TMDC: A Two-Stage Modality Denoising and Complementation Framework for Multimodal Sentiment Analysis with Missing and Noisy Modalities

Yan Zhuang[*1], **Minhao Liu**[*1,2], **Yanru Zhang** [1,2], **Jiawen Deng**[1†], **Fuji Ren**[1,2†]

[1]University of Electronic Science and Technology of China
[2]Shenzhen Institute for Advanced Study, UESTC
202211081370@std.uestc.edu.cn, {minhaoliu,yanruzhang,dengjw,renfuji}@uestc.edu.cn

Abstract

Multimodal Sentiment Analysis (MSA) aims to infer human sentiment by integrating information from multiple modalities such as text, audio, and video. In real-world scenarios, however, the presence of missing modalities and noisy signals significantly hinders the robustness and accuracy of existing models. While prior works have made progress on these issues, they are typically addressed in isolation, limiting overall effectiveness in practical settings. To jointly mitigate the challenges posed by missing and noisy modalities, we propose a framework called **Two-stage Modality Denoising and Complementation (TMDC)**. TMDC comprises two sequential training stages. In the Intra-Modality Denoising Stage, denoised modality-specific and modality-shared representations are extracted from complete data using dedicated denoising modules, reducing the impact of noise and enhancing representational robustness. In the Inter-Modality Complementation Stage, these representations are leveraged to compensate for missing modalities, thereby enriching the available information and further improving robustness. Extensive evaluations on MOSI, MOSEI, and IEMOCAP demonstrate that TMDC consistently achieves superior performance compared to existing methods, establishing new state-of-the-art results.

Code — https://github.com/YetZzzzzz/TMDC
Extended version — https://arxiv.org/abs/2511.10325

Introduction

Multimodal Sentiment Analysis (MSA) leverages data from multiple modalities, such as text, video, and audio, to predict the emotional state (Zadeh et al. 2016). With the rapid advancements in multimedia and multimodal learning (Zadeh et al. 2018; Liang, Zadeh, and Morency 2022), more and more researchers are focusing on MSA. However, practical applications often face two key challenges: missing modalities, which is caused by privacy concerns (Jaiswal and Provost 2020; Zhao, Li, and Jin 2021) or incomplete data collection (Liu et al. 2021), and noisy inputs from real-world sensors. These factors severely hinder the reliability and effectiveness of MSA systems.

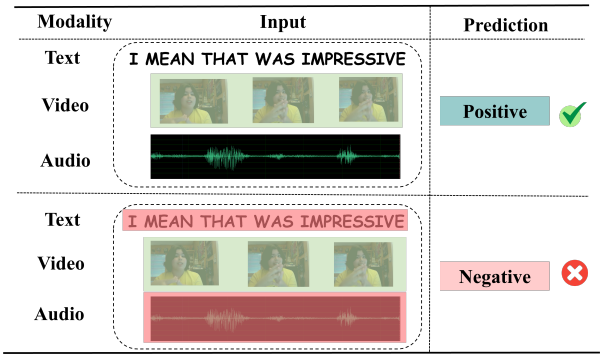


Figure 1: The existing model yields correct predictions when the input contains only missing modalities (highlighted in green), but fails when both missing (green) and noisy modalities (red) are present.

To tackle these challenges, most existing studies treat missing and noisy modalities as separate problems and develop isolated solutions. For noisy inputs, many works adopt information-theoretic approaches, such as the use of the information bottleneck (Mai, Zeng, and Hu 2022), to suppress irrelevant noise. On the other hand, methods targeting missing modalities often design reconstruction mechanisms to restore missing signals from available ones. For example, MPLMM (Guo, Jin, and Zhao 2024) distills knowledge from pre-trained models into learnable prompts to compensate for missing data. IMDer (Wang, Li, and Cui 2023) trains diffusion models on complete datasets and applies them at inference time to reconstruct absent modalities. DiCMoR (Wang, Cui, and Li 2023) creates category-specific flow-based generators for restoration, while MoMKE (Xu, Jiang, and Liang 2024) trains multiple modality-specific experts and combines their outputs to form joint representations. Despite their advances, these methods typically assume clean inputs and overlook the impact of noisy data. As illustrated in Figure 1, they often perform well under missing-modality conditions but fail when noise and missing data co-occur. Errors from noisy inputs or inaccurate reconstruction compound during training and inference, ultimately harming overall performance.

