ADRMSA: Adaptive Disentanglement for Robust Multimodal Sentiment Analysis

Anonymous ACL submission

Abstract

Random Missing Modality Features (RMFM) is a critical challenge in multimodal sentiment analysis. Existing methods tackling RMFM typically fuse corrupted representations directly, which inevitably causes representation pollution during cross-modal interaction. To address this, we propose ADRMSA, a unified framework designed for robust MSA under RMFM. Specifically, an Adaptive Disentanglement (AD) module decomposes each modality into clean and polluted representations, while a Pollution-Adaptive Gate (PAG) computes the Residual Semantic Reliability Coefficient γ to estimate their dependency. Based on this, an Adaptive Noise-Aware Fusion (ANAF) module progressively extracts reliable residual semantics from polluted representations under the guidance of γ. Moreover, a Clean Semantic Alignment (CSA) module aligns the refined incomplete representation with complete semantic representations to mitigate semantic drift. Extensive experiments on MOSI, MOSEI, and CH-SIMS show that ADRMSA achieves superior performance across various missing ratios, demonstrating its effectiveness and robustness in noisy environments.

1 Introduction

Multimodal Sentiment Analysis (MSA) (Zadeh et al., 2017a; Gandhi et al., 2023) aims to identify and understand human emotions more comprehensively and accurately by integrating multimodal information such as text, speech, and vision (Li et al., 2024). However, due to factors such as sensor failure, hardware limitations, or environmental interference, modality features often suffer from random loss in the time dimension, namely Random Modality Feature Missing (RMFM) (Zhong et al., 2025). Unlike modality-level overall missing, RMFM refers to discontinuous random missing within a modality, with missing patterns varying across samples (Tsai et al., 2019; Nguyen et al., 2025)., more details can be found in Appendix A. Recently, constructing a robust multimodal sentiment analysis system capable of addressing the RMFM problem has become a research hotspot in MSA (Lin and Hu, 2023; Zhang et al., 2025; Gong et al., 2026).

Although existing methods for RMFM differ in architectural design, they generally follow a similar paradigm: they first model inter-modal dependencies through cross-modal interaction mechanisms, and then compensate for missing information through feature reconstruction or robust representation learning (Zhao et al., 2025; Wu et al., 2026). For example, Yuan et al. (2021) establishes semantic correlations across modalities via cross-modal attention and introduces generative supervision to explicitly reconstruct missing features. Sun et al. (2025) incorporates convolutional layers into the cross-modal Transformer to capture local temporal dependencies.

However, these methods share a common limitation. During the core cross-modal interaction stage, they typically project modality representations with varying levels of completeness into a shared representation space and directly perform fusion. Since the noise induced by missing features has not been explicitly separated beforehand, it is injected into the shared semantic space during cross-modal interaction, leading to **representation pollution** (Chaudhuri et al., 2025; Su et al., 2025). As a result, subsequent reconstruction or representation learning modules can only operate on already polluted representations, making it difficult to preserve representation purity and stable cross-modal semantic alignment (Zhang, 2025; Chaudhuri et al., 2025). Consequently, these approaches tend to encounter performance bottlenecks under RMFM scenarios (Weng et al., 2026).

To address the representation pollution problem, we propose a robust multimodal sentiment analysis framework named ADRMSA, which consists

of three modules: (1) the **Adaptive Disentanglement (AD)** module performs initial separation of each modality into clean and polluted representations. (2) Based on the separated representations, the **Adaptive Noise-Aware Fusion (ANAF)** module progressively extracts residual sentiment semantics from polluted representations while suppressing noise propagation. (3) the **Clean Semantic Alignment (CSA)** module aligns the refined clean representations with complete semantic representations[1] to correct semantic drift accumulated during the process of ANAF. Through the collaborative optimization of these three stages, ADRMSA systematically alleviates representation pollution and improves robust cross-modal semantic alignment under RMFM conditions. Moreover, ADRMSA requires no explicit missing pattern assumptions, enabling generalization across diverse random missing scenarios. Meanwhile, the progressive purification strategy preserves informative sentiment semantics while suppressing interference from corrupted modalities.

The contributions of this paper are as follows:

- We propose ADRMSA, an adaptive disentanglement-based framework for robust multimodal sentiment analysis, which addresses the representation pollution issue caused by cross-modal interaction over corrupted modality representations under RMFM.

- We propose an Adaptive Disentanglement (AD) module to decompose each modality into initial clean and polluted representations, and we introduce an Adaptive Noise-Aware Fusion (ANAF) module and a Clean Semantic Alignment (CSA) module. ANAF progressively extracts reliable residual semantics from polluted representations while suppressing noise propagation. CSA aligns the refined incomplete representation with complete representations to correct semantic drift.

- Extensive experiments on CMU-MOSI, CMU-MOSEI and SIMS show that ADRMSA achieves new state-of-the-art performance under varying missing levels. We give a detailed analysis about the experiments. The code/data will be released on GitHub.

[1]Here, "complete semantic representations" refer to the representations extracted from the original input without any missing features, serving as a stable semantic anchor.

2 Related Work

Research on handling missing modalities in multimodal sentiment analysis (MSA) has mainly followed two paradigms: (1) **generative reconstruction** methods aim to recover missing features from available modalities. Early approaches relied on Variational Autoencoders (VAEs) to model data distributions in latent spaces (Kingma and Welling, 2014). Generative Adversarial Networks (GANs) were later introduced to improve reconstruction quality. For example, Suo et al. (2019) used GANs to generate missing modalities and combined metric learning to preserve cross-modal consistency. TFR-Net(Yuan et al., 2021) incorporates Transformer-based feature reconstruction to enhance representation learning, while EMT-DLFR (Sun et al., 2022) introduces a dual-level feature restoration mechanism that combines low-level reconstruction with high-level semantic constraints. More recently, diffusion models have demonstrated strong potential for modality completion due to their powerful distribution modeling capability, particularly in non-contiguous missing scenarios (Wang et al., 2023; Kebaili et al., 2025). However, under high missing rates or random missing patterns, these methods often struggle to maintain semantic consistency and may introduce sentiment-irrelevant artifacts that interfere with downstream prediction. (2) **joint representation learning** methods avoid explicit generation and instead focus on learning representations that remain stable under missing modalities. MISA (Hazarika et al., 2020) disentangles modality-invariant and modality-specific features to improve fusion quality. Self-MM (Yu et al., 2021) strengthens unimodal representations through self-supervised multitask learning, LNLN (Zhang et al., 2024) treats language as the dominant modality to compensate for missing information from other modalities. More recently, P-RMF(Zhu et al., 2025) learns a proxy modality in latent space with uncertainty estimation. TF-Mamba(Li et al., 2025) employs an efficient Mamba-based state-space architecture for multimodal fusion under missing conditions, and TCTR(Yang et al., 2026) introduces supervised contrastive learning to jointly supplement information in both feature and semantic spaces. LDSD (Shi et al., 2026) strengthens text features via auxiliary modalities and uses teacher–student learning with EMA updates to reconstruct missing data. Despite their success, most existing methods over-

