S^2Diff: Semantic–Sentiment Disentangled Diffusion for Missing-Modality Multimodal Sentiment Analysis

Anonymous ACL submission

Abstract

Multimodal Sentiment Analysis (MSA) integrates text, speech, and vision to infer sentiment, but real-world modalities are often incomplete, weakening cross-modal complementarity and degrading prediction performance. Although generative reconstruction can recover missing modalities in the representation space, it may be affected by sentiment distribution mismatch when affective cues are inconsistent across modalities. In such cases, generated features can be biased toward the sentiment tendency of observed modalities, which may impair downstream prediction. To mitigate this risk, we propose S^2Diff, a structured generative framework for missing-modality MSA. S^2Diff disentangles each modality representation into semantic and sentiment factors, where semantics preserve content consistency and alignability, while sentiment captures polarity, intensity, and expression style. Based on this factorization, S^2Diff employs a dual-branch diffusion model to perform conditional reconstruction in the semantic and sentiment subspaces, respectively. The two branches interact through partial parameter sharing, enabling complementary transfer while maintaining relative factor independence. As a result, reconstructed representations are encouraged to preserve cross-modal semantic coherence and reduce sentiment-space bias toward observed modalities. Experiments on MOSI, MOSEI, and IEMOCAP show consistent improvements and strong robustness across diverse missing-modality settings.

1 Introduction

Multimodal Sentiment Analysis (MSA) jointly models text, speech, and vision to infer sentiment, and has been widely applied in human–computer interaction, dialogue understanding, and social media mining (Wu et al., 2023; Zhang et al., 2023a). In practice, however, multimodal data are often incomplete due to modality asynchrony, occlusions, noisy signals, privacy constraints, and sensor failures. Such missing modalities disrupt cross-modal alignment and complementarity, making learned fusion strategies less reliable at inference and causing substantial performance degradation (Guo et al., 2024; Li et al., 2024). Therefore, robustness to missing modalities is essential for real-world MSA.

Existing missing-modality MSA methods mainly fall into four categories: (i) Robust Generalization, which mitigates missing-pattern-induced shifts via reweighting, alignment, or invariance constraints (Zhou et al., 2024; Zhang et al., 2024; Zhu et al., 2025); (ii) Expert Routing, which improves reliability through expert priors/routing, structural cues, or uncertainty modeling (Xu et al., 2024; Han et al., 2024; Yun et al., 2024); (iii) Prompt Adaptation, which encodes missing status as prompts or conditions to adapt across missing configurations (Guo et al., 2024; Zhao et al., 2025; Lang et al., 2025); and (iv) Generative Reconstruction, which learns conditional mappings from observed to missing modalities to recover complementary representations before fusion (Zhao et al., 2021; Wang et al., 2023b). Among them, generative reconstruction is particularly important in severe or fully missing cases (Dai et al.; Wu et al., 2024). Recently, diffusion models have emerged as a promising backbone due to stable training and strong generative capacity (Karras et al., 2022; Dhariwal and Nichol, 2021).

Despite the effectiveness of reconstruction-based paradigms, a potential risk remains in MSA: missing-modality reconstruction may be affected by sentiment distribution mismatch when affective cues are inconsistent across modalities (Liu et al., 2025; Wang et al., 2023a). As shown in Fig. 1, sentiment cues can differ across modalities in expression patterns, intensity scales, and temporal manifestations, and their consistency with global labels is not always guaranteed (Yu et al., 2021). When generating a missing modality conditioned

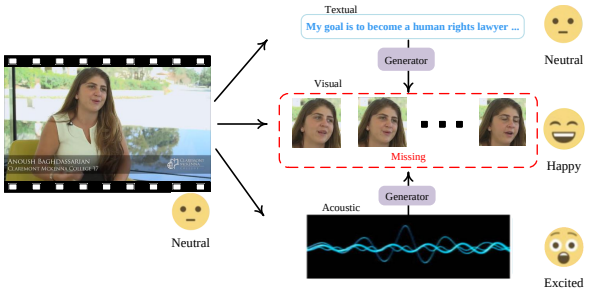

Figure 1: Effect of multimodal sentiment discrepancy on missing-modality reconstruction. A single utterance may exhibit different affective cues across text, vision, and audio, causing inconsistency between unimodal sentiments and the overall label. During reconstruction, the generator may inherit the observed modalities' sentiment bias and transfer it to the missing modality.

on observed modalities, the reconstructed representation may be pulled toward the sentiment tendency of the observed modalities, becoming more like a projection of observed affective cues than a faithful completion of the target-modality distribution. This bias can affect reconstruction quality and propagate through subsequent fusion, potentially undermining prediction reliability and robustness.

One possible reason for this risk is that many methods model semantic consistency and sentiment variation in a shared representation space, where factors with different properties may interfere with each other. Semantic information is usually more structured and alignable, whereas sentiment information is more abstract, less statistically stable, and more modality-dependent. Jointly modeling them may introduce competing constraints during reconstruction: semantic coherence can become entangled with biased sentiment transfer, making generated features more sensitive to the observed modalities' affective tendency. As a result, reconstruction may become vulnerable to sentiment bias, limiting the benefits of generative completion for downstream prediction.

To mitigate this risk, we present S^{2}Diff, a generative framework for missing-modality MSA. S^{2}Diff explicitly disentangles modality representations into semantic and sentiment factors: the semantic factor enforces content consistency and structural alignability, while the sentiment factor captures variations in polarity, intensity, and expression style. This factorization allows cross-modal completion to preserve semantic consistency while reducing the risk that sentiment factors are distorted by observation-induced bias. Based on this representation, S^{2}Diff adopts a dual-branch diffusion model, where the semantic and sentiment branches perform conditional diffusion and missing-modality reconstruction in their respective subspaces. Controlled interaction is implemented via partial parameter sharing, supporting complementary information exchange and consistency-aware conditioning while preserving relative factor independence. This design encourages completed representations to remain semantically coherent across modalities and less biased toward the observed sentiment tendency, leading to more discriminative representations for fusion-based prediction.

Extensive experiments are conducted on three standard benchmarks, MOSI (Zadeh et al., 2016), MOSEI (Zadeh et al., 2018), and IEMOCAP (Menze et al., 2015). Under diverse missing configurations and missing ratios, S^2Diff consistently yields stable improvements and stronger robustness. For example, on the large-scale MOSEI dataset, it surpasses the strongest prior method by 1.02% in ACC and 1.82% in F1 under fixed-missing settings, while showing markedly smaller degradation as the random missing rate increases.

Our contributions are summarized as follows:

- We motivate semantic–sentiment factorization for missing-modality MSA from the risk of observation-induced sentiment drift, where modality-specific affective cues may bias conditional reconstruction under cross-modal sentiment inconsistency.
- We propose S^{2}Diff, a dual-branch conditional diffusion framework that models semantic and sentiment factors separately while allowing controlled interaction through partial parameter sharing.
- Experiments on MOSI, MOSEI, and IEMOCAP under fixed- and random-missing protocols show that S^{2}Diff consistently improves robustness over strong missing-modality baselines, with ablations validating the contribution of semantic and sentiment guidance.

2 Related Work

2.1 Incomplete Modalities in MSA

MSA relies on complementary cues from language, vision, and acoustics, but real-world inputs are often incomplete due to sensor failures, occlusions/noise, privacy constraints, or network issues.