To jointly address both challenges, we propose a Two-

[*]These authors contributed equally.
[†]Corresponding authors.

stage Modality Denoising and Complementation (TMDC) framework. TMDC adopts a two-stage training paradigm. In the Intra-Modality Denoising Stage, TMDC is trained on complete data to capture both denoised modality-specific and modality-common representations for each modality. To handle noise, TMDC includes two denoising modules. A modality-specific denoising module, composed of a Variational Information Bottleneck (VIB) (Alemi et al. 2022), attention layers (Vaswani et al. 2017), and fully connected networks, filters noise while preserving distinctive features of each modality. A modality-common module learns to extract noise-robust features shared across modalities. In the Inter-Modality Complementation Stage, the denoised outputs from available modalities are used to complement missing ones by leveraging both shared and specific information extracted during the first stage. The final emotion prediction is made by integrating all observed and reconstructed representations through a fully connected layer.

The key contributions of this paper are as follows:

- TMDC, a framework that simultaneously addresses both missing and noisy modalities is proposed.
- TMDC complements missing modalities using both modality-invariant and modality-specific information from available modalities.
- Extensive experiments on multiple datasets, including scenarios with varying levels of data noise, demonstrates that TMDC achieves state-of-the-art performance.

Related Work

Multimodal Sentiment Analysis

Multimodal Sentiment Analysis (MSA) integrates text, video, and audio signals to predict sentiment state (Zadeh et al. 2016). Most existing methods assume the availability of all three modalities and focus on designing sophisticated fusion networks to integrate these heterogeneous sources effectively (Zong et al. 2023; Hazarika, Zimmermann, and Poria 2020; Lin and Hu 2023; Wang et al. 2022; Mai et al. 2023; Zeng et al. 2023; Zhu et al. 2024; Zhuang et al. 2025b). Some approaches prioritize modality alignment. For instance, MISA (Hazarika, Zimmermann, and Poria 2020) and FactorCL (Liang et al. 2024) map each modality into modality-shared and modality-specific representations, or task-relevant and task-irrelevant components. Similarly, MULT (Tsai et al. 2019) and AcFormer (Zong et al. 2023) employ cross-modal attention mechanisms to align pairs of modalities. Other methods refine fusion at a more granular level. GLoMo (Zhuang et al. 2024) utilizes MoE networks to extract fine-grained local information for enhanced integration, while PS2RI (Fang et al. 2024) incorporates sarcasm-aware cues to aid sentiment prediction. KEBR (Zhu et al. 2024), on the other side, explores common sentimental knowledge in unlabeled videos to enrich representation. Although these models perform well when all modalities are available, their effectiveness deteriorates significantly when one or more modalities are missing, a common scenario in real-world applications, which limits their practical usability (Guo, Jin, and Zhao 2024; Li et al. 2024b,a; Wei, Luo, and Luo 2023).

Incomplete Multimodal Learning

With the growing demand for robust multimodal models, researchers have increasingly focused on incomplete multimodal learning. Most existing methods attempt to reconstruct the missing modality using pre-trained external knowledge. For example, MMIN (Zhao, Li, and Jin 2021) involves pre-training on complete datasets before transferring or fine-tuning in situations with missing modalities. MPLMM (Guo, Jin, and Zhao 2024) leverages external datasets and distills cross-modal information into learnable prompts to supplement missing data. IMDer (Wang, Li, and Cui 2023) train diffusion models on complete datasets for each modality, later using them for reconstruction when a modality is missing. Similarly, DiCMoR (Wang, Cui, and Li 2023) employs category-specific flow-based models to restore absent modalities. MoMKE (Xu, Jiang, and Liang 2024) takes a different approach by training modality-specific MoE networks on complete datasets to capture modality-specific representations and then fusing them into a joint representation.

While these methods have achieved notable success, they assume that the collected data is noise-free. However, real-world multimodal data often contains inherent noise (Mai, Zeng, and Hu 2022; Gao et al. 2024), and errors introduced during reconstruction further degrade performance. This reduces the robustness of these models, particularly in noisy environments. To address this limitation, we propose TMDC, a framework that explicitly considers both intrinsic noise and errors introduced by missing modalities. By employing a denoising-first learning approach, TMDC enhances representation robustness, making it more suitable for real-world applications.

Methodology

Problem Definition

Given a dataset $D = \{X_i^A, X_i^T, X_i^V\}_{i=1}^N$ comprising three modalities (e.g. text ('T'), video ('V'), and audio ('A')), each element X_i^m represents the representation of modality m for i^{th} instance, $m \in \{A, T, V\}$. Specifically, $X_i^m \in \mathcal{R}^{L_m \times D_m}$, where L_m denotes the sequence length, and D_m represents the feature dimension of modality m. For simplicity, we omit the subscript i and use X^m to denote the representation of modality m. In real-world scenarios, some modalities may be missing due to various factors. To denote missing modalities, we use $\hat{X^m}$. Additionally, modality representations inherently contain noise, which we do not explicitly annotate. The goal of MSA with missing and noisy modalities is to train effective and robust models under different missing-modality and noisy-modality conditions.