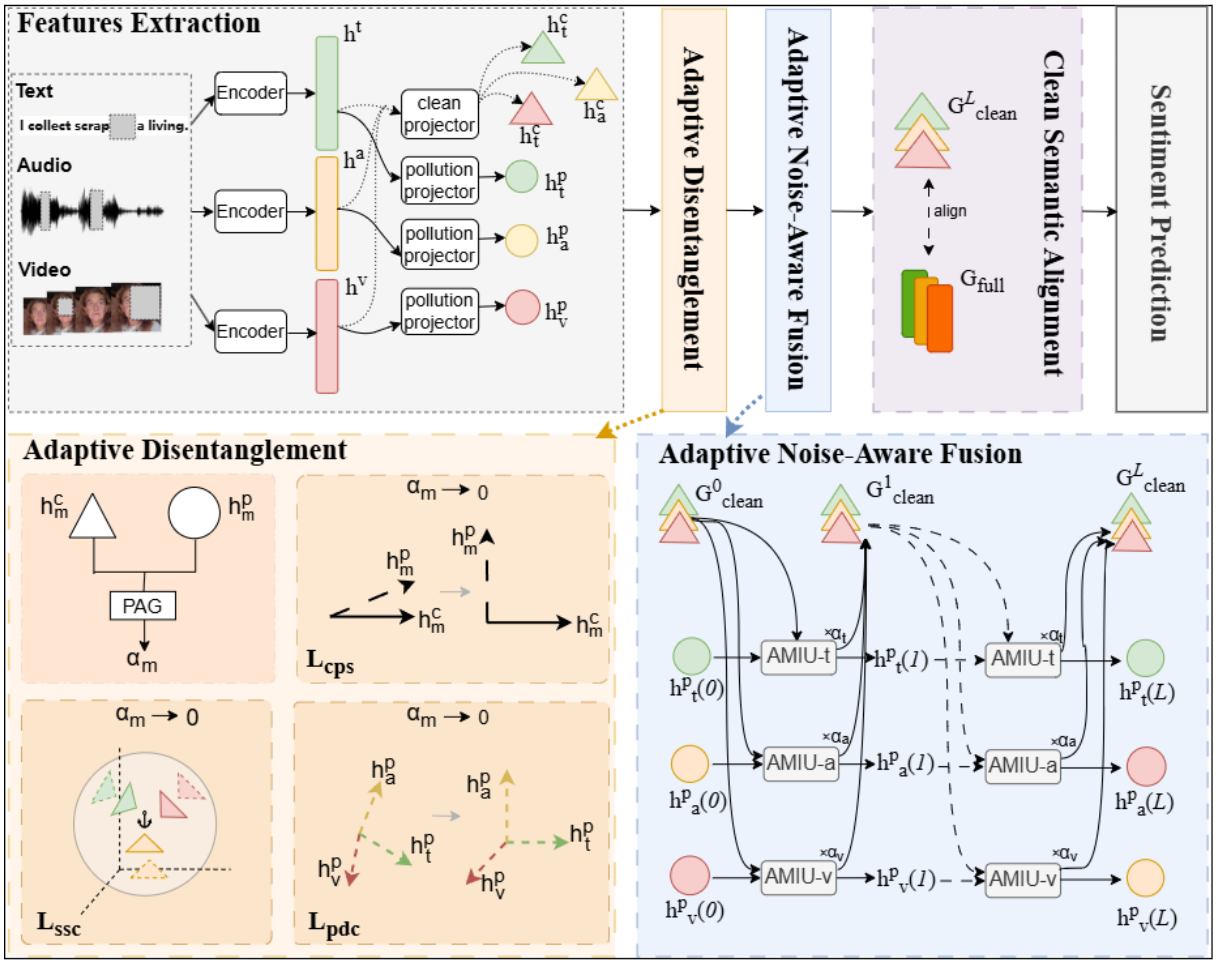

Figure 1: ADRMSA Model Structure.

look the noisy features introduced by random local corruption in RMFM scenarios and lack explicit mechanisms to suppress noise propagation during cross-modal interaction. *Our work falls into the category of joint representation learning. Specifically, ADRMSA adopts a disentangle–fuse–reconstruct tri-stage collaborative optimization strategy that explicitly separates modality-shared and modality-specific features, and combines noise-aware adaptive fusion with local-global semantic reconstruction. This design effectively suppresses noise propagation under Random Modality Feature Missing (RMFM), leading to more robust multimodal sentiment analysis.*

3 Method

Fig. 1 shows the ADRMSA framework. It consists of three core modules that together form a hierarchical noise-aware compensation architecture.

3.1 Data Preprocessing & Feature Extraction

We employ BERT (Devlin et al., 2019) and BiLSTM (Huang et al., 2015) to extract shallow features from the textual, visual, and acoustic modalities, denoted as $\tilde{U}_t \in \mathbb{R}^{l_t \times d_t}$, $\tilde{U}_v \in \mathbb{R}^{l_v \times d_v}$, and $\tilde{U}_a \in \mathbb{R}^{l_a \times d_a}$, respectively, where l and d denote the sequence length and hidden dimension. To simulate Random Modality Feature Missing (RMFM), we adopt a multimodal masking strategy during training. Specifically, for the text modality, tokens within the sequence are randomly replaced with the [UNK] token according to a predefined missing rate (MR). For audio and visual modalities, temporal positions within the feature sequences are randomly masked, and the corresponding feature vectors are set to zero. Formally, for each modality, a binary mask $g_m \in \{0,1\}^{l_m}$ is randomly generated to construct corrupted feature sequences as $U_m = g_m \odot \tilde{U}_m$, where $m \in \{t, a, v\}$. The missing rate varies from 0% to 90%, covering diverse RMFM scenarios. The cor-

rupted features are then encoded into utterance-level representations: $h_v = \text{BiLSTM}(U_v), \quad h_a = \text{BiLSTM}(U_a), \quad h_t = \text{BERT}(U_t)$.

3.2 Adaptive Disentanglement (AD)

To alleviate representation pollution caused by missing noise before cross-modal interaction under RMFM, we propose an Adaptive Disentanglement (AD) module, which first performs polluted semantic separation to explicitly distinguish stable sentiment semantics from corrupted information before interaction.

Given modality representation h_m, $m \in \{t, a, v\}$,we employ two independent lightweight projectors to map the input representation into clean and polluted semantic subspaces: $h_m^c = f_c^m(h^m)$,$h_m^p = f_p^m(h^m)$, Here, h_m^c is encouraged to preserve stable cross-modal sentiment semantics via semantic cohesion constraints, while h_m^p is driven to capture residual corrupted information and unstable modality-specific perturbations through decorrelation and separation objectives, where $f_c^m(\cdot)$ and $f_p^m(\cdot)$ are implemented as two-layer MLPs with non-shared parameters. The two projection encoders disentangle modality representations by projecting the input into two approximately orthogonal subspaces:

$$(h_m^c)^\top h_m^p \approx 0. \tag{1}$$

Since different samples exhibit varying corruption levels under RMFM, we further introduce a Pollution-Adaptive Gate (PAG) to compute a Residual Semantic Reliability coefficient: $\gamma_m = \sigma\big(W_g[h_m^c; h_m^p] + b_g\big)$, where $\sigma(\cdot)$ denotes the sigmoid function. γ_m[2] $\in (0, 1)$ is learned from the interaction pattern between h_m^c and h_m^p, serving as a learned proxy that estimates how much reliable residual sentiment semantics are preserved in the polluted representation. Under severe corruption, the discrepancy between h_m^c and h_m^p becomes larger, the polluted branch tends to contain fewer stable sentiment semantics, leading to a smaller γ_m. Consequently, stronger disentanglement constraints are adaptively imposed in subsequent optimization.

Based on PAG, we construct three adaptive objectives. First, to enhance cross-modal consistency among clean representations, we introduce the Shared Semantic Cohesion Loss:

$$D = cos(h_m^c, h_n^c) - cos(h_m^c, h_m^p), \tag{2}$$

where $cos(\cdot, \cdot)$ denotes the cosine distance. Before distance computation, all representations are L_2-normalized to stabilize semantic similarity estimation across different missing patterns.

$$L_{ssc} = \frac{1}{6} \sum_{m \neq n} (1 - \gamma_m) \log(1 + \exp(D)), \tag{3}$$

this objective encourages clean representations from different modalities to remain close while being separated from polluted representations. Smaller γ_m further strengthens semantic aggregation when modality semantics are more unreliable.