Early imputation methods, such as zero padding, mean filling, or statistic-based replacement, are simple but may confuse missingness with weak evidence and induce distribution shift (Zhang et al., 2020; Parthasarathy and Sundaram, 2020). Existing methods mainly fall into four categories: Robust Generalization mitigates missing-pattern shifts via reweighting, alignment, calibration, or invariance regularization (Zhou et al., 2024; Zhang et al., 2024; Zhu et al., 2025); Expert Routing enhances fusion with structures, priors, or uncertainty, e.g., *MoMKE* (Yun et al., 2024); Prompt Adaptation encodes modality availability as prompts, e.g., *MPLMM* (Guo et al., 2024); and Generative Reconstruction recovers missing representations before fusion, from *MMIN* to diffusion-based *IMDer* (Zhao et al., 2021; Wang et al., 2023b). Unlike prior diffusion-based completion with generic conditioning, our method explicitly separates semantic consistency and sentiment factors to reduce reconstruction-induced sentiment shift.

2.2 Condition Injection and Controllable Generation in Diffusion Models

Conditional diffusion models require effective condition propagation throughout denoising to maintain controllability (Chen et al., 2025; Zhang et al., 2023c; Ho, 2022). Existing methods commonly adopt cross-attention, which uses condition embeddings as keys/values (Rombach et al., 2021; Saharia et al., 2022; Hertz et al., 2022), or feature modulation, which injects conditions through normalization or residual modulation (Peebles and Xie, 2022; Zhang et al., 2023b; Dhariwal and Nichol, 2021). Sampling-time guidance further strengthens condition adherence during inference (Dhariwal and Nichol, 2021; Ho, 2022; Saharia et al., 2022). Branch-based or decoupled designs improve controllability by injecting conditions across layers or scales, especially for structured and alignable conditions (Zhang et al., 2023b; Mou et al., 2023; Zhao et al., 2023). However, in missing-modality MSA, semantic cues are relatively structured, whereas sentiment cues are more abstract and modality-dependent. A single pathway may therefore introduce competing constraints and impair distribution matching. We address this by using a semantic–sentiment dual-branch diffusion framework with partial parameter sharing, enabling complementary transfer while preserving factor independence.

3 Model

3.1 Problem Formulation

We study multimodal sentiment analysis with modalities $\mathcal{M} = \{T, A, V\}$, where $k \in \mathcal{M}$ denotes a modality index. At inference, some modalities may be missing. We use an availability indicator $\boldsymbol{\alpha} \in \{0,1\}^{|\mathcal{M}|}$ and define the observed and missing modality sets as $\mathcal{M}_{\text{obs}} = \{k : \alpha_k = 1\}$ and $\mathcal{M}_{\text{mis}} = \{k : \alpha_k = 0\}$, with $\mathcal{M}_{\text{obs}} \neq \emptyset$. We denote an observed modality by $o \in \mathcal{M}_{\text{obs}}$ and a missing modality by $m \in \mathcal{M}_{\text{mis}}$.

Given incomplete inputs, the goal is to predict sentiment y by reconstructing missing modalities. For each $m \in \mathcal{M}_{\text{mis}}$, we learn a conditional generator to approximate $p(X_m \mid X_o)$ and obtain a reconstructed feature $\widehat{X}_m$. The final prediction is then produced by a fusion model:

$$\widehat{y} = f(X_o, \widehat{X}_m) \tag{1}$$

3.2 Overall Framework

As shown in Fig. 2, S^{2}Diff is a generative completion framework for missing-modality MSA. It contains four stages: feature extraction, semantic–sentiment decomposition with sentiment transformation, dual-branch diffusion completion, and multimodal fusion.

(1) Feature Extraction. Given raw text (T), audio (A), and vision (V), each modality k is encoded by a modality-specific encoder $\text{Enc}_k(\cdot)$ into a sequence feature

$$X_k = \text{Enc}_k(\cdot), \qquad X_k \in \mathbb{R}^{L \times d} \tag{2}$$

where L is the unified sequence length and d is the hidden dimension.

(2) Semantic–Sentiment Decomposition and Sentiment Transformation. To reduce observation-induced sentiment bias, we decompose observed features into semantic and sentiment factors:

$$(H_o^{\text{sem}},\, H_o^{\text{sent}}) = \text{Decmp}(X_o);\, H_o^{\text{sem}},\, H_o^{\text{sent}} \in \mathbb{R}^{L \times d}, \tag{3}$$

where H_o^{sem} captures cross-modal semantic content and H_o^{sent} captures sentiment cues such as polarity, intensity, and expression style. For each missing modality $m \in \mathcal{M}_{\text{mis}}$, a sentiment transformation module maps the observed sentiment factor into the target sentiment subspace:

$$\widehat{H}_m^{\text{sent}} = \text{Trans}_m\big(H_o^{\text{sent}}\big);\, \widehat{H}_m^{\text{sent}} \in \mathbb{R}^{L \times d} \tag{4}$$

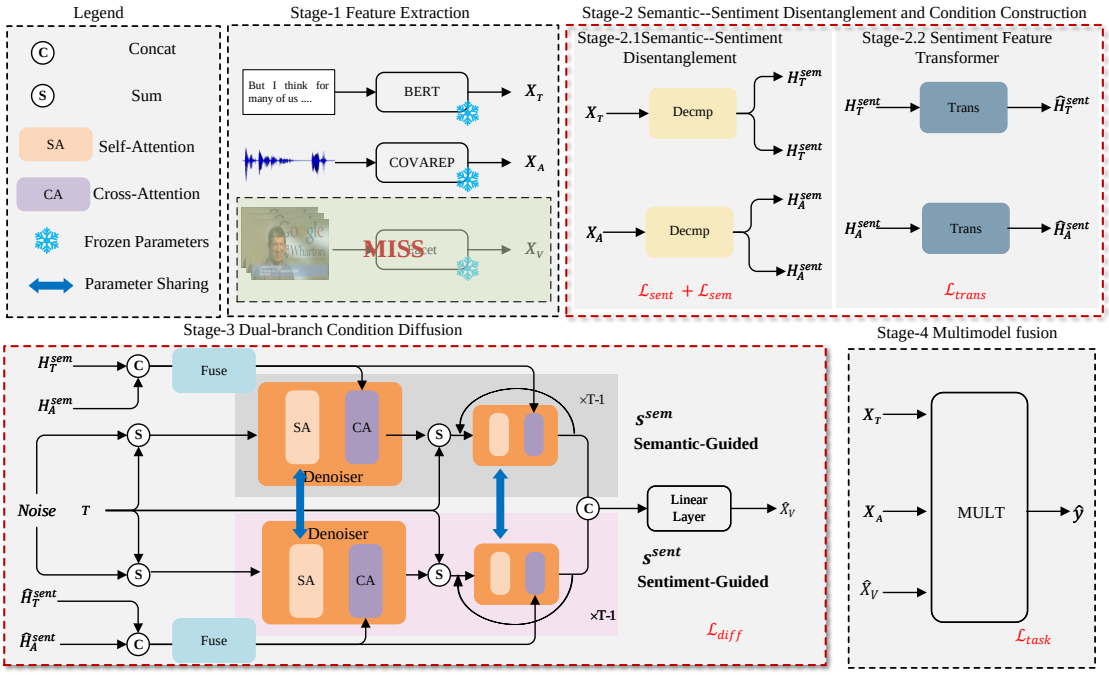

Figure 2: Overview of S^{2}Diff with text/audio observed and video missing as an example. In stage (2), semantic–sentiment disentanglement constructs semantic and sentiment conditions, supervised by $\mathcal{L}_{\text{sem}}$, $\mathcal{L}_{\text{sent}}$, and $\mathcal{L}_{\text{trans}}$. In stage (3), a dual-branch conditional diffusion model imputes the missing video representation under semantic and sentiment guidance, optimized by $\mathcal{L}_{\text{diff}}$. The final prediction is supervised by $\mathcal{L}_{\text{task}}$.

