λ_{reg} are hyperparameters that weight the components.

4 Experiments

4.1 Datasets

We conduct experiments on three widely used multimodal sentiment analysis benchmarks: MOSI, MOSEI, and SIMS. As in prior work, sentiment prediction is formulated as a regression task, and both regression and derived classification metrics are reported. Dataset statistics and splits are summarized in Table 1.

Table 1: Summary of evaluated datasets.

Dataset	Language	Total	Train	Valid	Test	Range	Notes
MOSI	English	2199	1284	229	686	[−3, 3]	Benchmark
MOSEI	English	22,856	16,326	1871	4659	[−3, 3]	Large-scale
SIMS	Chinese	2281	1368	456	457	[−1, 1]	Uni-modal labels

4.2 Implementation Details

For fair comparison, we follow the missing-modality protocol of TF-Mamba (Li et al., 2025 [10]). During training, modalities are randomly dropped according to a probabilistic scheme. During evaluation, the missing rate r is varied from 0 to 0.9 with a step size of 0.1, and the final results are averaged over all tested missing rates. We exclude $r = 1.0$ because all modalities would be removed, making prediction undefined. To provide a comprehensive measure of model robustness, the reported results are averaged across all tested missing rates.

All experiments are implemented in Python 3.11.7 with PyTorch 2.2.1. We use standard regression and classification metrics, including Mean Absolute Error (MAE), Pearson Correlation (Corr), binary accuracy (Acc-2), multi-class accuracy (Acc-7 for MOSI/MOSEI and Acc-5/Acc-3 for SIMS), and weighted F1-score (F1).

For MOSI and MOSEI, Acc-2 and F1 are reported under two common settings: Non0, which excludes samples with a sentiment score of 0, and Has0, which retains zero-score samples and treats them as non-negative. Thus, the values before and after "/" correspond to negative vs. positive and negative vs. non-negative settings, respectively. Predictions are clipped to the valid label ranges when computing MAE and Corr. Detailed hyper-parameter settings are listed in Table 2.

Table 2: Hyper-parameter settings.

Hyper-Parameter	MOSI	MOSEI	SIMS
Text dim (d_t)	768	768	768
Audio dim (d_a)	5	74	25
Visual dim (d_v)	20	35	177
Batch size	32	32	32
Epochs	150	150	150
Optimizer	Adam	Adam	Adam
Aligned sequence length (T)	8	8	8
Learning rate	1×10^{-5}	1×10^{-5}	1×10^{-5}
Hidden dimension (d)	128	128	128

4.3 Baselines

To evaluate the robustness of our method, we conduct fair comparisons with a series of state-of-the-art (SOTA) approaches under consistent modality-missing settings, including methods specifically designed for missing-modality scenarios, such as TFR-Net (Yuan et al., 2021 [28]), MCTN (Pham et al., 2019 [29]), TransM (Wang et al., 2020 [30]), SMIL (Ma et al., 2021 [31]), and CorrKD (Li et al., 2024 [23]). The

baselines also include representative multimodal methods, such as MISA (Hazarika et al., 2020 [32]), Self-MM (Yu et al., 2021 [33]), MMIM (Han et al., 2021 [11]), CENET (Wang et al., 2022 [34]), CubeMLP (Sun et al., 2022 [35]), as well as recent sequence modeling approaches, including BI-Mamba (Yang et al., 2024 [36]) and TF-Mamba (Li et al., 2025 [10]).

4.4 Experimental Results

We evaluate the performance of TC-DSC under both intra-modal and inter-modal missing conditions. Specifically, the intra-modal setting models partial information loss within individual modalities, while the inter-modal setting simulates the absence of entire modalities. We compare TC-DSC with existing methods to assess its performance across different missing scenarios.

Comparison for Intra-Modal Missingness: The experimental results on the MOSI and MOSEI datasets are summarized in Table 3, while the results of SIMS are reported in Table 4. The best results are in bold, and the second-best are underlined. Overall, the proposed method achieves competitive performance across both datasets, demonstrating its effectiveness in fine-grained sentiment analysis.

Table 3: Performance comparison under intra-modal missing settings on MOSI and MOSEI datasets.