Next, to suppress pollution propagation across modalities, we introduce the Hilbert-Schmidt Independence Criterion (HSIC):

$$\text{HSIC}(h_i^p, h_j^p) = \frac{1}{(n-1)^2} \operatorname{tr}(K_i H K_j H), \tag{4}$$

where H denotes the centering matrix, and K_i and K_j are kernel matrices computed using the Gaussian RBF kernel:

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right), \tag{5}$$

where the bandwidth σ is determined using the median heuristic on each mini-batch. All features are normalized before HSIC computation.The Pollution Decorrelation Loss is formulated as:

$$L_{pdc} = \frac{1}{3} \sum_{i \neq j} (1 - \gamma_i \gamma_j) \ \text{HSIC}(h_i^p, h_j^p), \tag{6}$$

this objective reduces dependency among pollution patterns from different modalities, thereby mitigating pollution propagation during cross-modal interaction.

We further separate clean and polluted representations via the Clean-Pollution Separation Loss:

$$L_{cps} = \frac{1}{3} \sum_{m} (1 - \gamma_m)^2 \ \text{HSIC}(h_m^c, h_m^p), \tag{7}$$

the quadratic gating term provides a smoother transition from soft to hard disentanglement by gradually increasing separation strength only when corruption becomes severe.

The final objective of AD is defined as:

$$L_{AD} = L_{ssc} + L_{pdc} + L_{cps}. \tag{8}$$

Through these designs, AD dynamically suppresses representation pollution before cross-modal interaction and provides a cleaner semantic foundation for subsequent robust semantic interaction.

[2] In this work, γ_m is a scalar gate shared across all feature dimensions rather than a vector-valued gate.

3.3 Adaptive Noise-Aware Fusion

Polluted representations h_m^p may still preserve partial sentiment-related semantics, while simultaneously containing severe noise interference. Direct cross-modal fusion would inject noise into the shared space, causing accumulation.

To address this, we propose the Adaptive Noise-Aware Fusion (**ANAF**) module, which aims to progressively extract reliable residual semantics from polluted representations.

Fig. 6 in Appendix B illustrates the Adaptive Modal Interaction Unit (AMIU), which serves as the fundamental interaction component of ANAF. The AMIU provides the full implementation details of the interaction process, while the formulation in the main text presents a compact abstraction for clarity.

Each AMIU takes as input the current clean global context $G_{\text{clean}}^{[i]}$ and the polluted representation $h_m^p(i)$ of modality m, and outputs an update $\Delta G_m^{[i]}$ that captures residual sentiment-related semantics.

Meanwhile, γ_m generated by the AD module is incorporated into the AMIU to adaptively control the contribution of residual information during interaction.

To achieve deep cross-modal fusion, ANAF stacks multiple AMIUs in a progressive manner. Early layers primarily extract coarse residual semantics from polluted features, while deeper layers refine sentiment representations based on progressively purified global context.

Specifically, the clean representations generated by the AD module are first concatenated to construct the initial global sentiment context: $G_{\text{clean}}^{[0]} = \text{concat}\left(h_t^c, h_v^c, h_a^c\right)$ .The interaction at the i-th layer can be written in compact form as:

$$\Delta G_m^{[i]} = \text{AMIU}_{m\to\text{clean}}\left(G_{\text{clean}}^{[i]}, h_m^p(i)\right), \quad (9)$$

where $\text{AMIU}_{m\to\text{clean}}(\cdot)$ denotes the full cross-modal interaction defined in Appendix B, which retrieves informative residual semantics from polluted representations to update the clean context.

Considering that different modalities exhibit different corruption levels, we further employ γ_m as a reliability-aware coefficient to dynamically regulate residual semantic injection:

$$G_{\text{clean}}^{[i+1]} = G_{\text{clean}}^{[i]} + \sum_{m\in\{t,a,v\}} \gamma_m \cdot \Delta G_m^{[i]}. \quad (10)$$

Through this progressive interaction, ANAF continuously extracts residual semantic information from polluted representations while refining the global context, ultimately producing a more robust sentiment representation $G_{\text{clean}}^{[L]}$.

3.4 Clean Semantic Alignment (CSA)

After L layers of ANAF interaction, the model produces a progressively refined global representation $G_{\text{clean}}^{[L]}$. Although the AD module isolates noise and ANAF further suppresses noise propagation while extracting residual semantics, due to the randomness of RMFM, the final representation may still suffer from semantic drift and partial sentiment loss.

To further improve semantic consistency, we introduce the **Clean Semantic Alignment (CSA)** module, which performs semantic-level alignment between refined incomplete representations and complete semantic representations. Specifically, the complete multimodal input $(\tilde{U}_t, \tilde{U}_a, \tilde{U}_v)$ is first encoded into the full semantic representation:$G_{\text{full}} = \text{concat}(\tilde{h}_t, \tilde{h}_a, \tilde{h}_v)$. To alleviate semantic deviations caused by random missing features, CSA adopts a SimSiam-style one-way semantic alignment objective, which pulls the refined representation toward the complete semantic representation without enforcing the reverse alignment. More details about SimSiam could be found in Appendix C.This design avoids perturbing the stable semantic structure of complete representations.

The refined representation is first projected by a projector $z(\cdot)$ and a predictor $p(\cdot)$: $z_c = z(G_{\text{clean}}^{[L]})$, $z_f = z(G_{\text{full}}), p_c = p(z_c)$. The semantic alignment objective is defined as:

$$L_{\text{CSA}} = -\frac{p_c \cdot \text{sg}(z_f)}{\|p_c\| \cdot \|\,\text{sg}(z_f)\|}, \quad (11)$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operation, the projector $z(\cdot)$ maps heterogeneous representations under different missing patterns into a unified semantic space, reducing modality distribution discrepancies, while the predictor $p(\cdot)$ further enhances the semantic fitting ability of the refined representation.

Through this one-way semantic alignment strategy, the complete semantic representation serves as a stable anchor that progressively guides $G_{\text{clean}}^{[L]}$ toward semantically consistent sentiment structures, thereby improving robustness under severe RMFM conditions.

It is worth noting that CSA is only employed during training as an auxiliary semantic supervision strategy. During inference, the model relies solely on incomplete multimodal inputs without requiring any additional complete-modal information.

3.5 Training Objective

During training, ADRMSA processes both complete modality input $((\tilde{U}_t, \tilde{U}_a, \tilde{U}_v))$ and noisy input $((U_t, U_a, U_v))$. Therefore, the task loss is:

$$L_{task} = \sum_{i=1}^{N} \left(|y_i - \hat{y}_i| + |y_i - \hat{\tilde{y}}_i| \right), \quad (12)$$

where N is the number of samples, y_i is the sentiment label, and $\hat{y}_i$ and $\hat{\tilde{y}}_i$ are the predicted sentiments for the complete and noisy modalities of sample i, respectively. The overall objective is defined as:

$$L = L_{\text{task}} + \alpha L_{AD} + \beta L_{\text{CSA}}. \quad (13)$$

By jointly integrating AD, ANAF, and CSA, the proposed framework systematically mitigates RMFM-induced performance degradation from the perspectives of pollution isolation, robust interaction, and semantic alignment.

4 Experimentals

4.1 Experimental-Settings

We use the CMU-MOSI and CMU-MOSEI datasets with fully observed modalities. For each metric, we compute the Area Under Indicators Line Chart (AUILC) to evaluate the model's overall performance under varying levels of modality missing. Given increasing missing rates $\{p_0, p_1, ..., p_t\}$ and corresponding metric values $\{v_0, v_1, ..., v_t\}$, AUILC is defined as:

$$AUILC = \sum_{i=0}^{t-1} \frac{1}{2} (v_i + v_{i+1}) (p_{i+1} - p_i). \quad (14)$$

More details about implementations and baselines are shown in Appendix D.