Preliminaries

Given a modality representation X^m with inherent noise, VIB approximates the information bottleneck by learning a compressed encoding X_s^m that retains essential task-relevant information while filtering out unnecessary information. The optimization objective is formulated as:

$$\mathcal{L}^m = I(X_s^m, y) - \beta I(X_s^m, X^m). \tag{1}$$

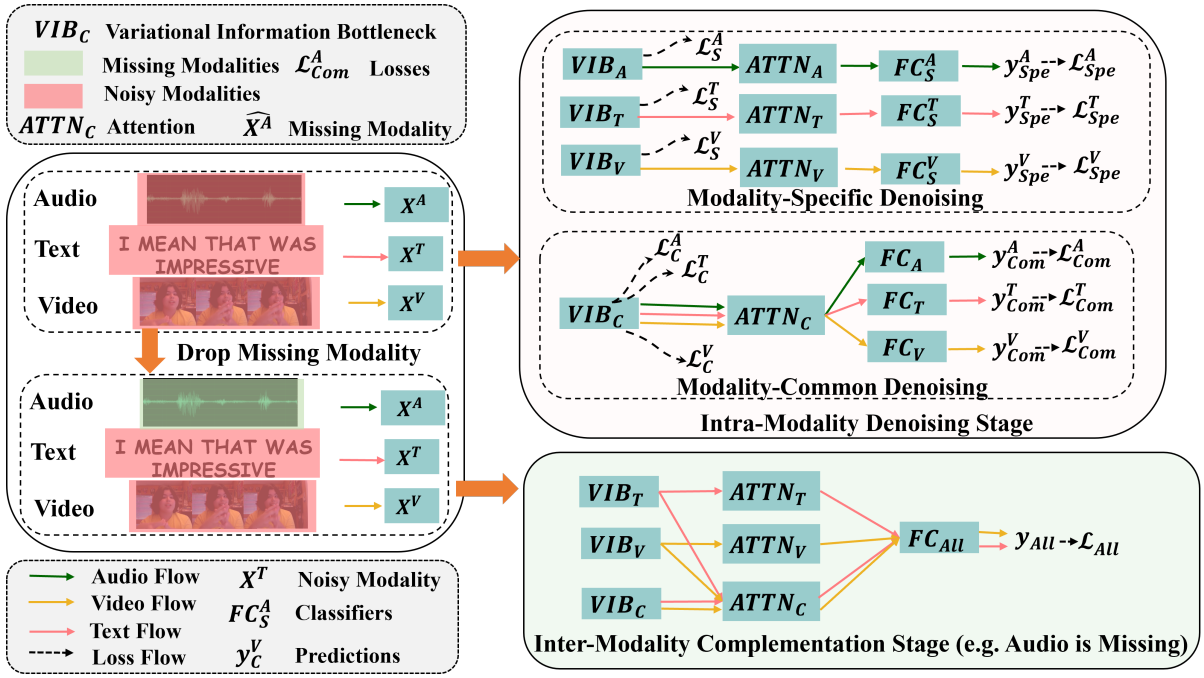


Figure 2: Illustration of the proposed TMDC. TMDC includes two training stages. In the first stage, TMDC learns from complete modality information using two denoising modules. The modality-specific denoising module applies separate networks to each modality to remove noise while preserving unique modality information. Simultaneously, the modality-common denoising module employs a shared network to filter noise across multiple modalities and extract common information. In the second stage, the learned shared information is used to supplement missing modalities.

Here y is the ground truth, $I(\cdot,\cdot)$ denotes mutual information, measuring the correlation between representations. A higher mutual information value indicates stronger relevance. X_s^m represents the denoised modality representation, and β is a Lagrange multiplier.

The mutual information $I(X_s^m, y)$ can be implemented using task-related loss $\mathcal{L}_{TASK}(y^m, y)$ and $I(X_s^m, X^m)$ can be approximated by employing KL divergence, leading to the revised objective:

$$\mathcal{L}^m = \mathcal{L}_{TASK}(y^m, y) + \beta KL(p(e_s^m|e^m)||\mathcal{N}(0, \mathbf{I})). \quad (2)$$