(3) Dual-branch Diffusion Completion. For each missing modality m, we reconstruct its feature using a conditional dual-branch diffusion model, conditioned on the semantic context H_o^{sem} and the target-aligned sentiment prompt $\widehat{H}_m^{\text{sent}}$:

$$\widehat{X}_m = \text{S}^2\text{Diff}_m\big(\mathbf{n};\, H_o^{\text{sem}}, \widehat{H}_m^{\text{sent}}\big);\ \widehat{X}_m \in \mathbb{R}^{L\times d} \tag{5}$$

where $\mathbf{n} \in \mathbb{R}^{L\times d}$ is Gaussian noise. The semantic and sentiment branches model the two factors separately and interact through partial parameter sharing.

(4) Multimodal Fusion and Sentiment Prediction. Finally, the observed features X_o and reconstructed features $\widehat{X}_m$ are fused by MULT (Tsai et al., 2019) for sentiment prediction:

$$\widehat{y} = \text{MULT}(X_o, \widehat{X}_m) \tag{6}$$

3.3 Semantic–Sentiment Decomposition and Sentiment Transformation

In missing-modality MSA, cross-modal semantics are often more consistent than sentiment, while sentiment expressions usually exhibit modality-specific biases in intensity, style, and temporal patterns. Directly conditioning generation on observed features may therefore transfer biased sentiment signals to the missing modality. To alleviate this issue, we adopt a decompose-then-transform strategy: each modality is decomposed into semantic and sentiment components, and only the sentiment component is mapped across modalities to form a target-aligned sentiment prompt.

Given a modality feature $X_k \in \mathbb{R}^{L\times d}$, we first extract an utterance-level sentiment vector

$$z_k = \text{SEnc}_k(X_k) \in \mathbb{R}^d \tag{7}$$

which captures the modality-specific sentiment state. Then, for each hidden state $h_{k,t} \in \mathbb{R}^d$, we normalize the sentiment direction as

$$u_k = \frac{z_k}{\|z_k\|_2 + \varepsilon} \tag{8}$$

and decompose $h_{k,t}$ by orthogonal projection:

$$h_{k,t}^{\text{sent}} = (h_{k,t}^{\top} u_k)u_k,\ h_{k,t}^{\text{sem}} = h_{k,t} - h_{k,t}^{\text{sent}} \tag{9}$$

Accordingly, the sentiment and semantic feature sequences are $H_k^{\text{sent}} = \{h_{k,t}^{\text{sent}}\}_{t=1}^{L}, H_k^{\text{sem}} = \{h_{k,t}^{\text{sem}}\}_{t=1}^{L}$, and $H_k^{\text{sent}},\ H_k^{\text{sem}} \in \mathbb{R}^{L\times d}$.

For a missing modality m, its sentiment feature cannot be directly obtained. We therefore transform only the sentiment components from observed

modalities. For each observed modality o, a sentiment transformer maps

$$\widehat{H}^{\text{sent}}_{o\to m} = \text{Trans}_{o\to m}\big(H^{\text{sent}}_{o}\big) \in \mathbb{R}^{L\times d} \tag{10}$$

and, during training, $\widehat{H}^{\text{sent}}_{o\to m}$ is supervised by the ground-truth target sentiment feature H^{sent}_{m}. When multiple modalities are observed, the transformed prompts are fused as

$$\widehat{H}^{\text{sent}}_{m} = \text{Fuse}\big(\{\widehat{H}^{\text{sent}}_{o\to m}\}\big) \in \mathbb{R}^{L\times d} \tag{11}$$

while the semantic features are fused into a global semantic condition:

$$H^{\text{sem}}_{\text{obs}} = \text{Fuse}\big(\{H^{\text{sem}}_{o}\}\big) \in \mathbb{R}^{L\times d} \tag{12}$$

The resulting semantic condition $H^{\text{sem}}_{\text{obs}}$ provides cross-modal contextual guidance, while $\widehat{H}^{\text{sent}}_{m}$ serves as a target-aligned sentiment prior for subsequent diffusion completion.

3.4 Dual-branch Conditional Diffusion Completion

To complete a missing modality, we build a dual-branch conditional diffusion model on top of the semantic–sentiment decomposition. The semantic condition $H^{\text{sem}}_{\text{obs}}$ provides content-level constraints, whereas the sentiment prompt $\widehat{H}^{\text{sent}}_{m}$ guides alignment to the target sentiment distribution. Modeling them in a single pathway may cause interference; therefore, we use two denoising branches with partial parameter sharing.

Let m denote the missing modality and $X^{(0)}_{m} \in \mathbb{R}^{L\times d}$ its clean feature. We aim to model $p(X_m \mid H^{\text{sem}}_{\text{obs}}, \widehat{H}^{\text{sent}}_{m})$ under a score-based diffusion framework. At diffusion step t, the noisy sample is

$$X^{(t)}_{m} = X^{(0)}_{m} + \sigma_t z, \qquad z \sim \mathcal{N}(0, I) \tag{13}$$

where σ_t is the noise scale. Let $C_s = H^{\text{sem}}_{\text{obs}}$, $C_a = \widehat{H}^{\text{sent}}_{m}$, and $C_m = \{C_s, C_a\}$. The conditional score is decomposed into semantic and sentiment guidance:

$$\begin{aligned}&\nabla_{X_m} \log p_t(X^{(t)}_{m} \mid C_m) \approx \\ &\nabla_{X_m} \log p_t(X^{(t)}_{m} \mid C_s) + \nabla_{X_m} \log p_t(X^{(t)}_{m} \mid C_a)\end{aligned} \tag{14}$$

which motivates a dual-branch score network $s_m(\cdot)$ with a semantic branch $s^{\text{sem}}_{m}(\cdot)$ and a sentiment branch $s^{\text{sent}}_{m}(\cdot)$:

$$\begin{aligned}&s_m\Big(X^{(t)}_{m}, t;\, H^{\text{sem}}_{\text{obs}}, \widehat{H}^{\text{sent}}_{m}\Big) = \\ &\quad s^{\text{sem}}_{m}\Big(X^{(t)}_{m}, t;\, H^{\text{sem}}_{\text{obs}}\Big) + s^{\text{sent}}_{m}\Big(X^{(t)}_{m}, t;\, \widehat{H}^{\text{sent}}_{m}\Big)\end{aligned} \tag{15}$$

Both branches are built with stacked Transformer denoising blocks. The self-attention modules are shared to model common temporal structures of the missing modality, while branch-specific cross-attention modules inject $H^{\text{sem}}_{\text{obs}}$ and $\widehat{H}^{\text{sent}}_{m}$, respectively. Therefore, the additional parameters mainly arise from the branch-specific cross-attention layers and the final fusion projection. The two denoised outputs are concatenated and linearly projected to obtain $\widehat{X}_m \in \mathbb{R}^{L\times d}$.

Overall, the dual-branch design separates semantic consistency and sentiment alignment during denoising, while coupling them through partial parameter sharing to improve completion quality and downstream sentiment prediction.

3.5 Training Objective

The proposed model is trained end-to-end. Let $\widehat{y}$ denote the output of the multimodal fusion network and y the ground-truth label. The overall objective consists of a task loss and four auxiliary losses, including semantic alignment, sentiment alignment, sentiment transformation supervision, and diffusion loss.

