Sentiment Intensity-guided Conditional Diffusion Model for Incomplete Multimodal Sentiment Analysis

Anonymous ACL submission

Abstract

Missing modalities in multimodal sentiment analysis mean that one or more modalities are entirely unavailable. This paper proposes a Sentiment Intensity-guided Conditional Diffusion (SIGCD) model to effectively generate missing modalities for multimodal sentiment analysis. SIGCD consists of two modules: (1) The Multimodal Sentiment Intensity Quantification Module (MSIQM) quantifies the sentiment intensity of available modalities to generate sentiment intensity labels; (2) The Sentiment-Driven Diffusion Module (SDDM) generates missing modalities by incorporating sentiment intensity labels as a dynamic conditioning factor into the diffusion process. Experiments demonstrate that SIGCD achieves state-of-the-art performance results on public datasets, and the code will be released on GitHub.

1 Introduction

Multimodal Sentiment Analysis (MSA) aims to improve the accuracy and robustness of sentiment analysis by collaboratively modeling the complementary information from text, audio, and video modalities (Soleymani et al., 2017; Das and Singh, 2023). However, in real-world scenarios, multimodal data are often incomplete because an entire modality may be missing due to device failures, privacy constraints, or transmission losses (appendix A shows different missing-data scenarios) (Zhang et al., 2015; Chen and Zhang, 2020; Zhang et al., 2022; Choi and Lee, 2025). Effectively fusing the remaining modalities information and recognizing sentiment in the absence of certain modalities is one of the key challenges in MSA (Lai et al., 2023).

A mainstream approach to address missing modalities in MSA is the generative methods, which learn the distribution of data in a latent space to generate missing modalities similar to the distribution of available modalities. For instance,

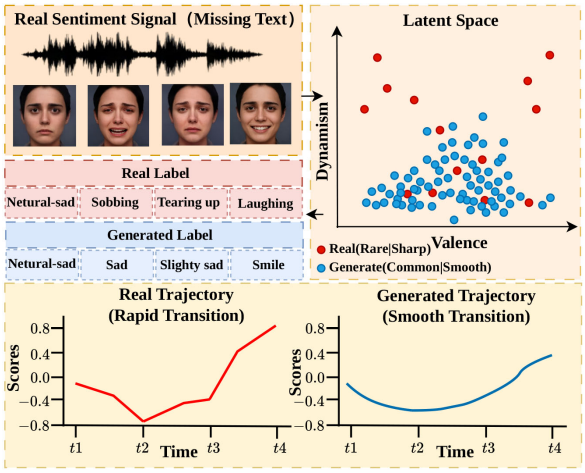

Figure 1: Over-smoothing Problem.

Du et al. (2018) model missing modalities as latent variables and use DNN-based Gaussian mappings to generate missing modalities from available modalities. Ma et al. (2021) reconstruct missing modalities by combining features from available modalities with features extracted from complete data via a learnable weighting network. However, distribution differences between modalities lead to sparse information in the latent space (Li et al., 2025), resulting in inadequate sentiment distribution modeling, causing the generated sentiment trajectories to be over-smoothing (Luo et al., 2021; Lu et al., 2024) (Figure 1 bottom right corner) and difficult to accurately capture the real sentiment trajectory (Figure 1 bottom left corner).

To mitigate the over-smoothing problem in generative methods, existing research has employed constrained strategies to enhance the authenticity of generated sentiment expression. For instance, Liu et al. (2025) employ multi-level cross-attention mechanisms for knowledge transfer and introduce consistency loss to ensure generated features closely match the distribution of real modality features. Zhang et al. (2025) enhance sentiment categories by weighting boundary samples,

reconstruct missing modalities, and reduce over-smoothing caused by ambiguous sentiment categories. Li et al. (2026b) addressed missing modality completion through a dynamic reliability constraint. However, the constraint signals of previous work are still primarily learned from the latent space, lacking fine-grained and quantifiable modeling, thereby hindering the model's ability to accurately learn and reconstruct the sentiment distribution of missing modalities.

To address the above issue, we propose the Sentiment Intensity-guided Conditional Diffusion (SIGCD) model, which uses explicit time-varying sentiment intensity signals to guide missing modality generation. Since sentiment information exhibits continuous and dynamic temporal variations (as shown in Fig 1), we employ a diffusion model with continuous-time modeling capabilities to generate missing modalities, enabling more accurate capture of temporal sentiment variations. SIGCD comprises two primary modules: (1) The Multimodal Sentiment Intensity Quantification Module (MSIQM) generates sentiment intensity labels with temporal information for available modalities and embeds them into the diffusion process. (2) The Sentiment-Driven Diffusion Module (SDDM) adopts a macro–micro dual-control mechanism. Sentiment intensity is used as a dynamic adjustment factor and integrated with available modality features to guide the diffusion process for missing modality generation. The macro level dynamically adjusts the noise schedule. The micro level performs cross-modal feature alignment. Through the collaboration of the MSIQM and SDDM modules, SIGCD can generate missing modality features that are closer to the real sentiment distribution. Finally, the generated modalities are fused with the available modalities for sentiment prediction.

We conducted experiments on three standard benchmarks under deterministic and stochastic missing scenarios, demonstrating that SIGCD outperforms mainstream models across various missing patterns. Additionally, SIGCD uses a parameter-sharing mechanism and simplified decoder structure, reducing computational resources for training and inference. The contributions are:

- We propose the Sentiment Intensity-guided Conditional Diffusion (SIGCD) model, which integrates sentiment intensity as a dynamic adjustment factor into the diffusion process to generate missing modalities.
- We introduce two key components in SIGCD: (1) a Multimodal Sentiment Intensity Quantification Module (MSIQM) that generates fine-grained sentiment intensity labels from available modalities to guide the diffusion process; (2) a Sentiment-Driven Diffusion Module (SDDM) featuring a macro-micro dual-control mechanism, which dynamically adjusts the noise schedule at the macro level and performs cross-modal feature alignment at the micro level.
- We conduct extensive experiments on the CMU-MOSI, CMU-MOSEI and CH-SIMS datasets under both deterministic and stochastic missing scenarios. Results demonstrate that SIGCD achieves state-of-the-art performance, demonstrating its strong robustness. A detailed experimental analysis is provided, and our code will be publicly released on GitHub.

2 Related Work

Missing-modality solutions in MSA mainly fall into two approaches: joint learning and generative. **Joint learning methods** learn joint feature spaces across modalities and construct latent representations of missing modalities from known modalities. Early approaches combined recurrent multi-stage fusion with LSTMs to capture temporal and cross-modal interactions (Liang et al., 2018). Subsequent works leveraged attention mechanisms and tensor fusion to model cross-modality dependencies (Hazarika et al., 2020). Recent advancements incorporate contrastive learning for text-vision alignment (Dao et al., 2025), utilize the Mamba architecture with multi-attention for feature generation and fusion (Liu et al., 2026), or adopt progressive alignment guided by modality reliability (Bao et al., 2026). **Generative methods** generate missing modalities by learning the latent distribution of data. Early studies utilize variational autoencoders (VAE) and generative adversarial networks (GAN) to generate missing modalities (Kingma and Welling, 2014; Goodfellow et al., 2014). With the development of self-supervised learning and diffusion models, Ho et al. (2020) proposes adding noise step by step and generating missing modalities during the denoising process. Guo et al. (2024a) uses prompt learning to generate missing modality features. Li et al. (2026b) addresses missing modalities through dynamic weighting and a dual-layer generative approach. Recent research focuses on modality het-