Here $e_s^m \in X_s^m$, $e^m \in X^m$, $p(e_s^m|e^m) \sim \mathcal{N}(\mu_s^m, (\sigma_s^m)^2\mathbf{I})$, and μ_s^m, σ_s^m denote the mean and variance of the denoised representation, which are predicted using fully connected layers through:

$$\mu_s^m = W_1^m e^m + b_1^m, \quad (3)$$

and:

$$\sigma_s^m = W_2^m e^m + b_2^m. \quad (4)$$

Following the reparameterization trick (Kingma 2013) as in (Mai, Zeng, and Hu 2022; Gao et al. 2024; Alemi et al. 2022), denoised representation can be constructed as:

$$X_s^m = \mu_s^m + \epsilon\sigma_s^m. \quad (5)$$

Here $\epsilon \in \mathcal{N}(0, \mathbf{I})$. y^m is the predicted label using denoised representation X_s^m through a linear layer:

$$y_s^m = W_3^m X_s^m + b_3^m. \quad (6)$$

Here W_1^m, W_2^m, W_3^m are trainable weights, and b_1^m, b_2^m, b_3^m are bias terms.

Model Overview

We propose TMDC, a two-stage modality denoising and complementation framework to handle missing and noisy modalities in MSA. As illustrated in Figure 2, TMDC consists of the Intra-Modality Denoising (IMD) Stage and the Inter-Modality Complementation (IMC) Stage. In the IMD Stage, TMDC learns denoised modality-specific and modality-invariant representations from complete multimodal data. The Modality-Specific Denoising (MSD) Module, which incorporates modality-specific VIB (Alemi et al. 2022) and attention layers (Vaswani et al. 2017), extracts modality-specific representations. Meanwhile, the Modality-Common Denoising (MCD) Module, utilizing shared VIB and attention layers, captures shared representations across modalities. While in the IMC Stage, TMDC processes incomplete data with missing modalities. The available modality-specific and modality-invariant representations are leveraged to complement missing information. The final integrated representation is then passed through a fully connected layer for sentiment prediction. The following sections provide a detailed introduction of each component and process within TMDC.

Intra-Modality Denoising Stage

This section details the first stage of TMDC. Following prior works (Wang, Li, and Cui 2023; Wang, Cui, and Li 2023; Guo, Jin, and Zhao 2024; Xu, Jiang, and Liang 2024),

TMDC is initially trained on the complete dataset to obtain comprehensive information from all modalities. However, unlike existing approaches that overlook inherent noise during training, TMDC explicitly addresses this issue by introducing two modules: the Modality-Specific Denoising (MSD) Module and the Modality-Common Denoising (MCD) Module. These modules separately extract modality-specific and shared information while filtering out noise. Below, we describe each module in detail.

Since different modalities have different dimensions, following existing studies(Tsai et al. 2019), we use the 1D temporal convolutional layer (Conv1D) with a kernel size of 3×3 to standardize each modality to the same dimension (D) and sequence length (T) through:

$$X^m = W^m_{3\times3}(X^m). \tag{7}$$

Here $X^m \in \mathcal{R}^{T\times D}$, and $W^m_{3\times3}$ is the trainable weights.

Modality-Specific Denoising Module. This module aims to reduce noise within each modality and extract modality-specific representations. To achieve this, TMDC employs a distinct yet structurally identical network for each modality. Inspired by prior methods (Mai, Zeng, and Hu 2022; Gao et al. 2024), we adopt VIB (Alemi et al. 2022) to simultaneously remove noise and redundant information, which is introduced in Preliminaries Section. Specifically, for each modality m, we use the VIB to obtain denoised representations X^m_s and get the predicted label y^m_s using Equations 3-6. Once the denoised representations are obtained, they are processed through a modality-specific attention network, which consists of a multi-head attention (MHA) layer and a residual fully connected layer, enabling interactions among different modality representations through:

$$X^m_{Spe} = MHA^m(X^m_s, X^m_s) + X^m_s, \tag{8}$$

and:

$$\hat{X}^m_{Spe} = X^m_{Spe} + W^m_4 X^m_{Spe} + b^m_4. \tag{9}$$

To ensure these representations retain modality-specific information, a fully connected layer predicts the label through:

$$y^m_{Spe} = W^m_5 \hat{X}^m_{Spe} + b^m_5. \tag{10}$$

Notably, each modality has its own independent network, as illustrated in Figure 2, and the network parameters are not shared across modalities.

Modality-Common Denoising Module. This module extracts denoised, modality-invariant features shared across all modalities. Its architecture is similar to the MSD Module, with a key difference: the parameters in Conv1D, VIB and Attention layers are shared, as shown in Figure 2.