(1) Task Loss. The task loss supervises the final prediction:

$$\mathcal{L}_{\text{task}} = \ell(\widehat{y}, y) \tag{16}$$

where $\ell(\cdot)$ denotes the cross-entropy loss for classification and the mean squared error for regression.

(2) Utterance-level Semantic Alignment Loss. To encourage semantic consistency across modalities, we first obtain an utterance-level semantic representation from each semantic sequence by pooling:

$$q^{\text{sem}}_{k} = \text{Pool}(H^{\text{sem}}_{k}) \in \mathbb{R}^{d} \tag{17}$$

where $\text{Pool}(\cdot)$ denotes average pooling over temporal steps. We then impose pairwise semantic alignment across modalities:

$$\mathcal{L}_{\text{sem}} = \sum_{i\neq j} \big(1 - \cos\big(q^{\text{sem}}_{i}, q^{\text{sem}}_{j}\big)\big)\,. \tag{18}$$

(3) Utterance-level Sentiment Alignment Loss. To improve the alignability of utterance-level sentiment representations, a modality-specific prediction head $g_k(\cdot)$ is attached to each sentiment vector z_k, producing $\widehat{y}_k = g_k(z_k)$. The corresponding loss is

$$\mathcal{L}_{\text{sent}} = \sum_{k\in\mathcal{M}} \ell(\widehat{y}_k, y) \tag{19}$$

Table 1: Comparison under the fixed-missing protocol. Results are reported on MOSI, MOSEI, and IEMOCAP under six available-modality settings $\{t\}$, $\{v\}$, $\{a\}$, $\{tv\}$, $\{ta\}$, $\{va\}$ and their average (avg). ACC and F1 are used for MOSI/MOSEI, while WA(weighted accuracy) and UA(unweighted accuracy) are used for IEMOCAP. Best results are in **bold**.

Dataset	Models	Available Modalities													
		$\{T\}$		$\{V\}$		$\{A\}$		$\{T,V\}$		$\{T,A\}$		$\{V,A\}$		avg	
		ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
MOSI	MULT	82.24	81.87	57.87	53.71	45.63	34.07	81.94	81.82	82.65	82.28	48.25	41.78	66.43	62.59
	SELFMM	82.24	82.01	49.49	29.93	45.19	32.86	82.94	82.65	82.38	83.16	47.52	36.46	64.96	57.85
	MMIN	82.23	82.80	58.48	57.67	54.75	52.77	81.57	81.45	82.13	82.19	50.75	50.67	68.32	67.92
	IMDER	82.63	82.65	58.13	57.74	58.25	58.26	83.36	82.31	83.01	83.07	59.12	58.55	70.75	70.43
	MPLMM	82.65	82.57	57.58	54.04	57.00	56.63	82.13	82.16	82.49	82.46	58.27	58.93	70.02	69.47
	MOMKE	81.86	81.72	58.08	57.95	58.77	58.18	82.57	82.56	82.13	82.08	58.33	58.29	70.29	70.13
	LNLN	81.05	81.06	58.75	57.67	59.75	57.67	83.29	83.20	82.22	82.21	59.79	58.89	70.81	70.12
	PRMF	81.66	81.63	59.33	59.32	59.64	59.40	83.25	83.22	83.33	83.13	59.76	59.64	71.16	71.06
	Ours	**83.64**	**83.13**	**60.14**	**60.04**	**60.21**	**60.22**	**83.60**	**83.45**	**84.07**	**83.67**	**60.00**	**59.73**	**71.94**	**71.71**
MOSEI	MULT	81.91	82.21	62.15	60.87	60.56	60.14	82.31	82.76	81.52	81.68	58.02	57.95	71.08	70.94
	SELFMM	80.41	79.62	52.63	54.81	45.48	46.10	72.59	73.85	74.76	75.87	55.10	55.67	63.50	64.32
	MMIN	82.44	82.78	68.90	65.60	60.61	62.04	80.79	81.30	80.60	81.10	66.37	66.40	73.29	73.20
	IMDER	83.97	82.39	70.66	62.08	64.19	61.88	83.79	83.11	83.43	83.67	66.47	66.20	75.42	73.22
	MPLMM	82.05	81.64	67.42	65.18	63.28	63.41	82.01	82.45	82.08	82.06	65.39	64.04	73.71	73.13
	MOMKE	82.05	81.66	67.38	65.11	62.11	60.15	82.44	82.41	81.94	81.99	63.97	63.14	73.32	72.41
	LNLN	79.65	80.21	65.53	63.04	67.83	63.45	83.25	83.71	83.07	82.74	64.24	64.49	73.93	72.94
	PRMF	83.64	83.55	66.48	60.11	67.99	64.23	83.97	83.88	82.44	82.49	65.57	64.34	75.02	73.10
	Ours	**84.23**	**83.97**	**69.89**	**66.86**	**69.64**	**65.12**	**84.34**	**84.34**	**83.36**	**83.41**	**67.26**	**67.35**	**76.45**	**75.18**
		WA	UA	WA	UA	WA	UA	WA	UA	WA	UA	WA	UA	WA	UA
IEMOCAP	MULT	74.93	74.17	60.82	60.06	58.42	55.95	65.62	66.64	71.47	71.85	64.91	64.49	66.03	65.53
	SELFMM	67.22	66.38	61.81	61.74	56.74	55.34	65.99	65.14	71.21	72.33	64.37	64.76	64.56	64.28
	MMIN	70.94	70.25	62.62	64.35	62.94	61.25	69.16	70.15	75.27	79.67	66.51	67.11	67.91	68.80
	IMDER	76.25	76.63	63.19	63.79	68.19	68.79	75.84	75.28	78.55	78.19	68.77	68.30	71.80	71.83
	MPLMM	72.86	70.31	62.41	62.37	65.17	65.94	74.96	74.83	77.54	78.02	67.14	67.24	70.01	69.79
	MOMKE	77.20	75.77	60.28	60.77	67.26	65.46	75.24	75.76	78.13	78.87	68.58	68.41	71.12	70.84
	LNLN	76.61	75.47	61.24	60.38	66.91	62.46	74.33	74.96	77.68	77.10	67.31	67.33	70.68	69.62
	PRMF	76.92	76.34	61.73	60.98	65.59	65.86	74.61	74.92	77.84	77.24	67.84	67.24	70.76	70.43
	Ours	**78.33**	**78.39**	**65.34**	**65.78**	**70.24**	**71.68**	**77.28**	**76.34**	**80.71**	**79.69**	**70.88**	**70.38**	**73.80**	**73.71**

(4) Sentiment Transformation Supervision Loss. During training, the transformed sentiment prompt is supervised by the target-modality sentiment feature. Let $\mathcal{P}$ denote the set of valid source–target modality pairs. The loss is defined as

$$\mathcal{L}_{\text{trans}} = \sum_{(o,m)\in\mathcal{P}} \left\| \widehat{H}^{\text{sent}}_{o\to m} - H^{\text{sent}}_{m} \right\|_2^2 \tag{20}$$

(5) Dual-branch Conditional Diffusion Loss. For the diffusion completion module, denoising score matching is adopted. Given a clean feature $X_m^{(0)}$ of modality m, we construct the noisy sample as $X_m^{(t)}$. Let $s_m^{(t)} = s_m(X_m^{(t)}, t; H^{\text{sem}}_{\text{obs}}, \widehat{H}^{\text{sent}}_m)$. The diffusion loss is

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{m,t,z}\left[\lambda(t)\left\| s_m^{(t)} + \frac{1}{\sigma_t} z \right\|_2^2\right], \tag{21}$$

where $\lambda(t)$ is a time-step weighting function.