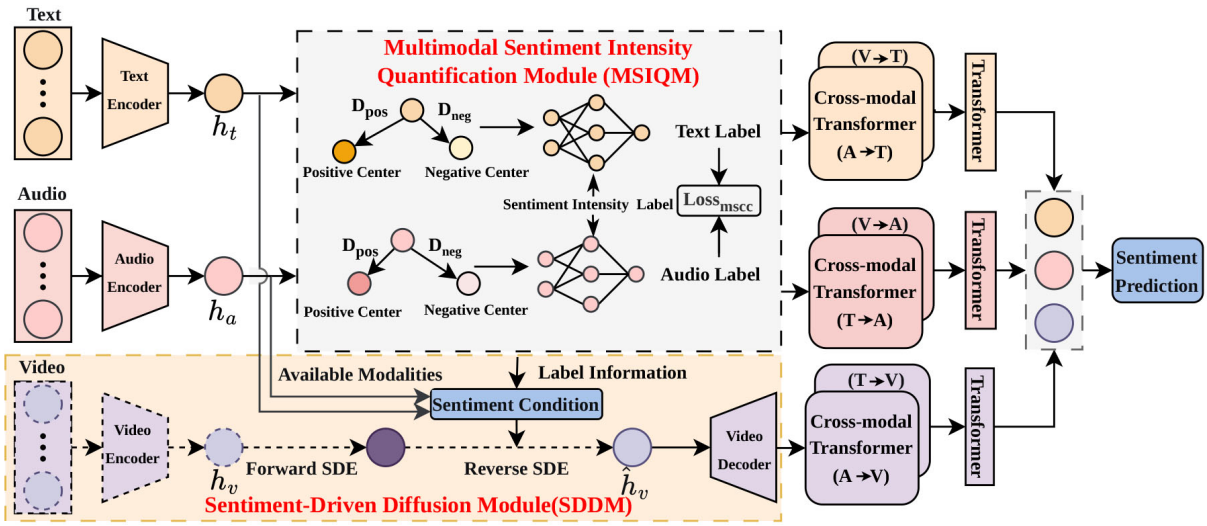

Figure 2: SIGCD architecture with video as the missing modality and it is randomly initialized.

erogeneity and noise issues. For instance, Zhuang et al. (2025) addresses missing and noisy modalities using a denoising-then-completion framework, utilizing modality heterogeneity and shared information to enhance robustness. Liu et al. (2025) reconstructs missing modalities via knowledge transfer and uses attention-based fusion to handle modality heterogeneity for robust sentiment prediction. *SIGCD belongs to the generative paradigm but differs from existing work. Specifically, it integrates sentiment intensity as a dynamic adjustment factor into the diffusion process and introduces a macro-micro dual-control mechanism, where sentiment intensity globally guides the noise schedule and locally fine-tunes feature alignment via cross-modal attention, thereby explicitly capturing the temporal dynamics of sentiment.*

3 Method

3.1 Problem Definition

In a multimodal task, each sample is represented as $X = \{x_1, \ldots, x_k, \ldots, x_M\}$, where x_k denotes the data of the k-th modality. We introduce a binary indicator vector $\alpha \in \{0,1\}^M$ to describe modality availability, where $\alpha_k = 1$ and $\alpha_k = 0$ indicate that the k-th modality is observed and missing, respectively. The corresponding available and missing modality sets are defined as $S_{\text{ava}} = \{k \mid \alpha_k = 1\}$ and $S_{\text{miss}} = \{k \mid \alpha_k = 0\}$. Assuming that all modalities follow a joint distribution $P(X)$, the missing-modality sentiment analysis task aims to infer missing modalities by estimating $P(x_m \mid x_a)$, where $m \in S_{\text{miss}}$ and $x_a = \{x_k\}_{k \in S_{\text{ava}}}$ denotes the observed modalities. The reconstructed complete multimodal feature is then used for sentiment classification.

3.2 Overall Framework

Figure 2 illustrates the architecture of the SIGCD model. Taking the absence of the video modality as an example, this section presents the workflow of the model. For the available text and audio modalities, feature representations are first extracted by encoders, and the Multimodal Sentiment Intensity Quantification Module (MSIQM) generates sentiment intensity labels and feature representations for each modality. MSIQM quantifies sentiment intensity by computing relative sentiment distance of sentiment center and generates corresponding labels, while an adaptive modulation mechanism is introduced to dynamically guide the feature generation process. Then, the encoded multimodal features and sentiment intensity labels are used as conditional inputs for the Sentiment-Driven Diffusion Module (SDDM). By modeling the reverse process of a conditional Stochastic Differential Equation (SDE), SDDM performs controlled feature diffusion and reconstruction in the latent space to generate the missing video modality. During the feature fusion stage, the model adopts a Transformer-based architecture and leverages attention mechanisms to facilitate multimodal feature interaction, followed by sentiment prediction.

3.3 Multimodal Sentiment Intensity Quantification Module

The Multimodal Sentiment Intensity Quantification Module (MSIQM) estimates unimodal sentiment scores by measuring relative distances to precomputed affective centers.

For textual feature extraction, we employ a BERT (Devlin et al., 2019) encoder to obtain the sentence representation $h_t = \text{BERT}(U_t) \in \mathbb{R}^{d_t}$. For audio and video modalities, a unidirectional Long Short-Term Memory network (LSTM) extracts temporal features, where the final hidden state yields the sequence representation. So $h_a = \text{sLSTM}(U_a) \in \mathbb{R}^{d_a}$, with h_v defined similarly.

During training, to enable MSIQM to learn affective information from all modalities as much as possible, the model is first fed with full-modality data. Two affective centers $\text{center}_m^{\text{pos}}$, $\text{center}_m^{\text{neg}}$ are computed for each modality based on its corresponding sentiment labels:

$$\text{center}_i = \frac{\sum_{j=1}^{N} h_i(j) \times \text{sgn}(y_i(j))}{\sum_{j=1}^{N} \text{sgn}(y_i(j))}, \quad (1)$$

where $i \in \{t, a, v\}$, $h_i(j)$ denotes the j-th sample of the i-th modality, and $\text{sgn}(\cdot)$ is the sign function. When $x > 0$, sentiment positive samples are selected to compute $\text{center}_m^{\text{pos}}$, and $\text{center}_m^{\text{neg}}$ is obtained similarly. The three hidden features are concatenated to form the fused representation $h_f = \text{concat}(h_t, h_a, h_v)$. Then, sentiment prediction for each modality is performed using Multi-layer Perceptrons (MLPs), producing the prediction $\hat{y}_m = \text{MLP}(h_m)$, where $m \in \{f, t, a, v\}$.

Based on the prediction $\hat{y}_m$, in the affective feature space, each modality's emotional representation is mapped to a distinct position. Due to disparities across modalities, absolute distances fail to capture inter-modal emotional relationships. We introduce relative distances: if a representation lies closer to the positive center and farther from the negative center, it tends to convey positive sentiment, and vice versa. This geometric method enhances cross-modal sentiment analysis. The distance to the positive center is defined as:

$$D_i^{\text{pos}} = \frac{\left\| h_i' - \text{center}_i^{\text{pos}} \right\|_2^2}{\sqrt{d_i}}, \quad (2)$$

similarly, the distance to the negative center D_i^{neg} is computed by replacing $\text{center}_i^{\text{pos}}$ with $\text{center}_i^{\text{neg}}$, where $i \in \{f, t, a, v\}$, d_i is the feature dimension of modality i, and h_i' is the hidden representation from the MLP. The relative sentiment intensity is calculated as

$$R_i = \frac{D_i^{\text{pos}} - D_i^{\text{neg}}}{D_i^{\text{pos}} + D_i^{\text{neg}} + \epsilon}, \quad (3)$$

where ϵ is a small positive constant to prevent division by zero.