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
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
TF-Mamba	30.18	33.90	71.57/**72.15**	71.59/**72.26**	1.094	**0.528**	45.66	46.64	77.34/77.61	77.18/77.43	0.673	0.578
TC-DSC	**33.95**	**37.64**	**71.77**/71.09	71.71/71.06	1.075	0.512	**46.95**	**48.08**	**78.47/78.27**	**77.90/77.56**	**0.664**	**0.593**

Table 4: Performance comparison under intra-modal missing settings the SIMS dataset.

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
BI-Mamba	31.90	54.95	70.79	69.26	0.529	0.345
TF-Mamba	34.46	55.51	**74.68**	**72.20**	0.512	**0.386**
TC-DSC	**35.51**	**57.26**	71.58	71.63	0.516	**0.386**

On MOSI, our method achieves 33.95% and 37.64% on Acc-7 and Acc-5, outperforming TF-Mamba by 3.77% and 3.74%, respectively. On MOSEI, it obtains the best results on most metrics, including Acc-2, F1, MAE, and Corr, demonstrating strong robustness and generalization on large-scale multimodal data. As shown in Table 4, our method also achieves competitive fine-grained classification performance on SIMS, with 35.51% on Acc-5 and 57.26% on Acc-3, surpassing TF-Mamba by 1.05% and 1.75%. These results confirm the effectiveness of the proposed hierarchical text-guided dual-stream interaction mechanism.

However, our method does not achieve the best performance on coarse-grained classification metrics, F1, and MAE on SIMS, where it is slightly inferior to some state-space modeling-based methods. This may be partly due to the limited scale of SIMS, which provides fewer training samples for learning stable cross-modal interactions. Since TC-DSC relies on hierarchical cross-modal interaction and semantic compensation distillation, its advantages may be less pronounced under data-scarce conditions. Nevertheless, the results show that TC-DSC is particularly effective in fine-grained sentiment modeling. Its more consistent improvements on the larger MOSEI dataset further suggest that the proposed framework is more suitable for medium- to large-scale multimodal sentiment analysis scenarios with sufficient data diversity.

Fig. 3 illustrates the performance trends of different methods across three datasets under varying modality missing rates. As the missing rate increases, the performance of all models generally declines, and regression errors gradually increase, highlighting the impact of modality incompleteness. On MOSI and MOSEI, TC-DSC maintains relatively competitive and stable performance across most missing rates, with a slower degradation trend, particularly under moderate missing conditions, indicating its ability to handle modality degradation effectively. On the SIMS dataset, due to its limited scale, all methods exhibit noticeable fluctuations under high missing rates; although TC-DSC does not consistently achieve the best results on all metrics, it maintains relatively stable trends and competitive performance under low to moderate missing rates.