Overall Objective. The total objective is

$$\begin{aligned}\mathcal{L} =& \lambda_{\text{task}}\mathcal{L}_{\text{task}} + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}} \\ &+ \lambda_{\text{sent}}\mathcal{L}_{\text{sent}} + \lambda_{\text{trans}}\mathcal{L}_{\text{trans}} + \lambda_{\text{diff}}\mathcal{L}_{\text{diff}}\end{aligned} \tag{22}$$

where $\lambda_{\text{task}}, \lambda_{\text{sem}}, \lambda_{\text{sent}}, \lambda_{\text{trans}}, \lambda_{\text{diff}} \geq 0$ are weighting coefficients. In practice, λ_{task} can be set to 1 by default.

4 Experiments

4.1 Datasets

We evaluate the proposed method on three public multimodal benchmarks: MOSI (Zadeh et al., 2016), MOSEI (Zadeh et al., 2018), and IEMOCAP (Menze et al., 2015), all of which provide text (T), vision (V), and audio (A) modalities. MOSI and MOSEI are English video sentiment datasets annotated with sentiment intensity in $[-3, +3]$, where MOSI is relatively small and MOSEI is a larger and more diverse extension. IEMOCAP is a multimodal conversational emotion dataset, which we follow prior work to formulate as a 4-way classification task. Their consistent modality setup enables systematic evaluation under different missing-modality settings.

4.2 Evaluation Settings and Criteria

We evaluate two inter-modal missing protocols: fixed-missing and random-missing. In fixed-

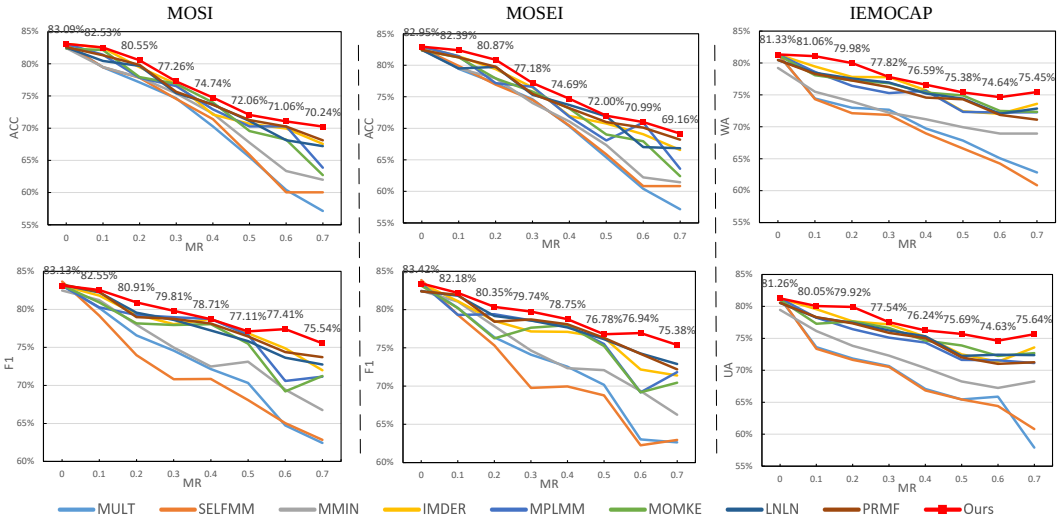

Figure 3: Performance degradation curves under the random-missing protocol. As the missing rate (MR) increases from 0.0 to 0.7, the figure shows the performance trends of different methods on MOSI, MOSEI, and IEMOCAP: ACC and F1 are reported for MOSI/MOSEI, while WA and UA are reported for IEMOCAP.

missing, six modality-availability settings, i.e., $\{T\}$, $\{V\}$, $\{A\}$, $\{T, V\}$, $\{T, A\}$, and $\{V, A\}$, are fixed during training and testing. In random-missing, available modalities are independently sampled for each instance, with the missing rate (MR) varied from 0 to 0.7 to assess robustness under increasing missingness. For MOSI and MOSEI, sentiment scores are converted into binary polarity labels, where scores ≤ 0 are negative and scores > 0 are positive, and accuracy (ACC) and F1-score are reported. For IEMOCAP, formulated as a 4-way classification task, we report weighted accuracy (WA) and unweighted accuracy (UA). All results are averaged over five runs unless otherwise stated. Detailed implementation settings, model parameters, and computational costs are provided in Appendices A and B.

4.3 Comparison with the State-of-the-Art Methods

We compare S^2Diff with representative MSA methods, including standard baselines (MULT (Tsai et al., 2019), SelfMM (Yu et al., 2021)), generative reconstruction methods (MMIN (Zhao et al., 2021), IMDer (Wang et al., 2023b)), prompt adaptation (MPLMM (Guo et al., 2024)), expert routing (MoMKE (Xu et al., 2024)), and robust generalization methods (LNLN (Zhang et al., 2024), PRMF (Zhu et al., 2025)). Table 1 and Fig. 3 report fixed-missing results and random-missing degradation trends, respectively, with additional confusion matrix analysis in Appendix D. Overall, S^2Diff remains stable across protocols and consistently outperforms strong baselines.

Performance under Fixed Missing Modality Protocol. Across the six fixed-missing settings, S^2Diff consistently achieves the best performance. On MOSI and MOSEI, S^2Diff improves over the strongest prior method by 0.78% and 1.02% in ACC, and by 0.65% and 1.82% in F1, respectively. On IEMOCAP, it yields average gains of 2.00% in WA and 1.88% in UA. Notably, the improvements are more pronounced when the language modality is missing. Since all methods exhibit substantially larger degradation under text-missing conditions across the three datasets—indicating a stronger sentiment shift induced by missing language—S^2Diff achieves, on average, 1.02% higher gains under text-missing protocols than under text-present protocols, demonstrating its effectiveness in suppressing sentiment shift.

Robustness under Random Missing Modality Protocol. As the missing rate increases from 0.0 to 0.7, S^2Diff shows markedly smaller performance degradation than all competing methods, demonstrating superior robustness to modality absence. On MOSEI, other methods drop by 9.55%–20.82% in ACC, whereas S^2Diff decreases by only 6.59%; on IEMOCAP, other methods degrade by 9.05%–20.59% in WA, while S^2Diff drops by only 5.88%. Further analysis suggests mild sensitivity to dataset scale: compared with the larger MOSEI, MOSI exhibits larger ACC drops under

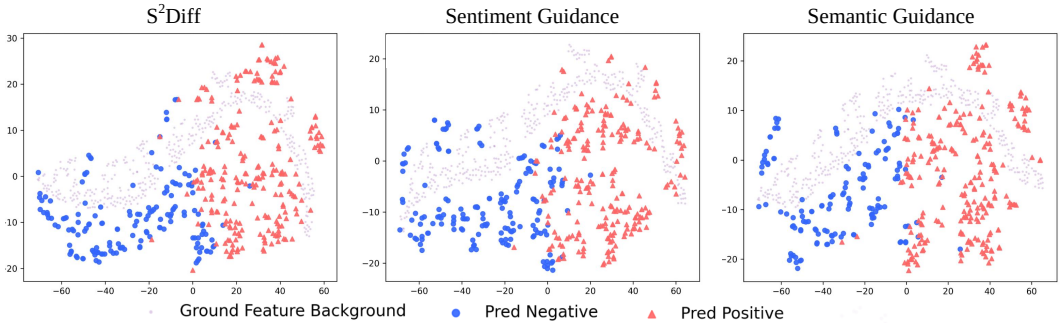

Figure 4: Visualization of recovered features under different guidance strategies. Ground-truth features are shown as the light-purple background, while recovered features are colored by predicted sentiment polarity.