Despite differences in sentiment expression across modalities, modalities from the same sample exhibit semantic coupling. Enlightened by the modal proportional correlation theory in Yu et al. (2021), we employ the affective center-based relative distance measurement and follow the proportional consistency assumption between modal sentiment labels and sentiment intensity for unimodal affective semantic modeling and reasoning. According to the above proportional assumption, we have $\frac{y_i}{y_f} \propto \frac{R_i}{R_f}$ and $y_i - y_f \propto R_i - R_f$, where $i \in \{t, a, v\}$. Combining these two proportional equations yields:

$$y_i = \frac{R_i + R_f}{2R_f} y_f + \frac{R_i - R_f}{2}. \quad (4)$$

During training, as the fused modality prediction $\hat{y}_f$ gradually fits the truth label y_f, the unimodal labels y_i derived from y_f via the proportional consistency assumption become correctly labeled, thereby generating unimodal sentiment labels. The loss function for MSIQM is:

$$d_m^j = \left| \widehat{y_m^j} - y_m^j \right|, \quad d_i^j = \left| \widehat{y_i^j} - y_i^j \right|, \quad (5)$$

$$L_{\text{MSCC}} = \frac{1}{N} \sum_{j}^{N} \left(d_m^j + \sum_{i}^{\{t,a,v\}} W_i^j \cdot d_i^j \right), \quad (6)$$

where N is the number of training samples, and $W_i^j = \tanh\left(y_i^j - y_m^j\right)$ represents the prediction loss weight for the j-th sample of the i-th modality. In missing-modality scenarios,, MSIQM not only enhances the model's ability to capture emotional information but also feeds the predicted unimodal sentiment label into the diffusion model to assist in generating missing modalities.

3.4 Sentiment-Driven Diffusion Module

Diffusion models operate through a dual-stage process: forward diffusion and reverse reconstruction(Ho et al., 2020). The Sentiment-Driven Diffusion Module (SDDM) couples these two processes: (1) The forward process iteratively injects Gaussian noise into the original data S_0 across time domain T via a Markov chain, gradually degrading it into isotropic Gaussian noise S_T. (2) The reverse process establishes a probability flow mapping from the noise space back to the data space via a learnable parameterized transition kernel. By estimating

the data distribution gradient under affective constraints using score matching, SDDM progressively removes noise to reconstruct high-quality samples. Its core objective is to recover the missing modality S_{miss} from the sampling space, conditioned on available modalities S_{ava} and their predicted sentiment scores.

Unlike discrete-time alternatives, SDDM adopts a continuous-time Stochastic Differential Equation (SDE) framework to precisely capture sentiment dynamics. The forward process is formalized as:

$$dS_{\text{miss}}(t) = f(S_{\text{miss}}(t), t)dt + g(t)dw(t), \quad (7)$$

where $f(\cdot)$ is the drift coefficient, $g(t)$ is the diffusion coefficient, and $w(t)$ is a standard Wiener process. We implement the Variance Preserving (VP) formulation, setting $f(S,t) = -\frac{1}{2}\beta(t)S$ and $g(t) = \sqrt{\beta(t)}$. As $t \to T$, $S_{\text{miss}}(T)$ converges to $\mathcal{N}(0, I)$, forming a Markovian continuous perturbation on the data manifold. To recover $S_{\text{miss}}(T)$, the reverse dynamics are derived based on Anderson's reverse-time SDE theory:

$$s(t) = \nabla_s \log p_t\big(S_{\text{miss}}(t) \mid S_{\text{ava}}, \hat{y}_m\big), \quad (8)$$

$$dS_{\text{miss}} = \Big[f_t(S_{\text{miss}}) - g_t^2 s(t)\Big]dt + g_t d\bar{w}, \quad (9)$$

where $\nabla_s \log p_t(\cdot)$ is the conditional score function guiding samples toward high-probability regions while embedding affective semantics (detailed in Appendix B). To parameterize and estimate this score function $s(t)$, we propose a **macro-micro dual-control architecture** as the foundational paradigm for our conditional score network $\text{score}_\theta(S_{\text{miss}}(t), t, S_{\text{ava}}, \hat{y}_m)$. Rather than deploying standard feature-fusion neural blocks, this architecture explicitly decouples the continuous affective guidance into two coordinated granularities to govern the reverse generation:

- **Macro-level Global Trajectory Guidance:** At the macro level, the model modulates the continuous-time trajectory by coupling temporal scheduling with global affective states. In practice, each residual block of the score network is conditioned on time embedding $e_t \in \mathbb{R}^d$ and sentiment embedding $e_y \in \mathbb{R}^d$ via the Dynamic Affect-Temporal Manifold Gate (**DATMGate**) (Appendix C). This adaptive gating enforces tighter semantic constraints under high sentiment intensity, while preserving structural diversity under low intensity.
- **Micro-level Cross-Modal Feature Alignment:** At the micro level, to mitigate modality heterogeneity and local semantic noise during trajectory evolution, the Sentiment-Guided Semantic Focusing Attention (**SG-SFA**) (Appendix D) is integrated within the network layers. It enforces local semantic alignment through sentiment-aware attention guidance, matching generated fine-grained details with available emotional cues.

Structurally, the network implements a multi-scale U-Net backbone with skip connections to carry out this dual-control design while effectively preserving multi-scale features.

The network is optimized by minimizing the weighted score matching loss:

$$s_\theta(t) = \text{score}_\theta(S_{\text{miss}}(t), t \mid S_{\text{ava}}, \hat{y}_m), \quad (10)$$

$$L_{\text{sm}} = \mathbb{E}_{t,\xi}\left[\lambda(t)\left\|s_\theta(t) + \frac{\xi}{\sigma(t)}\right\|_2^2\right], \quad (11)$$

where $\xi \sim \mathcal{N}(0, I)$, and $\lambda(t) = \sigma(t)^2$ provides time-adaptive regularization. Near $t \to 0$, $\lambda(t)$ suppresses updates to avoid model oscillation from initial noise estimation errors; as $t \to T$, $\lambda(t)$ increases to fully capture high-noise distribution features. During training, random modality masking forces global affective inference from incomplete inputs. Missing modalities are generated by inversely solving the SDE via an efficient Predictor-Corrector algorithm (Appendix E).

Finally, the recovered data $\hat{x}_{\text{miss}}$ is passed through a linear layer decoder to extract affective representations. The decoder is optimized using MSE loss $\mathcal{L}_{\text{rec}} = \sum_{k \in S_{\text{miss}}} \|\hat{x}_k - x_k\|_2^2$ against complete data. The total SDDM objective is:

$$\mathcal{L}_{\text{SDDM}} = \mathcal{L}_{\text{sm}} + \mathcal{L}_{\text{rec}}. \quad (12)$$

3.5 Modality Fusion and Sentiment Prediction

After generating the missing modality, multimodal features are fused for sentiment intensity prediction. First, they are projected into a unified feature space via projection layers $\tilde{H}_m = w_m H_m + b_m$, where $m \in \{t, a, v\}$, and H_m denotes the representation of either the available modality or the generated missing modality. Then, using each modality as an interaction center, a Transformer encoder is employed to obtain cross-modal interaction features. Taking text as an example, the text modality serves as the Query, while the audio and video

modalities serve as the Key and Value, capturing semantic associations between linguistic descriptions and audio-visual features. The interaction results are formulated as $H_t^a = \text{Transformer}(\tilde{H}_t, \tilde{H}_a, \tilde{H}_a)$ and $H_t^v = \text{Transformer}(\tilde{H}_t, \tilde{H}_v, \tilde{H}_v)$.

To further strengthen temporal dependencies and global semantic expression across modalities, a deep self-attention network is introduced, where the final representation is computed as $H_t^* = \text{Transformer}(H_t^a \oplus H_t^v)$, where $\oplus$ denotes concatenation along the feature dimension, and H_t^* is the final state capturing the enhanced modality memory features, serving as the global representation. Similarly, enhanced representations for the audio and video modalities H_a^*, H_v^* can be obtained.

Let $H_c = H_t^* \oplus H_a^* \oplus H_v^*$. The fused feature H_{fusion} is derived via a co-projection layer and residual connection as $H_{\text{fusion}} = \phi_2(\phi_1(H_c)) + H_c$, where ϕ_1, ϕ_2 are fully connected layers with ReLU and Dropout. The final sentiment intensity prediction is computed as (with w_{out} and b_{out} being learnable parameters):

$$\hat{y}_{\text{fusion}} = w_{\text{out}} H_{\text{fusion}} + b_{\text{out}} \tag{13}$$

The Mean Squared Error (MSE) loss is used for $\hat{y}_{\text{fusion}}$ as the task loss.