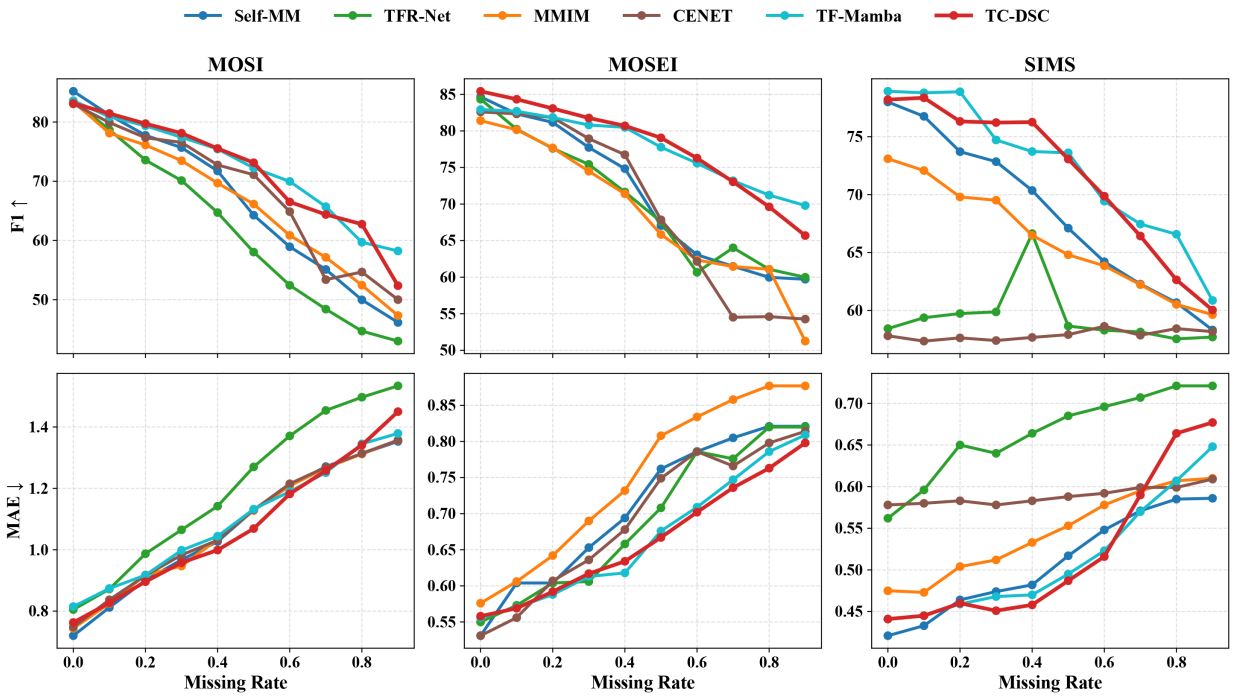

Figure 3: Performance comparison under different missing rates on MOSI, MOSEI, and SIMS datasets.

To further examine the statistical reliability of the reported improvements, we conducted a paired bootstrap significance test between TC-DSC and the strongest baseline TF-Mamba with 5000 resampling iterations on the official test sets. As summarized in Table 5, TC-DSC achieves statistically significant improvements on several key metrics. Specifically, significant gains are observed on classification metrics on SIMS and on MAE, Corr, Acc-2, and F1-related metrics on MOSEI. On MOSI, TC-DSC shows positive but non-significant improvements on most metrics. These results indicate that the performance gains of TC-DSC are statistically reliable on several key metrics rather than being solely caused by random fluctuations.

Table 5: Summary of paired bootstrap significance test results.

Dataset	Significant Results	Non-Significant Results
SIMS	4 metrics, p = 0.0028–0.0372	MAE, Corr
MOSI	–	Most metrics positive but $p > 0.05$
MOSEI	6 metrics, $p < 0.001$–0.0344	Acc-5, Acc-7

Comparison for Inter-Modal Missingness: To comprehensively evaluate the proposed model's cross-modal compensation capability under modality-missing scenarios, we construct multiple modality-availability settings on the MOSI and MOSEI datasets, including {t}, {t, a}, {t, v}, and {t, a, v}. This design follows a text-centric paradigm, with the textual semantic space serving as the anchor for cross-modal alignment, enabling us to analyze how auxiliary modalities (audio and visual) contribute to sentiment understanding and compensate for missing information.

We evaluate model performance across inter-modal missingness scenarios using the F1 score. To assess performance stability across modality combinations, we further report the average F1 score (denoted as Avg.).

As shown in Table 6, as the available modalities expand from text-only to bimodal and trimodal settings, all methods show improved overall performance, confirming the complementary value of multimodal information. The proposed TC-DSC maintains stable performance across all modality combinations and achieves superior average accuracy, reaching 83.24 on MOSI and 84.87 on MOSEI, outperforming CorrKD and TF-Mamba. Notably, in partial modality settings such as {t, a} and {t, v}, TC-DSC consistently delivers competitive results, demonstrating strong cross-modal compensation. This advantage is more pronounced on the larger MOSEI dataset, where TC-DSC further improves performance across all modality settings, highlighting its effectiveness in integrating cross-modal information and maintaining robust sentiment prediction under complex missing-modality conditions.

Table 6: Inter-modal missing comparison on MOSI and MOSEI datasets. The best results are in bold.