4.2 Comparison with Baselines

Table 1 reports the overall performance on MOSI and MOSEI under RMFM missing ratios ranging from 0.0 to 0.9. ADRMSA achieves the best overall performance on most evaluation metrics, particularly on Acc, F1, and Corr, demonstrating stronger robustness under severe feature corruption. On MOSI, ADRMSA consistently outperforms previous RMFM methods such as RMSAGF and TCTR while maintaining competitive MAE performance. On the more challenging MOSEI dataset, ADRMSA further achieves superior classification and correlation results, indicating that the proposed AD, ANAF, and CSA modules effectively mitigate representation pollution and suppress noise propagation during cross-modal interaction.

Table 2 further validates the effectiveness of ADRMSA on the multilingual CH-SIMS dataset. ADRMSA achieves the best performance on Acc-5, Acc-2, F1, and Corr, demonstrating strong cross-dataset generalization ability under multilingual RMFM scenarios. Additional comparisons on complete-modality settings and earlier baselines are provided in Appendix E and Appendix F, respectively.

4.3 Ablation Study

Table 3 shows ablation experiments.

"-AD". The AD module promotes cross-modal information sharing while preserving modality-unique features by separating each modality into shared and specific subspaces. Removing this module impairs the model's ability to align information across modalities, reducing sentiment analysis accuracy. AD mitigates the modality gap and provides richer emotional representations, balancing shared and modality-specific characteristics in multimodal processing.

"-ANAF". The ANAF module enables complementary integration of features across modalities via interactive learning of fine-grained representations, enhancing robustness under noisy and incomplete data. Removing ANAF significantly weakens performance in scenarios with missing or partial modalities, decreasing sentiment recognition accuracy. By dynamically adjusting the fusion of modality features, ANAF fills information gaps and stabilizes model performance.

"-CSA". The CSA Module improves the model's adaptability to incomplete or noisy data, effectively restoring missing features. Its removal leads to severe performance degradation under missing-modality conditions, particularly in realistic RMFM scenarios. By supplementing missing information, CSA enhances noise tolerance and ensures stable and accurate sentiment predictions in complex multimodal environments.

Method	MOSI						MOSEI					
	Acc-7	Acc-5	Acc-2	F1	MAE	Corr	Acc-7	Acc-5	Acc-2	F1	MAE	Corr
EMT-DLFR	32.54	35.61	69.69/70.39	69.69/70.34	110.69	48.61	47.91	48.83	76.43/76.91	75.24/75.92	66.51	54.61
P-RMF	30.80	34.72	64.46/65.62	64.60/65.88	**93.4**	40.50	40.16	41.27	71.02/70.48	72.31/71.48	**59.10**	53.60
TF-Mamba	31.52	34.87	71.28/70.54	70.95/70.12	108.20	50.20	43.89	44.76	75.61/75.93	75.24/75.52	72.10	53.70
DAST-GAN	33.19	36.01	69.70/70.46	69.66/70.49	110.83	49.04	48.05	**50.28**	77.13/77.99	76.27/77.27	64.00	56.30
RMSAGF	32.90	35.55	70.88/71.19	70.89/71.20	109.10	49.90	48.70	48.30	76.74/77.33	76.18/76.74	65.61	56.90
Meta-RoMSA	–	–	–	–	–	–	46.30	47.10	74.00/75.40	74.00/74.60	69.90	53.50
ADRMSA	**33.98**	**36.10**	70.49/**70.61**	**70.97/71.39**	110.35	**52.45**	**48.72**	50.26	**77.81/78.16**	**77.39/77.56**	66.89	**57.00**

Table 1: Comparison of the overall performance on MOSI and MOSEI under different missing rates (0.0-0.9), '–':No data are available in the original paper, and the results are not reproducible.

Method	Acc-5	Acc-3	Acc-2	F1	MAE	Corr
EMT-DLFR	32.78	**56.70**	70.56	70.41	51.00	39.40
P-RMF	31.42	49.46	66.38	67.29	**44.90**	37.80
TF-Mamba	30.11	54.27	70.14	70.31	55.60	33.50
DAST-GAN	–	–	–	–	–	–
RMSAGF	–	–	–	–	–	–
Meta-RoMSA	–	–	–	–	–	–
ADRMSA	**33.18**	54.40	**71.09**	**70.92**	54.03	**40.17**

Table 2: The comparison with baselines on CH-SIMS. '–' denotes either that the original paper only provides data for missing rates within [0.0, 0.5] or that no data are available.

Dataset	Model	ACC-2	F1	MAE	CORR
MOSI	ADRMSA	**78.97/79.05**	**78.00/79.15**	**90.17**	**66.88**
	-AD	76.43/76.88	76.73/76.79	92.84	65.26
	-ANAF	75.81/76.82	75.86/75.96	91.65	65.95
	-CSA	76.04/76.81	74.19/75.56	93.94	65.42
MOSEI	ADRMSA	**81.76/82.34**	**81.78/82.37**	**59.97**	**68.28**
	-AD	79.22/80.16	79.21/80.23	60.23	67.70
	-ANAF	79.41/80.63	79.44/80.73	60.61	67.43
	-CSA	79.69/80.85	79.59/80.79	60.19	67.84

Table 3: Ablation Experiment (missing rate 30%).

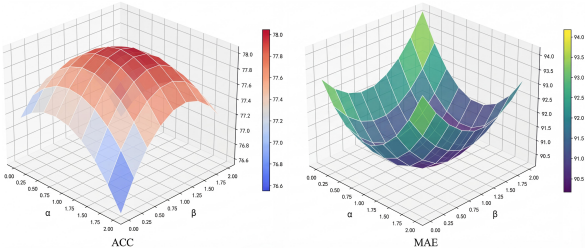

Figure 2: Parametric analysis of experimental results.

4.4 Hyper-parameter Study

Figure 2 shows the loss weights experiments. Based on prior experience and evaluation of the task loss function, the hyper-parameters are preliminarily set within the range [0, 2], ensuring that no single module dominates the total loss and maintaining a balanced contribution during training.

We start with a value of 0.5 for each parameter and increment by 0.5 to evaluate performance within this range, using ACC and MAE as primary metrics. To assess parameter sensitivity under noise, all analyses are conducted at a 30% missing rate. As shown in Figure 2, ACC and MAE values are visualized as 3D surfaces, where higher color saturation indicates better performance. We can observed that the increase of α value enhances the extraction of cross-modal shared features, and the cross-modal alignment strength and modal specificity retention reach a moderate balance around 1.0. However, exceeding a certain threshold inhibited modal-specific expression, and when $\alpha \geq 1.5$, the MAE of MOSI increased by 2.8 and the ACC of MOSEI decreased by 1.5%. β value determines the knowledge transfer ability of complete modal data, and when β=0.5, the model already has a certain semantic recovery ability. With the increase of value, the noise reduction ability is gradually enhanced. When $\beta \geq 1.5$, the model has a tendency to over-fit local noise, with a 1% drop in both MOSI and F1 on MOSEI, and the L1 loss gradient explosion risk of the MLP decoder also increases due to the increased computational overhead. Considering that the reconstruction loss weight β around 1.0 can not only ensure the noise immunity of the model, but also increase the generalization.