The overall loss function for the SIGCD model is:

$$\mathcal{L} = L_{\text{task}} + \mathcal{A} L_{\text{MSIQM}} + \mathcal{B} L_{\text{SDDM}}, \tag{14}$$

where $\mathcal{A}$ and $\mathcal{B}$ are loss weighting coefficients that regulate the influence of each module on the model.

4 Experiments

4.1 Datasets and Implementation Details

We performed a comprehensive experimental analysis on three widely used public datasets MOSI (Zadeh et al., 2016), MOSEI (Bagher Zadeh et al., 2018) and CH-SIMS (Yu et al., 2020). The details of the datasets are shown in Appendix F.

SIGCD reduces computational complexity and memory consumption through parameter sharing and a lightweight architecture. Specifically, within the U-Net-based score network, symmetric diffusion modules employ weight sharing to reduce redundant parameters, while the decoder adopts a lightweight linear projection layer instead of a complex network structure, thereby lowering both parameter size and computational cost while maintaining sentiment generation performance. Under this design, experiments conducted on an NVIDIA GB10 GPU with CUDA 13.0 show that the model requires only approximately 2704 MiB of GPU memory during training and inference, demonstrating its efficiency and low resource consumption.

To evaluate model robustness under missing scenarios, we consider two experimental settings. First is the Deterministic Missing Scenario (DMS), where the missing modalities are fixed across all samples. Six missing-modality settings are considered: $\{t\}$, $\{a\}$, $\{v\}$, $\{t, a\}$, $\{t, v\}$, and $\{a, v\}$. The second is the Stochastic Missing Scenario (SMS), where the missing modalities vary randomly across samples, with each sample missing one or two modalities. Following Lian et al. (2023), the Missing Rate (MR) is used to quantify the degree of random missingness, with MR ranging from 0.1 to 0.7. More details of MR are provided in Appendix G.

4.2 Baselines

To evaluate the performance of SIGCD, we compare it with state-of-the-art baselines under two modality missing scenarios on the MOSI, MOSEI, and CH-SIMS datasets. Details of all baselines are provided in Appendix H. Only recent methods are reported in the main text, while additional baseline results are provided in Appendix J. We use Accuracy (ACC) and F1-score (F1) for evaluation. Results are presented in Tables 1–3.

4.3 Results and Discussion

In the DMS scenario, as shown in Table 5, SIGCD achieves superior results across all six modality combinations, especially under text-missing conditions. On MOSI and MOSEI, it consistently outperforms existing methods in $\{a\}$, $\{v\}$, and $\{a, v\}$ settings, achieving the best average ACC/F1 of 74.70%/74.60% and 75.20%/75.30%, respectively. To further verify generalization and robustness, we evaluate it on CH-SIMS, where it obtains 81.27%/81.22% ACC/F1. This improvement is attributed to the proposed macro–micro dual-control mechanism, which integrates sentiment intensity to reconstruct missing modalities while preserving fine-grained emotional cues, alleviating over-smoothing and noise sensitivity.

In the SMS scenario, model performance gradually decreases as the missing rate increases from MR = 0.1 (mild missing) to MR = 0.7 (severe missing), indicating that random modality missing weakens cross-modal semantic complementarity

Dataset	Model	{t} ACC/F1	{a} ACC/F1	{v} ACC/F1	{t,a} ACC/F1	{t,v} ACC/F1	{a,v} ACC/F1	**Avg.** ACC/F1
MOSI	DERL	84.18/-	50.52/-	50.52/-	84.18/-	84.32/-	50.52/-	67.40/-
	MPLMM	83.80/83.80	54.60/42.00	54.70/42.10	83.70/83.60	83.80/83.70	57.30/42.60	69.70/63.00
	UniMF	85.30/85.20	61.60/59.30	61.70/61.60	85.40/85.20	85.60/85.40	62.80/62.70	73.70/73.20
	MulDeF	85.20/85.20	57.90/57.50	56.70/53.50	85.50/85.40	85.50/85.40	57.90/57.70	71.50/70.80
	IIE	85.50/85.30	62.50/61.50	62.50/62.30	85.80/85.60	85.80/85.80	65.20/64.20	74.60/74.10
	SIGCD(Ours)	85.70/85.60	62.30/62.10	62.40/62.50	86.10/86.10	86.20/86.00	65.50/65.20	**74.70/74.60**
MOSEI	MPLMM	84.60/84.60	61.80/49.40	62.20/53.20	85.10/85.10	85.20/85.10	63.20/51.00	73.70/68.10
	UniMF	84.80/84.80	62.40/55.70	62.10/58.10	85.10/85.20	85.50/85.30	63.40/60.10	73.90/71.50
	MulDeF	85.20/85.10	62.80/60.00	62.40/60.40	85.40/85.20	85.30/85.20	62.30/60.90	73.90/72.80
	IIE	85.50/85.30	61.90/57.80	65.20/63.40	85.70/85.60	85.90/85.90	65.20/63.70	74.90/73.60
	SIGCD(Ours)	85.70/85.40	62.20/63.50	64.80/64.70	86.20/86.00	86.10/86.40	66.40/65.90	**75.20/75.30**

Table 1: Results under DMS (%) on MOSI and MOSEI.

Dataset	Model	0.1 ACC/F1	0.2 ACC/F1	0.3 ACC/F1	0.4 ACC/F1	0.5 ACC/F1	0.6 ACC/F1	0.7 ACC/F1	**Avg.** ACC/F1
MOSI	GCNet	82.30/82.30	79.40/79.50	77.20/77.20	74.30/74.40	70.00/69.80	67.70/66.70	65.70/65.40	73.80/73.60
	MPLMM	80.60/80.70	77.70/77.90	75.00/75.10	70.00/69.80	67.40/66.60	63.40/61.80	59.20/56.20	70.50/69.70
	LNLN	81.40/81.40	78.70/78.70	74.70/74.80	70.30/69.90	67.10/66.30	62.50/60.50	61.00/58.30	70.80/70.00
	UniMF	83.70/83.40	80.80/80.80	79.10/79.10	75.80/75.30	72.90/72.60	71.70/71.30	70.30/68.70	76.30/75.90
	MulDeF	81.70/81.80	79.70/78.90	75.00/75.10	73.20/70.30	67.20/66.10	63.10/61.30	60.70/58.30	71.50/70.30
	IIE	85.10/84.90	82.30/82.30	81.70/81.70	79.30/78.80	77.00/76.50	73.60/72.90	73.20/72.20	**78.90**/78.50
	SIGCD(Ours)	84.50/84.50	83.40/83.40	81.10/81.10	79.40/79.30	76.20/75.30	74.10/73.30	73.90/73.90	**78.90/78.70**
MOSEI	GCNet	82.30/82.10	80.30/79.90	77.50/76.80	76.00/74.90	74.90/73.20	74.10/72.10	73.20/70.40	76.90/75.60
	MPLMM	83.70/83.50	81.40/80.80	79.20/78.30	77.10/76.10	75.60/73.80	72.70/70.60	70.50/69.20	77.20/76.00
	LNLN	83.10/83.00	81.70/81.00	79.10/78.10	76.90/75.60	75.10/72.40	72.50/68.60	70.60/65.20	77.00/74.80
	MulDeF	82.90/82.80	80.80/80.40	78.10/77.10	76.50/74.60	74.40/71.90	71.80/68.00	71.00/66.20	76.50/74.40
	IIE	84.20/83.80	82.20/81.60	79.60/78.90	77.40/76.30	75.70/74.50	71.60/70.00	71.60/70.00	77.50/76.40
	SIGCD(Ours)	83.80/84.00	82.60/81.80	80.50/79.20	78.30/78.10	76.10/75.50	74.20/73.40	71.40/69.10	**78.10/77.30**

Table 2: Results under SMS (%) on MOSI and MOSEI.