Specifically, denoised modality-invariant representation X^m_c for each modality m is obtained using Equations 3-5. And the representation after attention interaction is obtained, denoted as $\hat{X}^m_{Com}$ for modality m, by changing query and key to X^m_c in Equation 8 and changing X^m_{Spe} to X^m_{Com} in Equation 9. The predicted label y^m_c from the denoised modality-invariant representation X^m_c, and y^m_{Com} from the interaction representation $\hat{X}^m_{Com}$ for each modality m are then obtained through separate linear layers.

Inter-Modality Complementation Stage

To simulate real-world conditions where certain modalities may be unavailable, we randomly set the representation of the missing modality to a zero vector during training. Without loss of generality, we assume the audio modality (A) is missing in this case (as illustrated in Figure 2), resulting in model input of the form $\{\hat{X}^A, X^T, X^V\}$.

To compensate for the missing information, we leverage the available modalities through a structured fusion strategy. For the text and video modalities, we first extract two types of representations using VIB: (1) modality-specific representations X^T_s and X^V_s; and (2) modality-invariant representations X^T_c and X^V_c using Equations 3-5. To enhance the expressiveness of uni-modal representations, we integrate both type of representations with modality-specific attention layer trained in the first stage, where X^{m1}_s serves as the query, and X^{m1}_c acts as the key and value through:

$$X^{m1}_{All} = MHA^{m1}(X^{m1}_s, X^{m1}_c), \tag{11}$$

and:

$$\hat{X}^{m1}_{All} = X^{m1}_{All} + W^{m1}_{All} X^{m1}_{All} + b^{m1}_{All}. \tag{12}$$

Here $m1 \in \{T, V\}$.

To further compensate for the missing modality, we model cross-modal dependencies between the available modalities using a bidirectional attention mechanism. Specifically, we extract complementary features by swapping the roles of the query and key in the attention layers in Equation 11: X_{T2V} is obtained by setting the query to X^T_c and the key/value to X^V_s, and vice versa for X_{V2T}. The corresponding enhanced features $\hat{X}_{T2V}$ and $\hat{X}_{V2T}$ are computed using the same transformation as in Equation 12.

We then fuse the cross-modal features by adding $\hat{X}_{T2V}$ and $\hat{X}_{V2T}$ to produce a compensated representation $X_{Compensate}$. The final multimodal representation is constructed by concatenating the compensated features with the refined text and video representations, followed by a fully connected layer for sentiment prediction:

$$X = [X_{Compensate}, \hat{X}^T_{All}, \hat{X}^V_{All}], \tag{13}$$

$$y_{All} = W_{All} X + b_{All}. \tag{14}$$

In cases where multiple modalities are missing, for example, when only the audio modality is available, a different compensation strategy is applied. The available modality undergoes the same refinement as described above. For the missing modalities, we approximate their representations using self-attention within the available modality by changing query and key to X^A_c and X^A_s in Equation 11 and transform the obtained representation to get $\hat{X}_{A2A}$ through Equation 12. Finally, the multimodal representation in this scenario is formed by repeating the compensated audio representation:

$$X = [\hat{X}^A_{All}, \hat{X}_{A2A}, \hat{X}_{A2A}]. \tag{15}$$

Training Objectives

This section outlines the training objectives for both IMD and IMC stages in TMDC framework. Since MSA involves

both classification and regression tasks, we first introduce the task-specific loss function, denoted as $\mathcal{L}_{TASK}$. For regression tasks, we use Mean Squared Error (MSE) loss, while for classification tasks, the Cross-Entropy loss is applied.

In the IMD Stage, TMDC is trained on the complete dataset to learn both modality-specific and modality-invariant representations. The training objective includes two components: (1) VIB loss using Equation 2 to encourage denoised representation learning; and (2) task-specific losses applied to representations after modality interactions. In order to distinguish different task-specific losses, we define the loss of different representations as:

$$\mathcal{L}_b^m = \mathcal{L}_{TASK}(y_b^m, y). \tag{16}$$

Here $b \in \{Spe, Com\}$ denotes the predictions, and m denotes the modality. The overall loss function for IMD stage is formulated as:

$$\mathcal{L}_{IMD} = \sum_{m\in\{A,T,V\}} (\sum_{b\in\{Spe,Com\}} \mathcal{L}_b^m + \sum_{k\in\{s,c\}} \mathcal{L}_k^m). \tag{17}$$

Here k denotes the loss from the modality-specific (s) or modality-invariant (c) representations using Equation 2 in MSD or MCD module.