4.5 Analysis of the Residual Semantic Reliability Coefficient γ

Fig. 3 shows that γ consistently decreases as the missing ratio increases across all modalities, indicating that PAG effectively captures the degradation of residual semantic reliability under RMFM. When the missing ratio is low, polluted representations still preserve relatively stable sentiment-related semantics, resulting in larger γ values and allowing more residual information to participate in semantic interaction. As corruption becomes more severe, the semantic discrepancy between clean and polluted representations gradually increases,

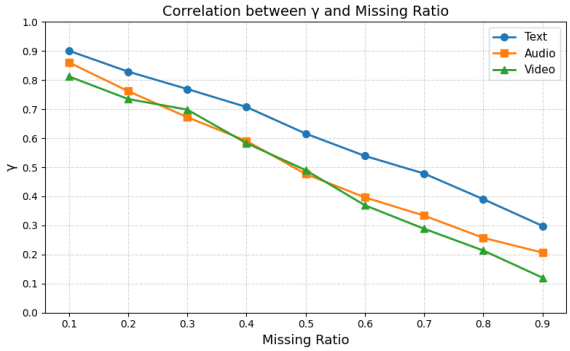

Figure 3: Correlation between the Residual Semantic Reliability Coefficient γ and Missing Ratio

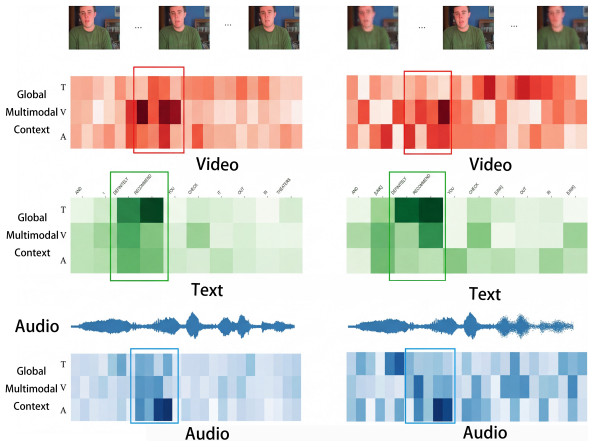

Figure 4: Cross-modal attention matrix heatmap.

causing γ to decrease and thereby reducing the influence of unreliable residual semantics during fusion.

Moreover, compared with audio and visual modalities, the text modality maintains relatively higher γ values across different missing ratios, suggesting that textual semantics are inherently more robust to random feature corruption.

4.6 Case Study

In order to further explore the modeling mechanism of the ADRMSA model for multimodal sequences in the RMFM noise environment, In this study, the signal capture ability of ANAF module is empirically analyzed by cross-modal attention weight visualization method. As shown in Figure 4, a typical sample with positive sentiment polarity in the MOSI dataset (original text: "AND I DEFINITELY RECOMMEND YOU CHECK IT OUT IN THEATERS") is taken for a case study. In order to simplify the visualization, this study only shows the distribution of attention of global multimodal contexts to each unimodal feature (i.e.,$m \rightarrow g$, $m \in \{t, a, v\}$). Given the large length of the speech and visual modal sequences (42 and 50 frames, respectively), an average pooling operation with a step size of 2 was used for downsampling.

Figure 4 The left half shows the distribution of attention without noise, where color saturation is positively correlated with attention weight. The analysis shows that the ANAF module successfully captures the global-local cross-modal interaction pattern with significant semantic association. Specifically, the performance is as follows: (1) in the video mode, the global context shows significant attention to several facial expression frames that represent pleasant emotions; (2) The peak of attention in speech mode is concentrated in the segments with high fundamental frequency of acoustic characteristics. (3) Textual modality gives the greatest weight to the emotionally inclined vocabulary combination "DEFINITELY RECOMMEND".

The right half of the figure presents the results of the comparison experiment under RMFM noise interference (the drop rate is set to 0.3). With this setting, missing audio and video features are filled with zero vectors, and missing markers in text modality are replaced by BERT's [UNK] placeholder (the blurred background schematic in the figure indicates where the feature is missing, which may deviate from the actual time mask). Experimental results show that although the attention matrix is dispersed due to RMFM noise, ADRMSA can still effectively maintain the focus intensity of key cross-modal signals under noise-free conditions, showing strong noise robustness.

5 conclusion

We propose ADRMSA, a robust multimodal sentiment analysis framework for Random Modality Feature Missing (RMFM). ADRMSA alleviates representation pollution via adaptive disentanglement, noise-aware fusion, and clean semantic alignment. Specifically, it separates clean and polluted representations before interaction, extracts reliable residual semantics from corrupted features, and corrects semantic drift through alignment. Extensive experiments on benchmark datasets demonstrate its effectiveness and robustness under various missing conditions. Future work includes extending ADRMSA to more complex missing scenarios and improving its efficiency for real-time applications.

Limitations

Despite its strong performance, ADRMSA has a few limitations. First, under extremely high missing ratios (above 90%), the residual semantics in polluted representations becomes too sparse for ANAF to extract useful signals, leading to marginal gains over simpler methods. Second, the current CSA module relies on complete input as an anchor during training, which may not be available in fully unsupervised missing scenarios. Extending to self-supervised alignment without complete data is a promising future direction.

References

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, Melbourne, Australia. Association for Computational Linguistics.

Abhra Chaudhuri, Anjan Dutta, Tu Bui, and Serban Georgescu. 2025. A closer look at multimodal representation collapse. *Preprint*, arXiv:2505.22483.

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: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. 2023. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Inf. Fusion*, 91:424–444.

Peizhu Gong, Jin Liu, Xiliang Zhang, Xingye Li, Lai Wei, and Huihua He. 2026. Towards robust sentiment analysis with multimodal interaction graph and hybrid contrastive learning. *Pattern Recogn.*, 169(C).

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. MISA: modality-invariant and -specific representations for multimodal sentiment analysis. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 1122–1131. ACM.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *Preprint*, arXiv:1508.01991.

Aghiles Kebaili, Jérôme Lapuyade-Lahorgue, Pierre Vera, and Su Ruan. 2025. Amm-diff: Adaptive multimodality diffusion network for missing modality imputation. In *22nd IEEE International Symposium on Biomedical Imaging, ISBI 2025, Houston, TX, USA, April 14-17, 2025*, pages 1–4. IEEE.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Mingcheng Li, Dingkang Yang, Yuxuan Lei, Shunli Wang, Shuaibing Wang, Liuzhen Su, Kun Yang, Yuzheng Wang, Mingyang Sun, and Lihua Zhang. 2024. A unified self-distillation framework for multimodal sentiment analysis with uncertain missing modalities. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Xiang Li, Xianfu Cheng, Dezhuang Miao, Xiaoming Zhang, and Zhoujun Li. 2025. TF-mamba: Text-enhanced fusion mamba with missing modalities for robust multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11252–11267, Suzhou, China. Association for Computational Linguistics.

Yang Li, Bin Gao, Xiaobai Li, Linlin Li, Yutong Li, Shutian Liu, and Zhengjun Liu. 2026. Rmsagf: A synergistic generation and fusion framework for robust multimodal sentiment analysis. *Expert Systems with Applications*, 316:131841.

Ronghao Lin and Haifeng Hu. 2023. Missmodal: Increasing robustness to missing modality in multimodal sentiment analysis. *Trans. Assoc. Comput. Linguistics*, 11:1686–1702.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2247–2256. Association for Computational Linguistics.

Duong M. Nguyen, Trong Nghia Hoang, Thanh Trung Huynh, Quoc Viet Hung Nguyen, and Phi Le Nguyen. 2025. Learning reconfigurable representations for multimodal federated learning with missing data. *Preprint*, arXiv:2510.22880.

Lei Pan and Hongqiao Li. 2026. Meta-romsa: A reciprocal translation and modality balancing framework for robust multimodal sentiment analysis using meta-learning. *IEEE Access*, 14:2034–2044.

Peng Shi, Jin Wang, and Xuejie Zhang. 2026. Language-dominated fusion and self-distillation for multimodal sentiment analysis with incomplete modalities. *Pattern Recognition*, 172:112644.

Ge Su, Kaiping Zheng, Tiancheng Zhao, and Jianwei Yin. 2025. Clear: Addressing representation contamination in multimodal healthcare analytics. *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*.

Bin Sun, Li Jia, Yiming Cui, Na Wang, and Tao Jiang. 2025. Conv-enhanced transformer and robust optimization network for robust multimodal sentiment analysis. *Neurocomputing*, 634:129842.

Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. 2022. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *CoRR*, abs/2208.07589.

Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. 2024. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE Trans. Affect. Comput.*, 15(1):309–325.

Qiuling Suo, Weida Zhong, Fenglong Ma, Ye Yuan, Jing Gao, and Aidong Zhang. 2019. Metric learning on healthcare data with incomplete modalities. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 3534–3540. ijcai.org.