DMS	PRLF ACC/F1	MPLMM ACC/F1	MUMP ACC/F1	**SIGCD(Ours)** ACC/F1
{t}	80.28/-	79.75/78.74	81.82/80.99	80.42/79.72
{a}	81.74/-	65.93/77.10	69.51/82.01	81.94/82.38
{v}	81.63/-	66.02/78.86	69.73/82.79	82.47/82.86
{t,a}	80.57/-	77.45/77.84	80.71/80.36	79.40/79.89
{t,v}	80.36/-	77.97/77.95	80.27/80.48	81.44/80.83
{a,v}	82.58/-	65.28/74.02	68.73/78.54	81.96/81.61
Avg.	81.19/-	72.07/77.42	75.13/80.86	**81.27/81.22**

Table 3: Results under DMS (%) on CH-SIMS

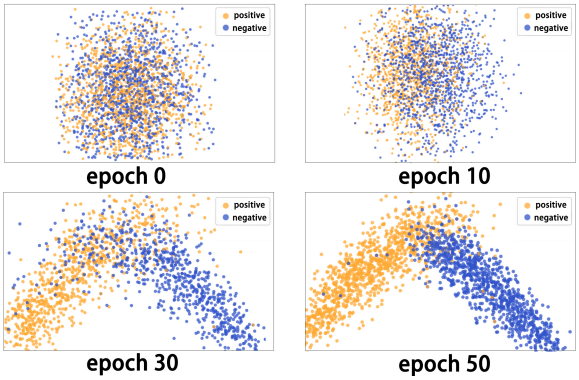

Figure 3: t-SNE of MOSI features at different epochs.

and increases the difficulty of sentiment recognition. Compared with existing methods, SIGCD still achieves good ACC/F1 results on MOSI and MOSEI under high missing rates. For example, on MOSI with MR = 0.7, SIGCD reaches 73.9/73.9. SIGCD achieves the best average performance on both datasets. These results show that SIGCD can effectively reduce semantic degradation caused by random missing modalities and has strong robustness and generalization in missing modality reconstruction and sentiment preservation.

Overall, DMS mainly evaluates a model's ability to exploit complementary information from specific modality combinations, whereas SMS focuses on its robustness to uncertain modality absence. By leveraging sentiment intensity guidance and conditional diffusion, SIGCD effectively addresses both challenges and achieves superior performance on the datasets.

4.4 Visualization

Figure 3 presents t-SNE visualizations of MOSI feature embeddings at epochs 0, 10, 30, and 50. Initially, positive (yellow) and negative (blue) samples are heavily mixed. With increasing training iterations, the two sentiment classes gradually separate, forming increasingly compact and distinct clusters. After 50 epochs, the features exhibit clear bimodal separation with minimal inter-class over-

Scenario	Dataset	SIGCD	w/o MSIQM	w/o SDDM
DMS	MOSI	74.7	73.2	72.8
	MOSEI	75.2	73.0	72.5
SMS	MOSI	78.9	75.5	74.8
	MOSEI	78.1	76.2	75.3

Table 4: Ablation Results of DMS and SMS (%).

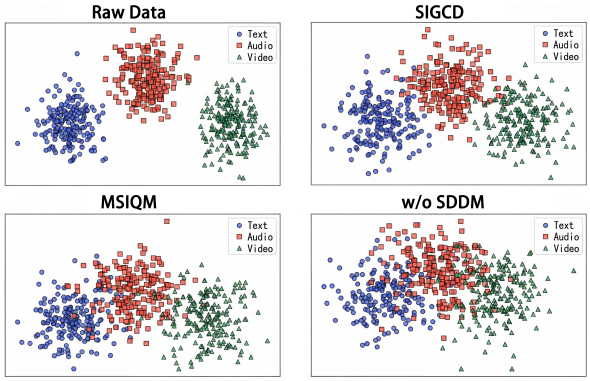

Figure 4: Data Distribution with Different Settings.

lap, confirming that the model effectively learns sentiment-discriminative representations.

4.5 Ablation Experiment

The previous main experiments demonstrated SIGCD's overall performance on missing modality data. This ablation study validates the contributions of its individual modules (MSIQM and SDDM). Results are shown in Table 4.

Results confirm MSIQM's sentiment labels are crucial for fine-grained calibration. SDDM refines generation via sentiment-guided diffusion. Removing either module degrades data alignment. Under random missing, MSIQM calibrates sentiment from available modalities, while SDDM maintains noise control. Without SDDM, stability and consistency suffer, particularly at high missing rates.

We use feature-space visualization to compare sample distributions across modalities and illustrate the impact of different modules on data generation(Figure 4). The results show that the full SIGCD closely matches the original data distribution, whereas models without MSIQM or SDDM exhibit clear deviations in both local and global structures. These results confirm the importance of each core module for multimodal data generation.

4.6 Parameter Analysis Experiment

To study MSIQM and SDDM, we perform parameter analysis on their loss coefficients. By varying the loss weights of MSIQM and key coefficients of SDDM, we evaluate their impact on sentiment recognition on CMU-MOSI and CMU-MOSEI and

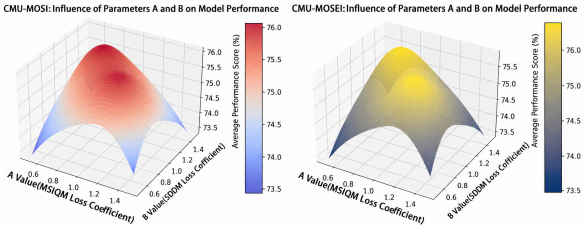

Figure 5: Analysis of hyper-parameters $\mathcal{A}$ and $\mathcal{B}$.

determine the optimal configuration. The results are shown in Figure 5.

In MSIQM, the coefficient $\mathcal{A}$ balances the contribution of the unimodal multi-task structure and sentiment intensity labels in the overall loss. A larger $\mathcal{A}$ improves cross-modal sentiment calibration and alignment under missing-modality settings, but an overly large value may overemphasize local information and weaken global integration. The results show that $\mathcal{A} = 1.0$ achieves the best trade-off and performance.

In SDDM, the coefficient $\mathcal{B}$ controls the balance between global noise scheduling and local feature alignment in sentiment-guided diffusion. A small $\mathcal{B}$ provides insufficient sentiment guidance and increases noise sensitivity, while an overly large value overemphasizes sentiment labels and reduces generation diversity, especially under ambiguous sentiments. The results show that $\mathcal{B} = 1.0$ achieves the best balance and performance.

5 Conclusion

This paper addresses the modality missing problem in multimodal sentiment analysis by proposing a Sentiment Intensity-guided Conditional Diffusion (SIGCD) model. It introduces a Multimodal Sentiment Intensity Quantification Module (MSIQM) and a Sentiment-Driven Diffusion Module (SDDM), achieving a "macro–micro" dual control of local sentiment details and global noise scheduling. This enables effective use of available modalities to generate high-quality missing data. Experimental results show that SIGCD achieves improved performance on the CMU-MOSI, CMU-MOSEI, and CH-SIMS datasets, with strong robustness under both fixed and random missing scenarios. Ablation studies and parameter analysis further validate the roles of MSIQM and SDDM in restoring data distribution and refining the generation process. Future work will explore more diverse missing patterns to improve generalization.

Limitations

Although the SIGCD achieves state-of-the-art performance on multimodal sentiment analysis, it has only been evaluated on limited datasets and tasks. Its generalization to other multimodal tasks, such as emotion recognition, sarcasm detection, rumor detection, and visual question answering, remains unclear. Future work will evaluate the proposed method on a wider range of tasks and datasets to further verify its effectiveness and robustness.

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.

Jindi Bao, Jianjun Qian, Mengkai Yan, and Jian Yang. 2026. Progressive representation learning for multimodal sentiment analysis with incomplete modalities. *Preprint*, arXiv:2603.09111.