Model	MOSI					MOSEI				
	{t}	{t, a}	{t, v}	{t, a, v}	Avg.	{t}	{t, a}	{t, v}	{t, a, v}	Avg.
CubeMLP	64.15	63.76	65.12	**84.57**	69.40	67.52	71.69	70.06	83.17	73.11
MCTN	75.21	77.81	74.82	80.12	76.99	75.50	76.64	77.13	81.75	77.76
TransM	77.64	82.07	80.90	82.57	80.80	77.98	80.46	78.61	81.48	79.63
SMIL	78.26	79.82	79.15	82.85	80.02	76.57	77.68	76.24	80.74	77.81
CorrKD	81.20	83.56	82.41	83.94	82.78	80.76	81.74	81.28	82.16	81.49
TF-Mamba	**82.07**	82.42	83.88	83.71	83.02	78.87	83.15	82.02	83.71	80.37
TC-DSC	81.40	**83.72**	**83.56**	83.90	**83.15**	**82.30**	**85.31**	**85.87**	**86.01**	**84.87**

4.5 Ablation Study

To evaluate the contribution of each core component, we conduct ablation studies on the MOSI dataset, as shown in Table 7. The best results are in bold.

Table 7: Ablation study results under different model settings on the MOSI dataset.

Method	MOSI					
	Acc-7	Acc-5	Acc-2	F1	MAE↓	Corr↑
w/o SCD	32.19	36.92	**72.03**/70.64	**71.86**/70.56	1.087	0.493
w/o BAF	15.46	15.46	55.56/44.27	55.62/44.17	1.350	0.399
w/o UFEE	33.24	37.62	71.80/70.29	71.80/70.26	1.100	0.507
w/o SCD & BAF	15.64	15.52	69.21/66.52	65.48/62.32	1.3146	0.46
w/o SCD & UFEE	33.54	36.82	71.10/70.07	71.11/69.99	1.0862	0.5058
Full Model	**33.95**	**37.64**	71.77/**71.09**	71.71/**71.06**	**1.075**	**0.512**

Removing BAF causes the most pronounced performance degradation, indicating its central role in cross-modal interaction. Without BAF, multimodal cues cannot be effectively integrated into the textual semantic space, causing the model to rely more heavily on unimodal text representations. Removing SCD also leads to clear performance drops, showing that semantic compensation distillation helps maintain robust and consistent high-level representations under incomplete or noisy inputs. Replacing UFEE results in a smaller but consistent decline, suggesting that multi-scale enhancement improves audio-visual feature quality for subsequent cross-modal fusion.

To further analyze component interactions, we conduct combined ablation experiments by removing multiple modules simultaneously. As shown in Table 7, removing both SCD and BAF causes a severe drop, with Acc-7 decreasing from 33.95 to 15.64 and MAE increasing from 1.075 to 1.3146, indicating that semantic compensation and bidirectional attention fusion jointly support effective feature alignment. In contrast, removing both SCD and UFEE leads to milder degradation, with Acc-7 decreasing to 33.54 and MAE increasing to 1.0862, suggesting that UFEE mainly provides complementary feature enhancement while SCD contributes more to semantic consistency.

To provide mechanism-level interpretability for TC-DSC, we conduct visualization analyses on the MOSI dataset from the perspectives of feature distribution and ablation-based contribution. As shown in Fig. 4, the raw incomplete multimodal features are scattered and highly overlapped, indicating obvious semantic ambiguity. After introducing UFEE, the feature distributions become more compact, suggesting that multi-scale enhancement improves the quality of audio-visual representations. After introducing BAF, the separation between sentiment categories is further improved, showing that BAF facilitates alignment with the textual semantic space. With the full TC-DSC model, the representations become the most compact and separable, demonstrating that UFEE and BAF progressively enhance representation discrimination under incomplete modality conditions.

In addition, since SCD is a training-stage distillation objective, its contribution is evaluated by removing the SCD loss during training. As shown in Fig. 5, the removal of SCD, BAF, or UFEE leads to clear performance degradation, indicating that these components are effective for semantic compensation, cross-modal interaction, and robust representation learning under incomplete modality conditions.

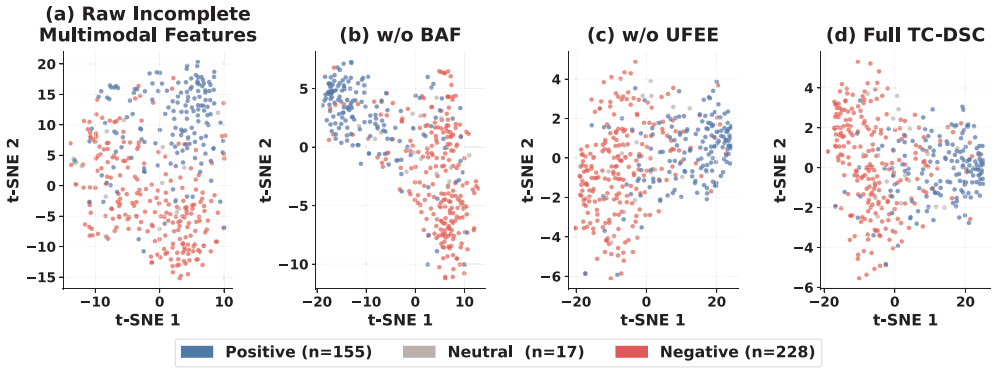

Figure 4: t-SNE visualization of feature distributions on MOSI.