Wenlong Tan and Bo Zhang. 2025. Dast-gan: An adversarial learning multimodal sentiment analysis model based on dynamic attention and spatio-temporal fusion. *Sensors*, 25(23).

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6558–6569. Association for Computational Linguistics.

Yuanzhi Wang, Yong Li, and Zhen Cui. 2023. Incomplete multimodality-diffused emotion recognition. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Pengcheng Weng, Yanyu Qian, Yangxin Xu, and Fei Wang. 2026. Purify-then-align: Towards robust human sensing under modality missing with knowledge distillation from noisy multimodal teacher. *Preprint*, arXiv:2604.05584.

Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. 2026. Deep multimodal learning with missing modality: A survey. *Preprint*, arXiv:2409.07825.

Zhihao Yang, Qing He, Minghao Yu, Nisuo Du, and Yijie Lu. 2026. Tctr: Text-guided contrastive learning with token-level reconstruction network for missing modalities in multimodal sentiment analysis. *Information Fusion*, 126:103571.

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, Online. Association for Computational Linguistics.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. *CoRR*, abs/2102.04830.

Ziqi Yuan, Wei Li, Hua Xu, and Wenmeng Yu. 2021. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 4400–4407. ACM.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017a. Tensor fusion network for multimodal sentiment analysis. *Preprint*, arXiv:1707.07250.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017b. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1103–1114. Association for Computational Linguistics.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *Preprint*, arXiv:1606.06259.

Duoyi Zhang. 2025. *Deep Learning Frameworks for Addressing Feature-Level Bias in Multimodal Fusion and Classification*. Ph.D. thesis, Queensland University of Technology.

Haoyu Zhang, Wenbin Wang, and Tianshu Yu. 2024. Towards robust multimodal sentiment analysis with incomplete data. In *Advances in Neural Information Processing Systems*, volume 37, pages 55943–55974. Curran Associates, Inc.

Ting Zhang, Bin Song, Zhiyong Zhang, and Yajuan Zhang. 2025. Multimodal sentiment analysis based on multi-stage graph fusion networks under random missing modality conditions. *IET Image Processing*, 19(1):e13310.

Kai Zhao, Mingsheng Zheng, Qingguan Li, and Jianing Liu. 2025. Multimodal sentiment analysis—a comprehensive survey from a fusion methods perspective. *IEEE Access*, 13:64556–64583.

Guowei Zhong, Ruohong Huan, Mingzhen Wu, Ronghua Liang, and Peng Chen. 2025. Towards robust multimodal emotion recognition under missing modalities and distribution shifts. *CoRR*, abs/2506.10452.

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Ning An. 2025. Proxy-driven robust multimodal sentiment analysis with incomplete data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22123–22138, Vienna, Austria. Association for Computational Linguistics.

A Missing Modality Scenarios

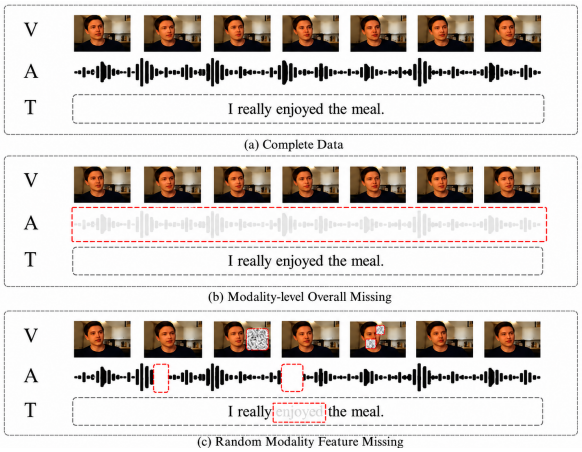

Figure 5: Illustration of Missing Modality Scenarios in Multimodal Sentiment Analysis

As illustrated in Fig. 5, we categorize missing data scenarios into two typical types to simulate real-world challenges encountered in multimodal systems.

Modality-level Overall Missing. In this setting, an entire modality (e.g., audio) is completely lost, while the remaining modalities (visual and text) stay intact. Such a situation usually stems from sensor failure or recording errors, and it necessitates recovering the missing signal from the available modalities.

Random Modality Feature Missing. This is a more fine-grained and realistic setting, where data loss occurs sporadically within each modality over time, such as randomly dropped video frames or intermittent gaps in an audio waveform. Accordingly, the model is required to handle incomplete temporal sequences and to reconstruct the fragmented emotional cues.

B AMIU

Specifically, given two input sequences H_p and H_q, after processing by $AMIU_{m\leftrightarrow n}$, the mutually enhanced outputs $H_{p\to q}$ and $H_{q\to p}$ are obtained. The arrow ($\to$) indicates the direction of information flow. Taking $q \to p$ as an example:

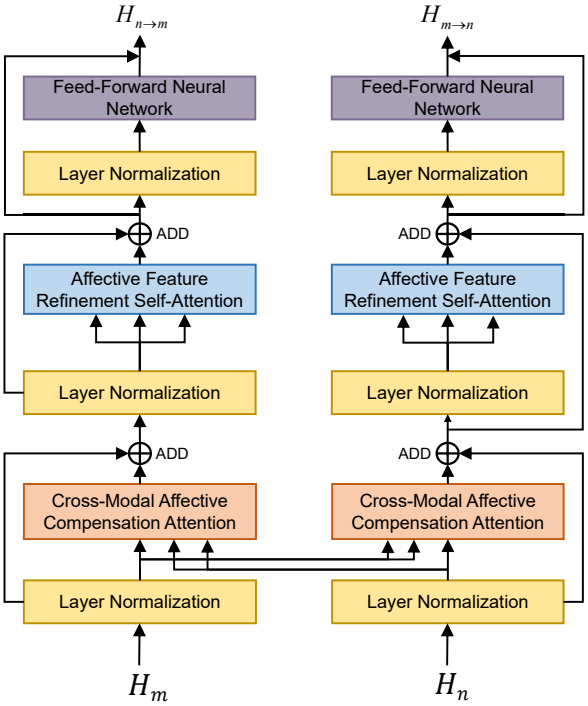

Figure 6: Basic Structure of AMIU

$$H'_{q\to p}=CSCA\left(LN\left(H_p\right), LN\left(H_q\right)\right)+H_p, \quad (15)$$

$$H''_{q\to p}=SFRSA\left(LN\left(H'_{q\to p}\right)\right)+H'_{q\to p}, \quad (16)$$

$$H_{q\to p}=FFN\left(LN\left(H''_{q\to p}\right)\right)+H''_{q\to p}, \quad (17)$$

where $CSCA(\cdot)$ denotes Cross-modal Sentiment Compensatory Attention, which captures cross-modal semantic associations via a query–key–value mechanism. $SFRSA(\cdot)$ denotes Sentiment Features Refined Self-Attention, which further models long-range intra-modal dependencies and suppresses temporal noise in local features. LN denotes Layer Normalization, and FFN is a position-wise feed-forward network that enhances feature discrimination through non-linear transformations.

To balance model capacity and generalization, we design a triple parameter-sharing mechanism: AMIU-level, modality-level, and layer-level sharing.

AMIU-level sharing implements bidirectional parameter sharing within a single AMIU unit (i.e.,$AMIU_{n\to m}$ and $AMIU_{m\to n}$ share parameters), ensuring consistency in inter-modal interactions and stability in feature fusion.

Modality-level sharing shares parameters when local representations from different modalities interact with the global context(i.e., $AMIU^{i}_{m\leftrightarrow g}$).