Jiayi Chen and Aidong Zhang. 2020. HGMF: heterogeneous graph-based fusion for multimodal data with incompleteness. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 1295–1305. ACM.

Suk Hoon Choi and Jong Min Lee. 2025. Multimodal learning with missing modality for chemical process system. *Comput. Chem. Eng.*, 201:109196.

Phuong Q. Dao, Mark Roantree, and Vuong M. Ngo. 2025. An enhanced dual transformer contrastive network for multimodal sentiment analysis. *CoRR*, abs/2510.23617.

Ringki Das and Thoudam Doren Singh. 2023. Multimodal sentiment analysis: A survey of methods, trends, and challenges. *ACM Comput. Surv.*, 55(13s):270:1–270:38.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Changde Du, Changying Du, Hao Wang, Jinpeng Li, Wei-Long Zheng, Bao-Liang Lu, and Huiguang He. 2018. Semi-supervised deep generative modelling of incomplete multi-modality emotional data. In *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*, pages 108–116. ACM.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial networks. *Preprint*, arXiv:1406.2661.

Zirun Guo, Tao Jin, and Zhou Zhao. 2024a. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 1726–1736. Association for Computational Linguistics.

Zirun Guo, Tao Jin, and Zhou Zhao. 2024b. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1726–1736, Bangkok, Thailand. Association for Computational Linguistics.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. MISA: modality-invariant and -specific representations for multimodal sentiment analysis. *CoRR*, abs/2005.03545.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Preprint*, arXiv:2006.11239.

Ruohong Huan, Guowei Zhong, Peng Chen, and Ronghua Liang. 2024. Unimf: A unified multimodal framework for multimodal sentiment analysis in missing modalities and unaligned multimodal sequences. *IEEE Transactions on Multimedia*, 26:5753–5768.

Ruohong Huan, Guowei Zhong, Peng Chen, and Ronghua Liang. 2025. Muldef: A model-agnostic debiasing framework for robust multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 27:2304–2319.

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*.

Songning Lai, Xifeng Hu, Haoxuan Xu, Zhaoxia Ren, and Zhi Liu. 2023. Multimodal sentiment analysis: A survey. *Displays*, 80:102563.

Xiang Li, Xiaoming Zhang, Dezhuang Miao, Xianfu Cheng, Dawei Li, Honggui Han, and Zhoujun Li. 2026a. Mixture of disentangled experts with missing modalities for robust multimodal sentiment analysis. *Preprint*, arXiv:2602.01833.

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

Zuhe Li, Zhenwei Huang, Xiaojiang He, Jun Yu, Haoran Chen, Chenguang Yang, and Yushan Pan. 2025. Representation distribution matching and dynamic routing interaction for multimodal sentiment analysis. *Knowledge-Based Systems*, 316:113376.

Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. 2023. Gcnet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8419–8432.

Paul Pu Liang, Ziyin Liu, Amir Zadeh, and Louis-Philippe Morency. 2018. Multimodal language analysis with recurrent multistage fusion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 150–161. Association for Computational Linguistics.

Siyuan Liu, Hongkun Zhao, Yang Chen, Fanmin Kong, and Kang Li. 2026. Dcgrm-net: Dual-channel guided reconstruction mamba network for robust multimodal sentiment analysis. *Inf. Process. Manag.*, 63(3):104491.

Weide Liu, Huijing Zhan, Hao Chen, and Fengmao Lv. 2025. Multimodal sentiment analysis with missing modality: A knowledge-transfer approach. *Preprint*, arXiv:2401.10747.

Nannan Lu, Zhen Tan, and Jiansheng Qian. 2024. Mrsln: A multimodal residual speaker-lstm network to alleviate the over-smoothing issue for emotion recognition in conversation. *Neurocomputing*, 580:127467.

Huaishao Luo, Lei Ji, Yanyong Huang, Bin Wang, Shenggong Ji, and Tianrui Li. 2021. Scalevlad: Improving multimodal sentiment analysis via multi-scale fusion of locally descriptors. *Preprint*, arXiv:2112.01368.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. 2021. SMIL: multimodal learning with severely missing modality. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 2302–2310. AAAI Press.

Mohammad Soleymani, David García, Brendan Jou, Björn W. Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image Vis. Comput.*, 65:3–14.

Jiajia Tang, Kang Li, Xuanyu Jin, Andrzej Cichocki, Qibin Zhao, and Wanzeng Kong. 2021. CTFN: Hierarchical learning for multimodal sentiment analysis using coupled-translation fusion network. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5301–5311, Online. Association for Computational Linguistics.

Yuanzhi Wang, Zhen Cui, and Yong Li. 2023. Distribution-consistent modal recovering for incomplete multimodal learning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21968–21977.

Zhongliang Wei, Ruofan Chen, and Jing Sun. 2025. Msaf-cf: A multimodal sentiment analysis framework based on feature enhancement and cross-fusion. *IEEE Access*, 13:157011–157025.

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.

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.

Jiandian Zeng, Tianyi Liu, and Jiantao Zhou. 2022. Tag-assisted multimodal sentiment analysis under uncertain missing modalities. *Preprint*, arXiv:2204.13707.

Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao. 2022. M3care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 2418–2428, New York, NY, USA. Association for Computing Machinery.

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.

Kangkang Zhang, Ruben Gonzalez, Biao Huang, and Guoli Ji. 2015. Expectation-maximization approach to fault diagnosis with missing data. *IEEE Trans. Ind. Electron.*, 62(2):1231–1240.

Yuqing Zhang, Dongliang Xie, Dawei Luo, and Baosheng Sun. 2025. Emotional boundaries and intensity aware model for incomplete multimodal sentiment analysis. *Digit. Signal Process.*, 160:105023.

Jinming Zhao, Ruichen Li, and Qin Jin. 2021. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2608–2618, Online. Association for Computational Linguistics.

Yan Zhuang, Minhao Liu, Yanru Zhang, Jiawen Deng, and Fuji Ren. 2025. TMDC: A two-stage modality denoising and complementation framework for multimodal sentiment analysis with missing and noisy modalities. *CoRR*, abs/2511.10325.

Yan Zhuang, Yanru Zhang, Jiawen Deng, and Fuji Ren. 2026. Intra-sample and intra-modal enhancement for multimodal sentiment analysis with missing modalities. *IEEE Transactions on Multimedia*, 28:1847–1859.

A Missing Modality Scenarios

Figure 6: Illustration of Missing Modality Scenarios in Multimodal Sentiment Analysis

As shown in Fig. 6, we consider two typical missing-data scenarios in MSA.

Modality-level overall missing. An entire modality (e.g., audio) is completely absent while others remain available, typically due to sensor failure or recording errors.

Random modality feature missing. Fine-grained missingness occurs within each modality, such as randomly dropped video frames or intermittent gaps in audio signals.

B Conditional Score Function

To jointly model affective and semantic information during reverse diffusion, the conditional score function is defined as

$$\nabla_s \log p_t(\cdot) = \mathbb{E}_\epsilon \left[\frac{\epsilon - S_{\text{miss}}(t)}{\sigma(t)^2} \,\middle|\, S_{\text{ava}}, \widehat{y}_m \right], \tag{15}$$

where

$$\sigma(t) = \sqrt{\int_0^t \beta(s)\, ds}. \tag{16}$$

The score function estimates the gradient of the conditional data distribution with respect to the generated state. Under the conditions of the available modalities S_{ava} and sentiment intensity $\widehat{y}_m$, it provides guidance toward high-density regions of the target distribution and enables affective-semantic control during generation.

C Dynamic Affect-Temporal Manifold Gate (DAT-MGate)

At the initial stage of the reverse process, the Dynamic Affect-Temporal Manifold Gate (DAT-MGate) is introduced to dynamically integrate temporal and affective information. It conditions each residual block on the time-step embedding $e_t \in \mathbb{R}^d$ and the sentiment intensity embedding $e_y \in \mathbb{R}^d$.