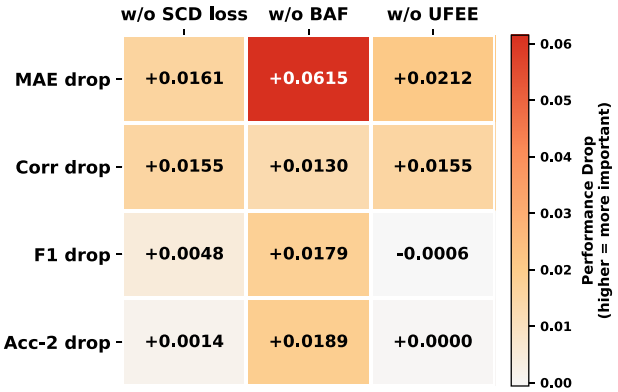

Figure 5: Ablation-based contribution heatmap of TC-DSC on MOSI.

5 Conclusion

In this paper, we propose TC-DSC, a text-centric hierarchical dual-stream interaction framework for incomplete multimodal sentiment analysis. TC-DSC avoids explicit generative reconstruction and instead performs feature-space interaction between textual and audio-visual semantic streams to improve cross-modal alignment and robustness. It integrates multi-scale feature enhancement, hierarchical proxy-based bidirectional interaction, and semantic compensation distillation to enhance representation quality and maintain semantic consistency under modality missing. Experiments on benchmark datasets demonstrate that TC-DSC achieves robust performance under both complete and incomplete settings, and ablation studies validate the effectiveness of its key components.

6 Limitations and Future Work

Despite consistent improvements across multiple datasets, the proposed method still has several limitations. First, its performance is sensitive to data scale, with reduced advantages on small-scale datasets due to insufficient training samples for stable cross-modal alignment. Second, the framework relies on text as the primary semantic anchor, and its robustness may be limited when textual inputs are noisy, ambiguous, or misleading, which may weaken the contributions of the audio and visual modalities. Third, this work focuses on static missing-modality settings, whereas real-world scenarios often involve dynamic and evolving missing patterns. Fourth, the feature-space alignment and proxy-based interaction are mainly task-oriented and do not impose explicit constraints to guarantee modality-invariant or disentangled representations.

Additionally, the method's computational cost increases quadratically with the feature dimension D or sequence length T, which may affect efficiency in resource-constrained scenarios.

In future work, we will explore data-efficient learning methods to improve performance with limited training data. We will also investigate adaptive modality reliability estimation mechanisms to reduce the negative impact of unreliable textual inputs and develop dynamic missing-modality modeling strategies for more realistic, evolving incomplete multimodal scenarios. We will explore iterative refinement mechanisms and feedback-based cross-modal correction strategies to enable tighter, more tightly coupled optimization across modules. We will also consider explicit representation constraints, such as contrastive learning, adversarial alignment, or orthogonality-based regularization, to better encourage modality-invariant and disentangled multimodal representations.

Acknowledgement: None.

Funding Statement: This research was funded by the Key Research Project for Higher Education Institutions of Henan Province, grant number 25A520023.