In the IMC Stage, TMDC is optimized based only on the final fused representation. Here, the training objective focuses solely on the prediction loss for the concatenated multimodal representation:

$$\mathcal{L}_{IMC} = \mathcal{L}_{TASK}(y_{All}, y). \tag{18}$$

Experiment

Datasets and Evaluation Criteria

We evaluate TMDC on three benchmark datasets: MOSI (Zadeh et al. 2016), MOSEI (Zadeh et al. 2018), and IEMOCAP (Busso et al. 2008). MOSI and MOSEI are regression-based sentiment datasets with scores from -3 to +3. MOSI includes 2,199 video clips, while MOSEI has 22,856 clips. Following prior work (Xu, Jiang, and Liang 2024; Wang, Li, and Cui 2023; Guo, Jin, and Zhao 2024), we convert scores into binary labels and report accuracy and F1-score. IEMOCAP contains 5,531 utterances across five sessions. We follow common practice (Xu, Jiang, and Liang 2024) to classify four emotions: neutral, happy, sad, and angry. Performance is measured using weighted accuracy (WA) and unweighted accuracy (UA) under five-fold cross-validation.

Experiment Setups

Feature Extraction. Following previous studies (Lian et al. 2023; Xu, Jiang, and Liang 2024), we adopt the same feature extraction methods for all three datasets. For the text modality, we use the pre-trained DeBERTa-large model (He et al. 2021) to obtain textual representations. For the audio modality, we extract features using the pre-trained wav2vec-large model (Schneider et al. 2019). For the video modality, we apply the MTCNN face detection algorithm (Zhang et al. 2016) followed by the pre-trained MA-Net model (Zhao, Liu, and Wang 2021) to obtain video representations. The final feature dimensions for the text, audio, and video modalities are 1024, 512, and 1024, respectively.

Implementation Details. Consistent with prior work (Xu, Jiang, and Liang 2024; Guo, Jin, and Zhao 2024), we evaluate the performance of TMDC under fixed missing modality scenarios, where a specific modality is absent during training, validation, and testing. For example, in Table 1, 'A' indicates that only the audio modality is available. All experiments are implemented in PyTorch and conducted on a GTX 3090 GPU with CUDA 11.5. We use Torch version 1.12.1 for model training and Adam optimizer (Kingma and Ba 2014) across all datasets. More implementation details are shown in Extended Version (Zhuang et al. 2025a).

Comparison with State-of-the-art Methods

To comprehensively evaluate TMDC's performance, we compare it with several state-of-the-art methods, including MCTN (Pham et al. 2019), MMIN (Zhao, Li, and Jin 2021), GCNet (Lian et al. 2023), IMDer (Wang, Li, and Cui 2023), DiCMoR (Wang, Cui, and Li 2023), MPLMM (Guo, Jin, and Zhao 2024), MoMKE (Xu, Jiang, and Liang 2024), IF-MMIN (Zuo et al. 2023), and MRAN (Luo, Xu, and Lai 2023). Table 1 presents the results on the MOSI, MOSEI, and IEMOCAP datasets.

Our findings show that TMDC outperforms all baseline models in most scenarios, with only one exception: when only the audio modality is available in the MOSI dataset. However, in all other MOSI settings, TMDC surpasses existing methods. Moreover, TMDC consistently achieves state-of-the-art results across all missing-modality scenarios in the MOSEI and IEMOCAP datasets. Specifically, on MOSI, TMDC achieves an average accuracy of 77.64 and an F1-score of 77.35, outperforming the second-best model, MoMKE, by 0.59 and 0.89, respectively. On MOSEI, TMDC attains 81.22 accuracy and 80.76 F1-score, improving upon the strongest baseline by 0.78 in both metrics. For IEMOCAP, TMDC achieves a WA of 73.77 and a UA of 73.64, exceeding the best competing method by 0.42 and 0.86, respectively. These results highlight TMDC's effectiveness in learning robust representations.

Ablation Study

To investigate how each module and stage contributes to TMDC's effectiveness, we conduct an ablation study by evaluating four TMDC variants: (1) 'w/o IMD': Excludes the Intra-Modality Denoising stage, meaning TMDC is not trained on complete datasets before transitioning to the IMC stage. (2) 'w/o IMC': Removes Inter-Modality Complementation, where representations from available modalities are directly concatenated without additional compensation after the IMD stage. (3) 'w/o MCD': Excludes the Modality-Common Denoising Module in both stages, preventing any compensation for missing information. (4) 'w/o MSD': Removes the Modality-Specific Denoising Module in both stages, relying only on shared modality-invariant information for predictions.

Table 2 presents the average results of TMDC and its four variants across all seven missing-modality scenarios.