Table 2: Ablation study of the guidance mechanism. S^2 guidance denotes dual-branch semantic and sentiment guidance, while $S+S$ denotes single-branch guidance with concatenated semantic and sentiment conditions.

Methods	MOSI		MOSEI		IEMOCAP	
	ACC	F1	ACC	F1	WA	UA
	Fixed-missing					
w/ S^2 Guidance	71.94	71.71	76.45	75.18	73.80	73.71
w/ $S+S$ Guidance	69.81	69.80	74.73	74.47	71.80	71.75
w/ Sentiment Guidance	69.50	69.43	74.42	72.22	71.34	71.62
w/ Semantic Guidance	70.94	70.86	75.24	74.13	72.34	72.18
w/o Guidance	68.14	67.82	73.25	72.57	70.61	70.46
	Random-missing					
w/ S^2 Guidance	76.44	76.28	79.40	79.19	77.78	77.62
w/ $S+S$ Guidance	74.92	74.50	77.74	77.39	75.82	76.36
w/ Sentiment Guidance	74.20	74.01	77.36	77.08	75.63	75.41
w/ Semantic Guidance	75.67	75.45	78.04	77.89	76.13	76.04
w/o Guidance	73.35	73.26	76.84	76.33	74.76	74.58

increasing missingness, whereas the degradation of S^2Diff increases by only 2.26%. These results further verify the adaptability and stability of the proposed method under varying dataset scales and missing-rate intensities.

4.4 Ablation Study

Table 2 evaluates different guidance mechanisms in S^{2}Diff. The proposed S^2 guidance consistently achieves the best performance under both fixed-missing and random-missing protocols. Compared with no guidance, it improves the average score by 3.32 and 2.93 points, respectively, confirming the necessity of conditional guidance for missing-modality completion. Semantic guidance performs better than sentiment guidance, with average gains of 1.19 and 0.92 points under the two protocols, suggesting that semantic consistency is central to cross-modal reconstruction. Nevertheless, the full S^2 guidance further outperforms semantic-only guidance by 1.18 and 1.25 points, showing that sentiment guidance provides complementary affective constraints. Moreover, S^2 guidance surpasses $S+S$ guidance by 1.74 and 1.66 points, indicating that dual-branch modeling is more effective than simply concatenating semantic and sentiment conditions in a single branch. Additional loss-level ablations are reported in Appendix C.

4.5 Visualization Analysis

We visualize the recovered features under different guidance strategies using t-SNE, as shown in Fig. 4. The light purple points denote the ground-truth feature background, and the recovered features are colored by predicted sentiment labels. S^{2}Diff yields clearer separation between negative and positive samples while better matching the target feature distribution. In comparison, sentiment guidance mainly improves polarity separation but produces more scattered features, whereas semantic guidance better preserves the global feature structure but shows weaker sentiment separability in overlapping regions. These results indicate that semantic and sentiment guidance are complementary for robust missing-modality reconstruction. Additional visualization results under different missing-modality settings are provided in Appendix E.

5 Conclusion

This work proposes S^{2}Diff, a structured generative framework for missing-modality multimodal sentiment analysis. By decomposing modality representations into semantic and sentiment components and using two partially shared diffusion branches, S^{2}Diff mitigates observation-induced sentiment bias under cross-modal affective inconsistency. Experiments on MOSI, MOSEI, and IEMOCAP show consistent improvements under fixed- and random-missing protocols, and ablations confirm the complementarity of semantic and sentiment guidance.

6 Limitations

Despite its effectiveness, S^2Diff has several limitations. First, its semantic–sentiment decomposition and dual-branch diffusion formulation rely on an approximate factorization assumption. In natural multimodal utterances, semantic and affective information can be deeply entangled, and their interactions may be nonlinear. Thus, the projection-based decomposition and branch-wise score modeling may not fully separate semantic consistency from sentiment alignment in ambiguous cases. Second, our evaluation follows fixed- and random-missing protocols, while real-world missingness can be correlated with sentiment, speaker behavior, environment, or recording conditions, leading to missing-not-at-random biases beyond current settings. Third, reconstruction is performed in the representation space and optimized for sentiment prediction, so the recovered features are task-oriented rather than faithful modality reconstructions. Future work will explore more flexible disentanglement, more realistic missing-modality evaluation, and better balance between task utility and reconstruction fidelity.

References

Xin Chen, Gillian Dobbie, Xinyu Wang, Feng Liu, Di Wang, and Jingfeng Zhang. 2025. Robust learning of diffusion models with extremely noisy conditions. *ArXiv*, abs/2510.10149.

Ruiting Dai, Chenxi Li, Yandong Yan, Lisi Mo, Ke Qin, and Tao He. Unbiased missing-modality multimodal learning.

Prafulla Dhariwal and Alex Nichol. 2021. Diffusion models beat gans on image synthesis. *ArXiv*, abs/2105.05233.

Zirun Guo, Tao Jin, and Zhou Zhao. 2024. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. *ArXiv*, abs/2407.05374.

Xing Han, Huy Nguyen, Carl William Harris, Nhat Ho, and Suchi Saria. 2024. Fusemoe: Mixture-of-experts transformers for fleximodal fusion. *ArXiv*, abs/2402.03226.

Amir Hertz, Ron Mokady, Jay M. Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. *ArXiv*, abs/2208.01626.

Jonathan Ho. 2022. Classifier-free diffusion guidance. *ArXiv*, abs/2207.12598.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. 2022. Elucidating the design space of diffusion-based generative models. *ArXiv*, abs/2206.00364.

Jian Lang, Zhangtao Cheng, Ting Zhong, and Fan Zhou. 2025. Retrieval-augmented dynamic prompt tuning for incomplete multimodal learning. *ArXiv*, abs/2501.01120.

Mingcheng Li, Dingkang Yang, Yang Liu, Shunli Wang, Jiawei Chen, Shuai Wang, Jinjie Wei, Yue Jiang, Qingyao Xu, Xiaolu Hou, Mingyang Sun, Ziyun Qian, Dongliang Kou, and Lihua Zhang. 2024. Toward robust incomplete multimodal sentiment analysis via hierarchical representation learning. *ArXiv*, abs/2411.02793.

Xiaohao Liu, Xiaobo Xia, Jiaheng Wei, Shuo Yang, Xiu Su, See-Kiong Ng, and Tat-Seng Chua. 2025. Calibrated multimodal representation learning with missing modalities. *ArXiv*, abs/2511.12034.

Bjoern H Menze, András Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin S. Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth R. Gerstner, Marc-André Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, and 49 others. 2015. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34:1993–2024.

Chong Mou, Xintao Wang, Liangbin Xie, Jing Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. 2023. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI Conference on Artificial Intelligence*.

Srinivas Parthasarathy and Shiva Sundaram. 2020. Training strategies to handle missing modalities for audio-visual expression recognition. *Companion Publication of the 2020 International Conference on Multimodal Interaction*.

William S. Peebles and Saining Xie. 2022. Scalable diffusion models with transformers. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4172–4182.

Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *ArXiv*, abs/2205.11487.

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. *Proceedings of the conference. Association for Computational Linguistics. Meeting*, 2019:6558–6569.

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

Yuanzhi Wang, Yong Li, and Zhen Cui. 2023b. Incomplete multimodality-diffused emotion recognition. In *Neural Information Processing Systems*.

Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. 2024. Deep multimodal learning with missing modality: A survey.