The gating mechanism is defined as:

$$g = \sigma\big(W_g[e_t; e_y] + b_g\big), \tag{17}$$

$$e_{\text{fusion}} = g \odot e_t + (1 - g) \odot e_y. \tag{18}$$

where σ denotes the Sigmoid function, and $g \in (0, 1)$ is a learned gating coefficient. This mechanism adaptively balances temporal guidance and affective guidance. When sentiment intensity is high, it strengthens semantic constraints; when sentiment intensity is low, it relies more on temporal scheduling to preserve diversity.

D Sentiment-Guided Semantic Focusing Attention (SG-SFA)

To reduce noise induced by modality gaps, we propose Sentiment-Guided Semantic Focusing Attention (SG-SFA). It consists of two complementary sub-attention mechanisms that jointly integrate affective and semantic information.

Let $V_1 = W_{v1} h_{\text{ava}}$ and $V_2 = W_{v2} h'_{\text{ava}}$ denote the value vectors of the two attention branches. The first branch, Sentiment-Guided Attention (SGA), is defined as:

$$\text{SGA} = \text{softmax}\left(\frac{W_{Q1} e_y \cdot W_{k1} (h_{\text{ava}})^\top}{\sqrt{d}} \right) V_1, \tag{19}$$

where the sentiment intensity embedding e_y serves as the query to guide feature selection.

The second branch, Semantic Attention (SA), is defined as:

$$\mathrm{SA} = \mathrm{softmax}\left(\frac{W_{Q2}h_{\mathrm{miss}} \cdot W_{k2}(h'_{\mathrm{ava}})^{\top}}{\sqrt{d}}\right) V_2. \tag{20}$$

SGA focuses on sentiment-aware filtering of available modalities, producing h'_{ava} that is highly correlated with affective cues. SA further injects semantically enriched representations into the generation process, enhancing cross-modal fusion and robustness.

E Predictor-Corrector algorithm Attention

The specific content of the formula is as follows:

$$f_\theta(S_t, t) = -\tfrac{1}{2}\beta(t)S_t - \beta(t)\,\mathrm{score}_\theta\big(S_t, t, S_{\mathrm{ava}}, \hat{y}_m\big), \tag{21}$$

$$S_{t-\Delta t} = S_t + f_\theta(S_t, t)\,\Delta t + \sqrt{\beta(t)\Delta t}\,z_t, \tag{22}$$

$$g(S) = \nabla_s \log p_{t-\Delta t}\big(S \mid S_{\mathrm{ava}}, \hat{y}_m\big), \tag{23}$$

$$S_{t-\Delta t} \leftarrow S_{t-\Delta t} + \eta\, g(S_{t-\Delta t}) + \sqrt{2\eta}\,\psi. \tag{24}$$

Based on the current state S_t and the score estimate score_θ, the Euler-Maruyama method approximates the solution of the reverse SDE. Here, $z_t \sim \mathcal{N}(0, I)$, Δt is the time step length. This step provides an initial estimate of the generation direction by balancing deterministic drift and stochastic diffusion. The variance of the stochastic term $\sqrt{\beta(t)\Delta t}z_t$ adapts with $\beta(t)$. When sentiment intensity is low, the $\beta(t)$ schedule delays noise attenuation, preserving more possibilities in the generation path. Subsequently, Langevin dynamics are used for local optimization of the prediction: η is the correction step size, $\psi \sim \mathcal{N}(0, I)$. Gradient ascent refines generation details, such as adjusting subtle arcs in facial muscle movement or the precise location of speech formants.

F Datasets

The experiments are conducted on three benchmark datasets for multimodal sentiment analysis: CMU-MOSI (Zadeh et al., 2016), CMU-MOSEI (Bagher Zadeh et al., 2018), and CH-SIMS (Yu et al., 2020).

33**CMU-MOSI** consists of 2,199 segmented utterances extracted from 93 YouTube movie review videos involving 89 speakers. Following the standard split, 1,284 samples are used for training, 229 for validation, and 686 for testing.

CMU-MOSEI is a larger-scale dataset containing approximately 20,000 utterances collected from 3,228 videos spanning 250 topics and over 1,000 speakers. The dataset is divided into 16,326 training, 1,871 validation, and 4,659 test samples.

CH-SIMS contains 2,281 video clips collected from real-world scenarios, including movies, TV series, and talk shows. Following the standard split, 1,368 samples are used for training, 456 for validation, and 457 for testing.

G Missing Rate Definition

For the Stochastic Missing Scenario (SMS), the Missing Rate (MR) is defined as

$$\mathrm{MR} = 1 - \frac{\sum_{i=1}^{N} m_i}{N \times M}, \tag{25}$$

where N denotes the total number of samples, M is the number of modality types ($M = 3$ in this work), and m_i denotes the number of available modalities in the i-th sample. Since each sample contains at least one available modality ($1 \le m_i \le M$), the maximum missing rate satisfies

$$\mathrm{MR} \le \frac{M-1}{M}.$$

In our experiments, MR varies from 0.1 to 0.7.

H Baseline Methods

To evaluate the effectiveness of SIGCD, we compare it with the following representative multimodal sentiment analysis methods.

CTFN (Tang et al., 2021): a coupled translation fusion network that models bidirectional interactions through coupled learning.

MMIN (Zhao et al., 2021): a cross-modal imagination network based on CRA. TATE (Zeng et al., 2022): a label-assisted Transformer encoder network.

MSAF-CF (Wei et al., 2025): a framework that integrates multi-source information and enhances discriminative features to accurately determine sentiment orientation.

MPLMM (Guo et al., 2024b): a multimodal prompt learning method for missing modalities.

GCNet (Lian et al., 2023): a graph completion network for incomplete multimodal learning in conversations.

DicMoR (Wang et al., 2023): a distribution-consistent modal recovery method for incomplete multimodal learning.

UniMF (Huan et al., 2024): a unified multimodal framework for multimodal sentiment analysis that handles missing modalities and unaligned multimodal sequences.

MulDeF (Huan et al., 2025): a model-agnostic debiasing framework for robust multimodal sentiment analysis.

IIE (Zhuang et al., 2026): an intra-sample and intra-modal enhancement framework for multimodal sentiment analysis with missing modalities.

DERL (Li et al., 2026a): a disentangled expert representation learning framework for robust multimodal sentiment analysis with missing modalities.

LNLN (Zhang et al., 2024): a robust multimodal missing signal framework for multimodal sentiment analysis under uncertain signal missing.

I Variance Preserving (VP) Diffusion

The Variance Preserving (VP) formulation is a classical diffusion process design that preserves the variance of the data during forward corruption. It ensures a controlled noise injection process while gradually transforming the data distribution into a standard Gaussian distribution.

The VP diffusion process is defined as:

$$f(S,t) = -\frac{1}{2}\beta(t)S, \quad g(t) = \sqrt{\beta(t)}, \tag{26}$$

where $\beta(t)$ is the noise schedule function controlling the rate of noise injection.

Under this formulation, the terminal state satisfies:

$$S_{\text{miss}}(T) \sim \mathcal{N}(0, I). \tag{27}$$

This process can be interpreted as a continuous stochastic perturbation of the data manifold, where each state depends only on its previous state, satisfying the Markov property.

J Supplementary Results

We provide comparisons with baseline methods on the MOSI, MOSEI, and SIMS datasets, as shown in Tables 5, 6, and 7, respectively.