Author Contributions: Conceptualization, Changchang Fan, Jinjin Liu, Qiulu Guo, Yihao Xu and Penghui Ma; methodology, Changchang Fan and Jinjin Liu; software, Changchang Fan and Jinjin Liu; validation, Changchang Fan; formal analysis, Changchang Fan, Jinjin Liu, Qiulu Guo, Yihao Xu and Penghui Ma; investigation, Changchang Fan; resources, Jinjin Liu; writing—original draft preparation, Changchang Fan; writing—review and editing, Changchang Fan and Jinjin Liu. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets used in this study are publicly available at: https://github.com/thuiar/MMSA. Further inquiries can be directed to the corresponding author.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sun L, Lian Z, Liu B, Tao J. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. IEEE Trans Affective Comput. 2024;15(1):309–25. doi:10.1109/taffc.2023.3274829.
2. Sun Y, Liu Z, Sheng QZ, Chu D, Yu J, Sun H. Similar modality completion-based multimodal sentiment analysis under uncertain missing modalities. Inf Fusion. 2024;110(3):102454. doi:10.1016/j.inffus.2024.102454.
3. Liu Z, Zhou B, Chu D, Sun Y, Meng L. Modality translation-based multimodal sentiment analysis under uncertain missing modalities. Inf Fusion. 2024;101(3):101973. doi:10.1016/j.inffus.2023.101973.
4. Shi P, Hu M, Nakagawa S, Zheng X, Shi X, Ren F. Text-guided reconstruction network for sentiment analysis with uncertain missing modalities. IEEE Trans Affective Comput. 2025;16(3):1825–38. doi:10.1109/taffc.2025.3541743.
5. Tu G, Wu T, Luo X, Zeng X, Li W, Xu R. Meta-learning for incomplete multimodal sentiment analysis. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2025 Jul 13–18; Padua, Italy. p. 2911–5.
6. Lin R, Hu H. MissModal: increasing robustness to missing modality in Multimodal Sentiment analysis. Trans Assoc Comput Linguist. 2023;11(2):1686–702. doi:10.1162/tacl_a_00628.
7. Li M, Yang D, Lei Y, Wang S, Wang S, Su L, et al. A unified self-distillation framework for multimodal sentiment analysis with uncertain missing modalities. Proc AAAI Conf Artif Intell. 2024;38(9):10074–82. doi:10.1609/aaai.v38i9.28871.
8. Li M, Yang D, Liu Y, Wang S, Chen J, Wang S, et al. Toward robust incomplete multimodal sentiment analysis via hierarchical representation learning. Adv Neural Inf Process Syst. 2024;37:28515–36. doi:10.52202/079017-0895.

9. Yang Z, He Q, Yu M, Du N, Lu Y. TCTR: text-guided contrastive learning with token-level reconstruction network for missing modalities in multimodal sentiment analysis. Inf Fusion. 2026;126(10):103571. doi:10.1016/j.inffus.2025.103571.
10. Li X, Cheng X, Miao D, Zhang X, Li Z. TF-Mamba: text-enhanced fusion Mamba with missing modalities for robust Multimodal Sentiment analysis. arXiv:2505.14329. 2025.
11. Han W, Chen H, Poria S. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. arXiv:2109.00412. 2021.
12. Yu Y, Zhao M, Qi S, Sun F, Wang B, Guo W, et al. ConKI: contrastive knowledge injection for multimodal sentiment analysis. arXiv:2306.15796. 2023.
13. Qian F, Han J, He Y, Zheng T, Zheng G. Sentiment knowledge enhanced self-supervised learning for multimodal sentiment analysis. In: Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023; 2023 Jul 9–14; Toronto, ON, Canada. p. 12966–78.
14. Wang D, Guo X, Tian Y, Liu J, He L, Luo X. TETFN: a text enhanced transformer fusion network for multimodal sentiment analysis. Pattern Recognit. 2023;136(2):109259. doi:10.1016/j.patcog.2022.109259.
15. Lin R, Hu H. Multi-task momentum distillation for multimodal sentiment analysis. IEEE Trans Affective Comput. 2024;15(2):549–65. doi:10.1109/taffc.2023.3282410.
16. Feng X, Lin Y, He L, Li Y, Chang L, Zhou Y. Knowledge-guided dynamic modality attention fusion framework for multimodal sentiment analysis. arXiv:2410.04491. 2024.
17. Wang L, Peng J, Zheng C, Zhao T, Zhu LA. A cross modal hierarchical fusion multimodal sentiment analysis method based on multi-task learning. Inf Process Manag. 2024;61(3):103675. doi:10.1016/j.ipm.2024.103675.
18. Wang H, Ren C, Yu Z. Multimodal sentiment analysis based on multiple attention. Eng Appl Artif Intell. 2025;140(2):109731. doi:10.1016/j.engappai.2024.109731.
19. Xu D, Zhang C. Exploring feature interactions for multimodal sentiment analysis. Int J Cogn Inform Nat Intell. 2025;19(1):1–19. doi:10.4018/ijcini.383754.
20. Xiao L, Wu X, Wu W, Yang J, He L. Multi-channel attentive graph convolutional network with sentiment fusion for multimodal sentiment analysis. In: Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2022 May 23–27; Singapore, Singapore. New York, NY, USA: IEEE; 2022. p. 4578–82.
21. Zeng J, Liu T, Zhou J. Tag-assisted multimodal sentiment analysis under uncertain missing modalities. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2022 Jul 11–15; Madrid, Spain. p. 1545–54.
22. Zeng J, Zhou J, Liu T. Mitigating Inconsistencies in multimodal sentiment analysis under uncertain missing modalities. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing; 2022 Dec 7–11; Abu Dhabi, United Arab Emirates. p. 2924–34.
23. Li M, Yang D, Zhao X, Wang S, Wang Y, Yang K, et al. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA. p. 12458–68.
24. Xi Z, Yu B, Zhao S, Yang S. Bridging the gaps: a hierarchical distillation approach to multimodal sentiment analysis with missing modalities. Inf Fusion. 2026;127(3):103838. doi:10.1016/j.inffus.2025.103838.
25. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2019 Jun 2–7; Minneapolis, MN, USA. Minneapolis, MN, USA: Association for Computational Linguistics; 2019. p. 4171–86.
26. Baltrusaitis T, Zadeh A, Lim YC, Morency LP. OpenFace 2.0: facial behavior analysis toolkit. In: Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018); 2018 May 15–19; Xi'an, China. p. 59–66.
27. McFee B, Raffel C, Liang D, Ellis D, McVicar M, Battenberg E, et al. librosa: audio and music signal analysis in python. In: Proceedings of the 14th Python in Science Conference; 2015 July 6–12; Austin, TX, USA. p. 18–24.