Zehui Wu, Ziwei Gong, Jaywon Koo, and Julia Hirschberg. 2023. Multimodal multi-loss fusion network for sentiment analysis. In *North American Chapter of the Association for Computational Linguistics*.

Wenxin Xu, Hexin Jiang, and Xuefeng Liang. 2024. Leveraging knowledge of modality experts for incomplete multimodal learning. *Proceedings of the 32nd ACM International Conference on Multimedia*.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *AAAI Conference on Artificial Intelligence*.

Sukwon Yun, Inyoung Choi, Jie Peng, Yangfan Wu, Jingxuan Bao, Qiyiwen Zhang, Jiayi Xin, Qi Long, and Tianlong Chen. 2024. Flex-moe: Modeling arbitrary modality combination via the flexible mixture-of-experts. *ArXiv*, abs/2410.08245.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, E. Cambria, and Louis philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Annual Meeting of the Association for Computational Linguistics*.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31:82–88.

Changqing Zhang, Yajie Cui, Zongbo Han, Joey Tianyi Zhou, H. Fu, and Qinghua Hu. 2020. Deep partial multi-view learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:2402–2415.

Haoyu Zhang, Wenbin Wang, and Tianshu Yu. 2024. Towards robust multimodal sentiment analysis with incomplete data. *ArXiv*, abs/2409.20012.

Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. 2023a. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In *Conference on Empirical Methods in Natural Language Processing*.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023b. Adding conditional control to text-to-image diffusion models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3813–3824.

Zijian Zhang, Zhou Zhao, Jun Yu, and Qi Tian. 2023c. Shiftddpms: Exploring conditional diffusion models by shifting diffusion trajectories. In *AAAI Conference on Artificial Intelligence*.

Jinming Zhao, Ruichen Li, and Qin Jin. 2021. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Annual Meeting of the Association for Computational Linguistics*.

Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee Kenneth Wong. 2023. Uni-controlnet: All-in-one control to text-to-image diffusion models. *ArXiv*, abs/2305.16322.

Yihan Zhao, Wei Xi, Xiao Fu, and Jizhong Zhao. 2025. Enhancing multimodal model robustness under missing modalities via memory-driven prompt learning. In *International Joint Conference on Artificial Intelligence*.

Yu Zhou, Min Hu, and Xiaohua Wang. 2024. Robust multimodal sentiment analysis of image-text pairs by distribution-based feature recovery and fusion. *Proceedings of the 32nd ACM International Conference on Multimedia*.

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Ning An. 2025. Proxy-driven robust multimodal sentiment analysis with incomplete data. In *Annual Meeting of the Association for Computational Linguistics*.

A Experimental Setup

All models are implemented in PyTorch and trained on a single NVIDIA RTX 3090 GPU with 24GB memory. Experiments are conducted on MOSI, MOSEI, and IEMOCAP. For each dataset, we run all experiments with five random seeds: 1111, 2222, 3333, 4444, and 5555, and report the average results. For fair comparison, all methods use the same input features, data splits, missing-modality protocols, and evaluation metrics.

We adopt the Adam optimizer with a learning rate of 1×10^{-3} and set the feature embedding dimension to 32. For MOSI, the batch size is set to 32 with a dropout rate of 0.05. For MOSEI, the batch size is set to 64 with a dropout rate of 0.1. For IEMOCAP, the batch size is set to 32 with a dropout rate of 0.1. The multimodal fusion backbone is MULT (Tsai et al., 2019).

For each baseline, we follow the training strategy reported in its original paper. In particular, IMDer and MPLMM adopt a training protocol similar to ours: they are first trained under the full-modality setting and then further optimized from the full-modality checkpoint for each fixed-missing or random-missing configuration. For S^2Diff, we also first train the model under the full-modality setting, corresponding to $\mathrm{MR} = 0$, and then initialize each missing-modality training run from the $\mathrm{MR} = 0$ checkpoint. This strategy provides a stable full-modality representation space and allows the completion module to focus on reconstructing missing modality features under different missing patterns.

For the proposed S^2Diff, the total objective consists of five terms: the task loss, semantic alignment loss, sentiment alignment loss, sentiment transformation supervision loss, and diffusion loss. Their weights are set to $\lambda_{\text{task}} = 1,\ \lambda_{\text{sem}} = 0.5,\ \lambda_{\text{sent}} = 0.2,\ \lambda_{\text{trans}} = 0.2,\ \lambda_{\text{diff}} = 0.5$. These weights are kept fixed across all datasets and missing-modality settings.

B Computational Cost and Model Size

We further compare the computational cost and parameter size of S^2Diff with two simplified variants. As shown in Table 3, S^2Diff introduces moderate training overhead compared with the no-branch variant, increasing training time by 0.08h, 0.17h, and 0.24h on MOSI, MOSEI, and IEMOCAP, respectively. The additional testing cost is negligible, with at most 0.01h increase. Table 4 shows that S^2Diff adds 1.89M parameters on MOSI/MOSEI and 1.75M parameters on IEMOCAP compared with the no-branch variant. These results indicate that the proposed dual-branch design improves robustness with limited computational and parameter overhead.

Table 3: Computational cost comparison of S^2Diff and its simplified variants. Training and testing time are reported on MOSI, MOSEI, and IEMOCAP.

Methods	MOSI		MOSEI		IEMOCAP	
	train	eval	train	eval	train	eval
S2Diff	0.62h	0.01h	2.01h	0.05h	2.20h	0.04h
Single-Guidance	0.58h	0.01h	1.92h	0.04h	2.13h	0.03h
No-Guidance	0.54h	0.01h	1.84h	0.04h	1.96h	0.03h

Table 4: Parameter size comparison of S^2Diff and its simplified variants. The number of trainable parameters is reported in millions.

Methods	MOSI/MOSEI	IEMOCAP
S2Diff	112.89M	59.81M
Single-Guidance	111.48M	58.45M
No-Guidance	111.00M	58.06M

C Analysis of Multi-objective Training Objectives

To examine the contribution of different training objectives, we remove each loss term from the full S^2Diff objective and report the results in Table 5. The average scores are computed over all reported metrics on MOSI, MOSEI, and IEMOCAP under each missing-modality protocol.

Overall, the full objective achieves the best performance under both fixed-missing and random-missing settings, indicating that the losses provide complementary supervision. The diffusion loss $\mathcal{L}_{\text{diff}}$ is the basic generative objective. Removing it causes the largest degradation, reducing the average score from 74.21 to 49.83 under fixed missing and from 73.57 to 52.38 under random missing. This confirms that diffusion supervision is essential for learning the target-modality feature distribution.

Beyond the generative objective, semantic and sentiment losses contribute most to representation quality. Removing $\mathcal{L}_{\text{sem}}$ leads to average drops of 8.97 and 9.08 points under fixed and random missing, respectively, showing the importance of cross-modal semantic consistency. Removing $\mathcal{L}_{\text{sent}}$ causes slightly larger drops of 9.62 and 10.33 points, suggesting that sentiment supervision is crucial for preserving discriminative affective information and reducing sentiment bias.