Models	Modalities						
	A	T	V	A,V	A,T	T,V	T,A,V
Results on MOSI							
	ACC/F1	ACC/F1	ACC/F1	ACC/F1	ACC/F1	ACC/F1	ACC/F1
MCTN	56.10/54.50	79.10/79.20	55.00/54.40	57.50/57.40	81.00/81.00	81.10/81.20	81.40/81.50
MMIN	55.30/51.50	83.80/83.80	57.00/54.00	60.40/58.50	84.00/84.00	83.80/83.90	84.60/84.40
GCNet	56.10/54.50	83.70/83.60	56.10/55.70	62.00/61.90	84.50/84.40	84.30/84.20	85.20/85.10
IMDer	62.00/62.20	84.80/84.70	61.30/60.80	63.60/63.40	85.40/85.30	85.50/85.40	85.70/85.60
DiCMoR	60.50/60.80	84.50/84.40	62.20/60.20	64.00/63.50	85.50/85.50	85.50/85.40	85.70/85.60
MPLMM	62.71/**63.65**	80.12/80.31	63.12/63.74	65.02/65.41	80.76/81.09	81.12/81.19	-
MoMKE	**63.19**/58.61	86.59/86.52	63.35/63.34	64.04/64.66	87.20/87.17	87.04/87.00	87.96/87.89
TMDC	62.35/60.24	**87.35/87.27**	**64.63/64.82**	**65.40/65.60**	**87.50/87.45**	**87.96/87.87**	**88.26/88.19**
Results on MOSEI							
	ACC/F1	ACC/F1	ACC/F1	ACC/F1	ACC/F1	ACC/F1	ACC/F1
MCTN	62.70/54.50	82.60/82.80	62.60/57.10	63.70/62.70	83.50/83.30	83.20/83.20	84.20/84.20
MMIN	58.90/59.50	82.30/82.40	59.30/60.00	63.50/61.90	83.70/83.30	83.80/83.40	84.30/84.20
GCNet	60.20/60.30	83.00/83.20	61.90/61.60	64.10/57.20	84.30/84.40	84.30/84.40	85.20/85.10
IMDer	63.80/60.60	84.50/84.50	63.90/63.60	64.90/63.50	85.10/85.10	85.00/85.00	85.10/85.10
DiCMoR	62.90/60.40	84.20/84.30	63.60/63.60	65.20/64.40	85.00/84.90	84.90/84.90	85.10/85.10
MPLMM	67.33/68.71	79.12/79.17	67.29/69.40	68.21/69.91	80.45/80.43	80.11/80.13	-
MoMKE	72.56/71.03	86.46/86.43	70.12/70.23	73.34/71.82	86.68/86.61	86.79/86.69	87.12/87.03
TMDC	**73.64/72.23**	**86.87/86.82**	**71.60/70.65**	**74.13/73.41**	**87.15/87.14**	**87.48/87.43**	**87.67/87.62**
Results on IEMOCAP							
	WA/UA	WA/UA	WA/UA	WA/UA	WA/UA	WA/UA	WA/UA
MCTN	49.75/51.62	62.42/63.78	48.92/45.73	56.34/55.84	68.34/69.46	67.84/68.34	-
MMIN	56.58/59.00	66.57/68.02	52.52/51.60	63.99/65.43	72.94/75.14	72.67/73.61	-
IF-MMIN	55.03/53.20	67.02/68.20	51.97/50.41	65.33/66.52	74.05/75.44	72.68/73.62	-
MRAN	55.44/57.01	65.31/66.42	53.23/49.80	64.70/64.46	73.00/74.58	72.11/72.24	-
MoMKE	70.32/71.38	77.82/78.37	58.60/54.70	68.85/67.65	79.89/79.53	77.87/77.84	80.13/79.99
TMDC	**70.45/71.40**	**77.88/78.44**	**59.18/55.33**	**70.21/69.91**	**79.99/81.45**	**78.20/78.29**	**80.48/80.69**

Table 1: Performance comparison under different modality combinations on three datasets. Best performance is bold.

We observe that removing any stage results in performance degradation, but the effects of removing modules vary across datasets. Overall, removing both denoising modules leads to performance degradation, with 'w/o MSD' having a more pronounced impact than 'w/o MCD'. This may be because modality-specific information is more discriminative than modality-invariant information, as the latter can be inferred from other modalities, whereas the former cannot. Consequently, losing modality-specific information results in a greater drop in performance. Notably, the 'w/o IMC' variant suffers the most significant drop across all datasets. This is likely because, while the IMD stage has access to all modalities, the IMC stage encounters missing modalities without compensation, leading to substantial performance deterioration. More detailed ablation results for each missing modality are shown in Extended Version (Zhuang et al. 2025a).