28. Yuan Z, Li W, Xu H, Yu W. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In: Proceedings of the 29th ACM International Conference on Multimedia; 2021 Oct 20–24; Chengdu, China. p. 4400–7.
29. Pham H, Liang PP, Manzini T, Morency LP, Póczos B. Found in translation: learning robust joint representations by Cyclic translations between modalities. Proc AAAI Conf Artif Intell. 2019;33(01):6892–9. doi:10.1609/aaai.v33i01.33016892.
30. Wang Z, Wan Z, Wan X. TransModality: an End2End fusion method with transformer for multimodal sentiment analysis. In: Proceedings of the Web Conference 2020; 2020 Apr 20–24; Taipei, Taiwan. p. 2514–20.
31. Ma M, Ren J, Zhao L, Tulyakov S, Wu C, Peng X. SMIL: multimodal learning with severely missing modality. Proc AAAI Conf Artif Intell. 2021;35(3):2302–10. doi:10.1609/aaai.v35i3.16330.
32. Hazarika D, Zimmermann R, Poria S. Misa: modality-invariant andspecific representations for multimodal sentiment analysis. In: Proceedings of the 28th ACM International Conference on Multimedia; 2020 Oct 12–16; Seattle, WA, USA. p. 1122–31.
33. Yu W, Xu H, Yuan Z, Wu J. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. Proc AAAI Conf Artif Intell. 2021;35(12):10790–7. doi:10.1609/aaai.v35i12.17289.
34. Wang D, Liu S, Wang Q, Tian Y, He L, Gao X. Cross-modal enhancement network for multimodal sentiment analysis. IEEE Trans Multim. 2023;25:4909–21. doi:10.1109/tmm.2022.3183830.
35. Sun H, Wang H, Liu J, Chen YW, Lin L. CubeMLP: an MLP-based model for multimodal sentiment analysis and depression estimation. In: Proceedings of the 30th ACM International Conference on Multimedia; 2022 Oct 10–14; Lisboa, Portugal. p. 3722–9.
36. Yang Z, Zhang J, Wang G, Kalra MK, Yan P. Cardiovascular disease detection from multi-view chest X-rays with bimamba. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention; 2024 Oct 6–10; Marrakesh, Morocco. Berlin/Heidelberg, Germany: Springer; 2024. p. 134–44.