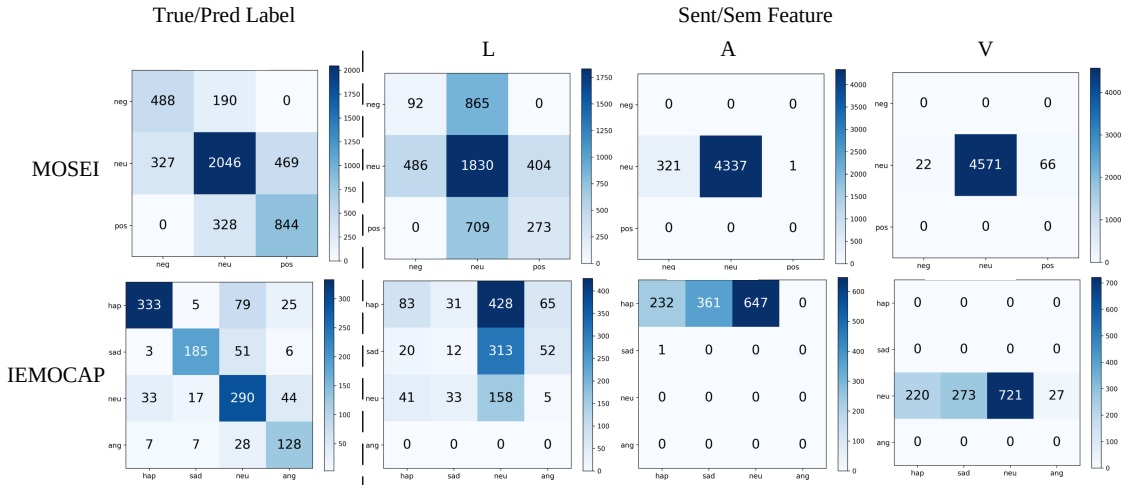

Figure 5: Confusion matrix analysis of final predictions and decomposed features. The left column shows ground-truth/predicted label confusion matrices, while the remaining columns probe sentiment-label distributions of decomposed sentiment/semantic features from language (L), acoustic (A), and visual (V) modalities. MOSEI scores are evenly divided into negative, neutral, and positive classes. Language features show richer affective patterns, with sentiment features being more polarized and semantic features more conservative.

The task loss $\mathcal{L}_{\text{task}}$ yields moderate but consistent gains, with average drops of 3.63 and 2.30 points when removed. This indicates that reconstructed features should remain aligned with downstream sentiment prediction. In contrast, $\mathcal{L}_{\text{trans}}$ has a smaller but stable effect, with drops of 1.12 and 0.94 points, suggesting that sentiment transformation supervision mainly acts as a calibration term for target-modality sentiment prompts. These results show a clear hierarchy among the objectives: $\mathcal{L}_{\text{diff}}$ provides the reconstruction basis, $\mathcal{L}_{\text{sem}}$ and $\mathcal{L}_{\text{sent}}$ control semantic and affective quality, $\mathcal{L}_{\text{task}}$ ensures task relevance, and $\mathcal{L}_{\text{trans}}$ refines cross-modal sentiment transfer.

Table 5: Ablation study of training objectives in S^{2}Diff. We remove each loss term from the full objective and report performance under fixed-missing and random-missing protocols.

Methods	MOSI		MOSEI		IEMOCAP	
	ACC	F1	ACC	F1	UA	WA
	fixed-missing					
S2Diff	**71.94**	**71.71**	**76.78**	**75.51**	**74.79**	**74.54**
w/o task	69.21	69.12	73.52	72.17	70.64	68.83
w/o sem	63.49	62.89	69.54	69.47	63.61	62.48
w/o sent	63.27	63.13	68.81	68.09	62.37	61.88
w/o trans	70.67	70.43	75.95	75.21	73.44	72.86
w/o diff	52.25	51.42	51.07	50.935	48.02	45.26
	random-missing					
S2Diff	**70.24**	**69.16**	**75.54**	**75.38**	**75.45**	**75.64**
w/o task	68.34	67.98	73.57	72.36	73.28	72.11
w/o sem	63.25	63.02	67.48	66.91	63.49	62.81
w/o sent	62.67	61.24	66.47	65.39	62.3	61.39
w/o trans	69.48	69.27	74.17	74.28	74.56	74.02
w/o diff	53.14	51.16	54.44	52.47	52.85	50.191

D Confusion Matrix Analysis

We provide a diagnostic confusion-matrix analysis in Fig. 5. The left column reports the confusion matrices between ground-truth and predicted labels, while the remaining columns show the label distributions predicted from decomposed semantic and sentiment features of different modalities. For MOSEI, sentiment scores are divided into negative, neutral, and positive categories for clearer visualization.

The prediction-level confusion matrices show that most samples are concentrated around the diagonal, indicating that the reconstructed multimodal representations remain discriminative for downstream prediction. On MOSEI, errors mainly occur between adjacent sentiment categories, while direct confusion between negative and positive samples is relatively rare. On IEMOCAP, misclassifications tend to appear among affectively related emotions, reflecting the inherent ambiguity of conversational emotion recognition.

The feature-level matrices reveal modality-dependent affective response patterns. Language features cover richer sentiment/emotion categories, suggesting that language provides more explicit affective cues. In contrast, acoustic and visual features are more concentrated in fewer categories, indicating relatively coarser affective patterns. More-

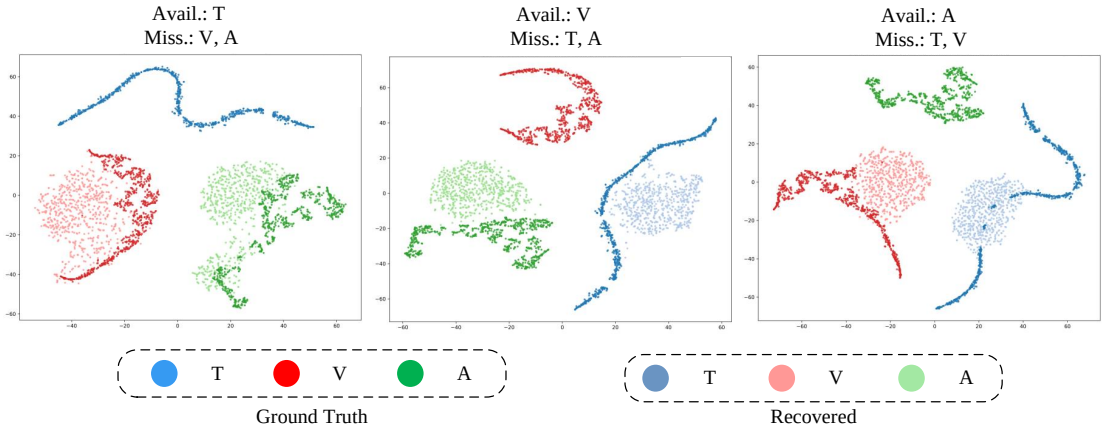

Figure 6: Visualization of feature generation with a single observed modality. When only text (T), only vision (V), or only audio (A) is available, the figure compares the generated missing-modality features (Recovered) with the corresponding ground-truth features (Ground Truth) in a low-dimensional space.

over, sentiment components generally produce more polarized predictions than semantic components, while semantic components tend to be more conservative and less category-specific. Since existing benchmarks do not provide modality-specific sentiment annotations, this analysis should be viewed as diagnostic rather than direct evidence of sentiment drift. Nevertheless, these observations are consistent with our motivation that semantic and sentiment factors exhibit different properties and that structured semantic–sentiment conditioning is beneficial for missing-modality reconstruction.

E Additional Missing-modality Visualizations

To assess imputation under extreme missing scenarios, we use t-SNE to visualize generated missing-modality features when only one modality is observed, as shown in Fig. 6. The recovered features generally cluster around the ground-truth features but still show moderate distribution shift. Compared with the more fine-grained manifolds of ground-truth features, the recovered features are smoother and less locally structured. This may result from multi-objective optimization, where generation tends to favor the coarse-grained task objective, yielding simpler geometric patterns in the embedding space.