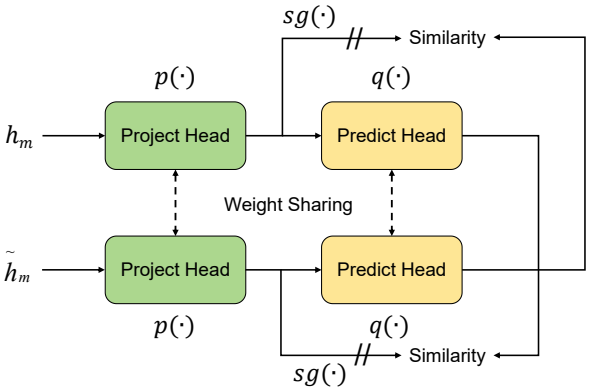

Figure 7: SimSiam Branch Structure

This cross-modal sharing encourages uniform feature extraction across modalities, forcing the model to learn modality-agnostic interaction patterns and better capture cross-modal shared semantics.

Layer-level sharing applies the same AMIU parameters across different layers. Regardless of which layer the features pass through, the interaction parameters remain consistent, reducing model complexity and enhancing feature transfer across layers. This enables low-level and high-level features to be co-optimized, improving representational capacity and convergence efficiency.

C SimSiam

Inspired by SimSiam (Simple Siamese Network), a self-supervised learning method based on a dual-branch network, its core idea is to maximize consistency between positive sample pairs (clean and noisy views processed by weight-shared encoders), suppressing semantic drift caused by missing noise. The SimSiam branch structure is shown in Fig 7. To adapt to multimodal specifics, a projection head $p(\cdot)$ maps high-dimensional global features to a low-dimensional contrastive space, implemented by two fully connected layers with ReLU activation. A prediction head $q(\cdot)$ (a single fully connected layer) enhances the model's adaptability to noise distributions through non-linear feature transformation.

The loss function uses the negative cosine similarity to measure positive pair consistency, employing a Stop-Gradient (sg) operation to prevent trivial solutions:

$$L_{sim}^{m} = -\frac{1}{2}\Big[\frac{q(p(\tilde{h}_m)) \cdot \mathrm{sg}(p(h_m))}{\|q(p(\tilde{h}_m))\| \cdot \|\mathrm{sg}(p(h_m))\|} + \frac{q(p(h_m)) \cdot \mathrm{sg}(p(\tilde{h}_m))}{\|q(p(h_m))\| \cdot \|\mathrm{sg}(p(\tilde{h}_m))\|}\Big]. \quad (18)$$

To promote interaction within the ANAF module, feature reconstruction is also applied to the global multimodal context $G^{[L]}$:

$$L_{sim}^{g} = -\frac{1}{2}\Big[\frac{q(p(\tilde{G}^{[L]})) \cdot \mathrm{sg}(p(G^{[L]}))}{\|q(p(\tilde{G}^{[L]}))\| \cdot \|\mathrm{sg}(p(G^{[L]}))\|} + \frac{q(p(G^{[L]})) \cdot \mathrm{sg}(p(\tilde{G}^{[L]}))}{\|q(p(G^{[L]}))\| \cdot \|\mathrm{sg}(p(\tilde{G}^{[L]}))\|}\Big]. \quad (19)$$

D Experimental Settings

.

During training, the Adam optimizer is employed in combination with a StepLR scheduler. For the BERT model, the initial learning rates are set to 5e-5 for CMU-MOSI and 2e-5 for CMU-MOSEI, while the remaining model parameters are initialized with learning rates of 1e-3 and 1e-4, respectively. Early stopping with a patience of five is applied to prevent overfitting. All experiments are conducted on an NVIDIA RTX 4090 GPU to ensure fair comparison.

To evaluate the performance of ADRMSA, we conduct a comparative analysis with six benchmark models. LMF (Liu et al., 2018) improves TFN (Zadeh et al., 2017b) by performing low-rank multimodal tensor fusion to enhance efficiency. MulT (Tsai et al., 2019) learns representations directly from unaligned multimodal streams by modeling attention between modalities, addressing heterogeneity and misalignment in multimodal language sequences. MISA (Hazarika et al., 2020) projects each modality into two subspaces to provide a comprehensive view of multimodal data. SelfMM (Yu et al., 2021) obtains independent emotion labels for each modality using a self-supervised learning strategy to assist multimodal emotion prediction. TRF-Net (Yuan et al., 2021) employs a feature extractor that models both intra- and inter-modality interactions to enhance robustness against randomly missing non-aligned modalities. EMT-DLFR (Sun et al., 2024) leverages an efficient cross-modal interaction mechanism and a feature-view repair module to improve model robustness under randomly missing modality features. TF-Mamba (Li et al., 2025) A text-enhanced fusion Mamba framework

that treats text as the dominant modality, integrating text-aware enhancement into efficient Mamba-based architectures for robust MSA under missing modality conditions. P-RMF (Zhu et al., 2025) A data-centric robust fusion method that learns stable unimodal representations from the latent Gaussian space of missing data and generates a robust proxy modality by incorporating quantized uncertainty to guide dynamic multimodal fusion. DAST-GAN (Tan and Zhang, 2025) A GAN-enhanced framework that combines dynamic attention for adaptive modality weighting and unified spatio-temporal attention for joint modeling of temporal and cross-modal interactions, with adversarial training to learn robust representations from incomplete data. Meta-RoMSA (Pan and Li, 2026) A meta-learning framework that employs triple modality translation to exploit cross-modal redundancy for imputing missing features, and dynamically balances modality contributions via unimodal annotation-guided weight learning. RMSAGF (Li et al., 2026) A synergistic dual-component framework featuring context-aware feature generation with dynamic reliability assessment to actively synthesize missing modality sequences, coupled with interdependent multimodal attention for deep iterative fusion.

The experiments are carried out on three widely used benchmark datasets for multimodal sentiment analysis.

CMU-MOSI(Zadeh et al., 2016) consists of 2,199 segmented monologue utterances extracted from 93 YouTube movie review videos involving 89 distinct speakers. Following the standard data split, 1,284 samples are used for training, 229 samples for validation, and 686 samples for testing.

CMU-MOSEI(Bagher Zadeh et al., 2018) is a larger-scale extension that comprises approximately 20,000 video utterances collected from 3,228 videos covering 250 diverse topics and spoken by over 1,000 YouTube contributors. Specifically, the dataset is divided into 16,326 training utterances, 1,871 validation utterances, and 4,659 test utterances.

CH-SIMS(Yu et al., 2020) serves as the primary benchmark for Chinese MSA, encompassing 60 videos from diverse sources including movies, TV series, and variety shows, totaling 2281 utterance-level segments. Unlike the English datasets, it supports both multimodal and unimodal annotations (though only multimodal annotations are utilized in this study). Emotional intensity annotations range from -1 (strongly negative) to 1 (strongly positive), emphasizing the subtleties of Chinese emotional expression and providing valuable cross-lingual validation for model robustness.

Both datasets provide continuous sentiment annotations on a Likert scale from -3 to $+3$, where the sign represents sentiment polarity (positive or negative), and the magnitude reflects the intensity of the expressed sentiment.

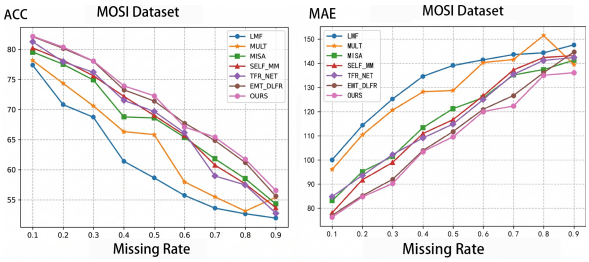

Figure 8: MAE and ACC on the MOSI dataset.

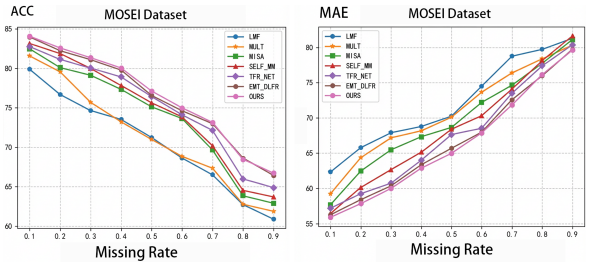

Figure 9: MAE and ACC on the MOSEI dataset.