Results under DMS								
Dataset	**Model**	{t} ACC/F1	{a} ACC/F1	{v} ACC/F1	{t,a} ACC/F1	{t,v} ACC/F1	{a,v} ACC/F1	**Avg.** ACC/F1
MOSI	MSAF-CF	82.50/-	59.60/-	61.10/-	81.10/-	80.90/-	59.10/-	70.70/-
	DERL	84.18/-	50.52/-	50.52/-	84.18/-	84.32/-	50.52/-	67.40/-
	MMIN	83.80/83.80	55.30/51.50	57.00/54.00	84.00/84.00	83.80/83.90	60.40/58.50	70.70/69.30
	MPLMM	83.80/83.80	54.60/42.00	54.70/42.10	83.70/83.60	83.80/83.70	57.30/42.60	69.70/63.00
	GCNet	83.70/83.60	56.10/54.50	56.10/55.70	84.50/84.40	84.30/84.20	62.00/61.90	71.10/70.70
	DicMoR	84.50/84.40	60.50/60.80	62.20/60.20	85.50/85.50	85.50/85.40	64.00/63.50	73.70/73.30
	UniMF	85.30/85.20	61.60/59.30	61.70/61.60	85.40/85.20	85.60/85.40	62.80/62.70	73.70/73.20
	MulDeF	85.20/85.20	57.90/57.50	56.70/53.50	85.50/85.40	85.50/85.40	57.90/57.70	71.50/70.80
	IIE	85.50/85.30	62.50/61.50	62.50/62.30	85.80/85.60	85.80/85.80	65.20/64.20	74.60/74.10
	SIGCD(Ours)	85.70/85.60	62.30/62.10	62.40/62.50	86.10/86.10	86.20/86.00	65.50/65.20	**74.70/74.60**
MOSEI	MSAF-CF	81.00/-	65.50/-	65.60/-	80.80/-	81.00/-	66.30/-	73.40/-
	DERL	85.20/-	60.40/-	60.50/-	85.20/-	85.60/-	60.10/-	72.80/-
	MMIN	82.30/82.40	58.90/59.50	59.30/60.00	83.70/83.30	83.80/83.40	63.50/61.90	71.90/71.80
	MPLMM	84.60/84.60	61.80/49.40	62.20/53.20	85.10/85.10	85.20/85.10	63.20/51.00	73.70/68.10
	GCNet	83.00/83.20	60.20/60.30	61.90/61.60	84.30/84.40	84.30/84.40	64.10/57.20	73.00/71.90
	DicMoR	84.20/84.30	62.90/60.40	63.60/63.60	85.00/84.90	84.90/84.90	65.20/64.40	74.30/73.80
	UniMF	84.80/84.80	62.40/55.70	62.10/58.10	85.10/85.20	85.50/85.30	63.40/60.10	73.90/71.50
	MulDeF	85.20/85.10	62.80/60.00	62.40/60.40	85.40/85.20	85.30/85.20	62.30/60.90	73.90/72.80
	IIE	85.50/85.30	61.90/57.80	65.20/63.40	85.70/85.60	85.90/85.90	65.20/63.70	74.90/73.60
	SIGCD(Ours)	85.70/85.40	62.20/63.50	64.80/64.70	86.20/86.00	86.10/86.40	66.40/65.90	**75.20/75.30**

Table 5: Results under DMS (%) on MOSI and MOSEI

Results under SMS									
Dataset	**Model**	0.1 ACC/F1	0.2 ACC/F1	0.3 ACC/F1	0.4 ACC/F1	0.5 ACC/F1	0.6 ACC/F1	0.7 ACC/F1	**Avg.** ACC/F1
MOSI	CTFN	79.70/-	77.10/-	74.60/-	71.20/-	67.40/-	63.20/-	61.50/-	70.70/-
	MMIN	81.80/81.80	79.00/79.10	76.10/76.20	71.70/71.60	67.20/66.50	64.90/64.00	62.80/61.00	71.90/71.50
	GCNet	82.30/82.30	79.40/79.50	77.20/77.20	74.30/74.40	70.00/69.80	67.70/66.70	65.70/65.40	73.80/73.60
	MPLMM	80.60/80.70	77.70/77.90	75.00/75.10	70.00/69.80	67.40/66.60	63.40/61.80	59.20/56.20	70.50/69.70
	DicMoR	83.20/83.20	81.60/81.30	78.40/77.90	76.20/74.70	72.60/72.70	71.30/71.40	69.10/69.20	76.10/75.80
	LNLN	81.40/81.40	78.70/78.70	74.70/74.80	70.30/69.90	67.10/66.30	62.50/60.50	61.00/58.30	70.80/70.00
	UniMF	83.70/83.40	80.80/80.80	79.10/79.10	75.80/75.30	72.90/72.60	71.70/71.30	70.30/68.70	76.30/75.90
	MulDeF	81.70/81.80	79.70/78.90	75.00/75.10	73.20/70.30	67.20/66.10	63.10/61.30	60.70/58.30	71.50/70.30
	IIE	85.10/84.90	82.30/82.30	81.70/81.70	79.30/78.80	77.00/76.50	73.60/72.90	73.20/72.20	**78.90**/78.50
	SIGCD(Ours)	84.50/84.50	83.40/83.40	81.10/81.10	79.40/79.30	76.20/75.30	74.10/73.30	73.90/73.90	**78.90/78.70**
MOSEI	CTFN	81.20/-	79.50/-	77.40/-	74.20/-	72.10/-	69.20/-	66.70/-	74.30/-
	MMIN	81.90/81.30	79.80/78.80	77.20/75.50	75.20/72.60	73.90/70.70	73.20/70.30	73.10/69.50	76.30/74.10
	GCNet	82.30/82.10	80.30/79.90	77.50/76.80	76.00/74.90	74.90/73.20	74.10/72.10	73.20/70.40	76.90/75.60
	MPLMM	83.70/83.50	81.40/80.80	79.20/78.30	77.10/76.10	75.60/73.80	72.70/70.60	70.50/69.20	77.20/76.00
	DicMoR	83.20/83.10	81.20/80.80	79.00/77.90	76.90/75.20	73.70/71.70	71.10/70.40	70.60/68.90	76.50/75.40
	LNLN	83.10/83.00	81.70/81.00	79.10/78.10	76.90/75.60	75.10/72.40	72.50/68.60	70.60/65.20	77.00/74.80
	UniMF	83.70/83.40	81.30/80.50	78.50/77.20	76.90/75.30	74.90/72.20	71.90/67.90	71.20/66.50	76.90/74.70
	MulDeF	82.90/82.80	80.80/80.40	78.10/77.10	76.50/74.60	74.40/71.90	71.80/68.00	71.00/66.20	76.50/74.40
	IIE	84.20/83.80	82.20/81.60	79.60/78.90	77.40/76.30	75.70/74.50	71.60/70.00	71.60/70.00	77.50/76.40
	SIGCD(Ours)	83.80/84.00	82.60/81.80	80.50/79.20	78.30/78.10	76.10/75.50	74.20/73.40	71.40/69.10	**78.10/77.30**

Table 6: Results under SMS (%) on MOSI and MOSEI

Models	**{t}** ACC/F1	**{a}** ACC/F1	**{v}** ACC/F1	**{t, a}** ACC/F1	**{t, v}** ACC/F1	**{a, v}** ACC/F1	**Avg.** ACC/F1
LNLN	79.72/-	81.91/-	81.91/-	80.11/-	79.69/-	81.91/-	80.88/-
EMOE	78.66/-	81.25/-	80.69/-	79.54/-	79.36/-	81.27/-	80.13/-
PRLF	80.28/-	81.74/-	81.63/-	80.57/-	80.36/-	82.58/-	81.19/-
MPMM	78.56/78.65	64.98/76.41	65.40/77.92	77.11/77.20	77.51/77.47	64.01/73.47	71.26/76.85
MPLMM	79.75/78.74	65.93/77.10	66.02/78.86	77.45/77.84	77.97/77.95	65.28/74.02	72.07/77.42
MUMP	81.82/80.99	69.51/82.01	69.73/82.79	80.71/80.36	80.27/80.48	68.73/78.54	75.13/80.86
SIGCD(Ours)	80.42/79.72	81.94/82.38	82.47/82.86	79.40/79.89	81.44/80.83	81.96/81.61	**81.27/81.22**

Table 7: Results under DMS (%) on CH-SIMS