E Comparison on Complete data (Noise-Free Data)

To evaluate the generalization ability of the ADRMSA model, we conduct comparative experiments on complete multimodal datasets. Related results can be found in Tables 4 and Tables 5.

F Comparison with Pre-2021 Baseline Methods

Figures 8 and 9 illustrate the changes in ACC and MAE metrics across missing rates ranging from 10% to 90% for different models. Overall, model performance decreases as the missing rate increases; however, our ADRMSA not only outperforms the baseline models but also maintains relatively stable performance even under high missing rates.

A closer analysis reveals three intervals in the performance trends. In the first interval (missing rate $< 50\%$), most models show relatively stable performance, except for LMF and MulT, which are sensitive to RMFM noise, indicating limited

Method	MOSI						MOSEI					
	Acc-7	**Acc-5**	**Acc-2**	**F1**	**MAE**	**Corr**	**Acc-7**	**Acc-5**	**Acc-2**	**F1**	**MAE**	**Corr**
EMT-DLFR	46.06	51.75	82.85/84.65	82.77/84.63	71.63	79.28	54.5	56.3	83.74/85.25	83.68/85.34	53.34	78.2
P-RMF	44.31	48.83	82.65/84.15	82.69/84.37	72.63	78.20	49.77	52.09	83.62/85.20	83.68/85.48	53.90	76.70
TF-Mamba	44.31	50.58	83.69/81.63	83.71/81.58	76.20	77.40	52.26	53.83	83.32/82.89	83.71/82.92	55.60	74.80
DAST-GAN	48.62	54.94	83.23/85.17	83.11/85.03	**69.81**	80.01	54.00	56.04	83.14/85.16	83.27/85.33	52.83	78.03
RMSAGF	47.80	54.50	83.86/85.57	83.74/85.43	69.83	80.20	54.80	56.72	83.93/86.41	84.04/86.32	**52.13**	77.94
Meta-RoMSA	–	–	–	–	–	–	53.40	55.30	80.50/84.70	81.10/84.48	53.90	76.90
ADRMSA	**48.90**	**55.46**	83.84/**86.57**	83.91/**86.49**	70.33	**80.38**	**55.08**	**56.98**	83.90/**86.43**	83.85/**86.43**	52.80	**79.80**

Table 4: Comparison Results on **Noise-Free Data**.– indicates that the original paper does not contain relevant data.

Method	Acc-5	Acc-3	Acc-2	F1	MAE	Corr
EMT-DLFR	43.59	67.46	80.12	80.11	39.60	62.31
P-RMF	38.95	60.61	78.34	79.69	44.10	55.00
TF-Mamba	37.86	61.93	79.65	78.92	44.10	54.80
DAST-GAN	43.61	67.52	80.53	80.51	39.80	**64.08**
RMSAGF	**43.85**	67.96	80.55	80.54	39.10	62.94
Meta-RoMSA	–	–	–	–	–	–
ADRMSA	43.45	**68.04**	**80.66**	80.98	**38.95**	62.13

Table 5: The comparison with baselines on CH-SIMS.

robustness. In the second interval (50% $\leq$ missing rate $\leq$ 80%), as more modality information is lost, performance declines sharply compared with the previous interval. On the MOSI dataset, due to its smaller size, models reach a performance bottleneck in this range. In the third interval (missing rate $>$ 80%), most emotional information is lost, and even on the larger MOSEI dataset, models cannot be trained effectively, resulting in substantial performance deterioration.

We present the comparison results with baseline methods proposed before 2021 in Tables 6, 7and 8.

Method	MOSI						MOSEI					
	Acc-7	Acc-5	Acc-2	F1	MAE	Corr	Acc-7	Acc-5	Acc-2	F1	MAE	Corr
LMF	27.84	31.34	78.31/79.57	78.22/79.55	94.32	65.76	48.00	–	80.16/81.23	80.10/81.23	58.25	72.15
MulT	37.61	42.71	78.86/80.03	78.84/80.06	89.61	70.75	52.69	54.19	81.25/83.16	81.27/83.11	56.76	73.39
MISA	40.09	45.92	82.13/83.78	82.10/83.81	76.42	78.63	–	–	82.23/84.55	82.23/84.52	55.14	74.45
Self-MM	45.92	52.48	83.24/84.82	83.19/84.82	71.62	79.32	53.69	55.40	83.14/84.86	83.13/84.86	54.09	76.80
TFR-Net	40.23	46.06	81.98/83.74	81.94/83.77	72.83	78.38	52.39	54.37	82.1/84.07	82.05/84.54	55.04	77.20
ADRMSA	**48.90**	**55.46**	**83.84/86.57**	**83.91/86.49**	**70.33**	**80.38**	**55.08**	**56.98**	**83.90/86.43**	**83.85/86.43**	**52.80**	**79.80**

Table 6: Comparison with classical methods prior to 2021 on **Noise-Free-Data**. Our model significantly outperforms these early works, which are included in the Appendix to serve as a historical performance reference.

Method	MOSI						MOSEI					
	Acc-7	Acc-5	Acc-2	F1	MAE	Corr	Acc-7	Acc-5	Acc-2	F1	MAE	Corr
LMF	21.50	22.70	59.72/59.34	56.44/56.14	131.00	29.90	45.31	45.72	72.21/73.97	69.54/69.42	71.80	44.70
MulT	23.11	24.65	63.13/63.23	60.75/61.01	126.30	34.80	46.38	46.89	74.46/75.18	72.93/72.66	70.00	50.40
MISA	25.75	27.43	63.93/67.73	59.05/58.83	120.20	40.51	45.11	45.70	75.21/75.71	74.41/74.07	69.84	51.41
Self-MM	27.81	30.36	66.97/67.56	65.43/66.22	116.20	44.40	46.76	47.38	75.18/75.44	73.74/72.98	68.56	50.71
TFR-Net	27.56	30.53	67.68/67.89	65.75/66.18	115.60	45.20	46.91	47.32	74.74/74.22	72.55/73.45	68.90	51.10
ADRMSA	**33.98**	**37.10**	**70.49/70.61**	**70.97/71.39**	**110.35**	**52.45**	**50.72**	**51.26**	**77.81/78.16**	**77.39/77.56**	**61.89**	**59.00**

Table 7: Overall performance comparison with pre-2021 baselines on CMU-MOSI and CMU-MOSEI under the random missing modality setting. Results are obtained under missing rates from 0.0 to 0.9.

Method	Noise-Free Complete Data						Missing Modality (0.0–0.9 AUILC)					
	Acc-5	Acc-3	Acc-2	F1	MAE	Corr	Acc-5	Acc-3	Acc-2	F1	MAE	Corr
LMF	39.30	65.12	77.43	77.44	43.81	57.87	31.53	56.87	72.71	66.03	53.94	34.84
MulT	40.02	65.71	78.27	78.54	44.21	58.11	32.28	56.75	72.81	68.13	50.86	37.66
MISA	40.55	63.38	77.51	76.61	44.75	56.35	31.81	52.76	69.86	66.12	54.40	33.90
Self-MM	43.15	66.14	78.64	78.64	41.11	60.17	32.28	56.87	72.81	68.43	**50.85**	37.65
TFR-Net	41.23	64.25	78.08	78.14	43.71	58.31	26.52	52.89	68.13	58.71	66.41	36.92
ADRMSA	**43.45**	**68.04**	**80.66**	**80.98**	**38.95**	**62.13**	**33.18**	**54.40**	**71.09**	**70.92**	54.03	**42.17**

Table 8: Performance comparison with classical baseline methods (proposed before 2021) on CH-SIMS under both noise-free and random missing modality settings.