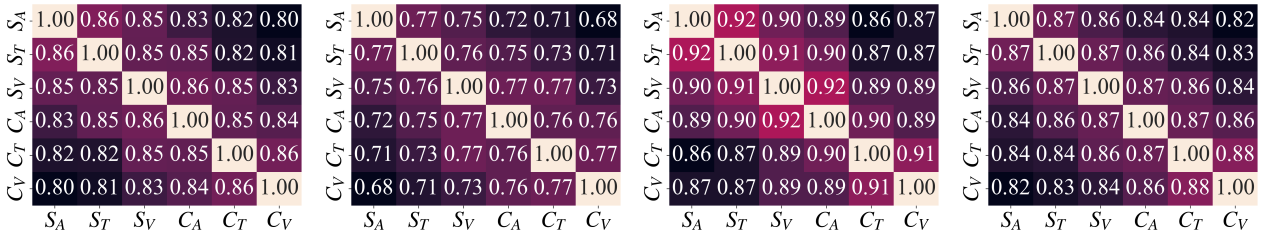
Further Analysis

In this section, we focus on TMDC's performance under noisy conditions, the training dynamics of different loss functions, and the relationships between representations under various missing modality scenarios. More results are shown in Extended Version (Zhuang et al. 2025a).

	MOSI	MOSEI	IEMOCAP
TMDC	**77.64/77.35**	**81.22/80.76**	**73.77/73.64**
w/o IMD	74.48/74.34	79.76/79.36	72.59/71.76
w/o IMC	74.17/73.97	79.74/79.41	66.05/65.94
w/o MCD	75.67/75.26	80.03/79.52	72.41/71.99
w/o MSD	74.76/74.16	80.03/79.92	71.68/71.24

Table 2: Ablation studies on different datasets. The averaged results of all seven conditions are reported. Best performance is bold.

Analysis of Experiments on Noisy Datasets. Here we introduce controlled noise to simulate realistic data collection challenges. Specifically, we add gaussian noise of varying intensities (Gao et al. 2024) to the existing modality representations across both training stages, as well as during validation and testing. Table 3 reports the average performance across seven missing modality scenarios. We observe a clear performance degradation as noise intensity increases. In IEMOCAP, TMDC's accuracy drops significantly from 73.8/73.6 to 37.1/32.2 when noise intensity reaches 20, while in MOSEI, the impact is less pronounced, with performance declining from 81.2/80.8 to 66.6/63.9. Despite the



(a) In IMD Stage (b) In IMC Stage with 'A,T,V' (c) In IMC Stage with 'A' (d) In IMC Stage with 'A,V'

Figure 3: Visualization of cosine similarity of representations on IEMOCAP dataset.

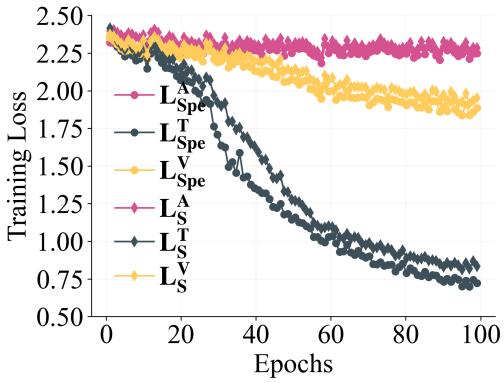


Figure 4: Losses in MSD Module

degradation, TMDC consistently outperforms MoMKE at all noise levels, particularly at a noise intensity of 10, where it surpasses MoMKE by an average of 10 points.

Methods	Gaussian Noise		
	$\epsilon = 5$	$\epsilon = 10$	$\epsilon = 20$
Results on Noisy MOSI			
MoMKE	55.9/55.8	53.9/54.1	52.7/52.9
TMDC	**68.4/67.8**	**60.8/60.2**	**53.0/53.0**
Results on Noisy MOSEI			
MoMKE	71.0/69.3	61.2/58.4	59.4/55.4
TMDC	**74.5/73.7**	**71.2/70.5**	**66.3/63.9**
Results on Noisy IEMOCAP			
MoMKE	46.1/44.2	34.4/30.3	31.1/28.3
TMDC	**60.0/57.9**	**51.0/48.5**	**37.1/32.2**

Table 3: Performance on noisy datasets. The averaged results of all seven conditions are reported.

Analysis of Training Convergence. To evaluate the training stability and convergence, we track the behavior of loss terms over 100 epochs on the MOSI dataset in Figure 4. The results show a clear correlation between the decline rate of each loss and the performance of its corresponding unimodal representation. As reported in Table 1, the text modality outperforms others on MOSI, while the audio modality performs worst. Consistently, text-related loss terms (e.g., $\mathcal{L}_S^T$ and $\mathcal{L}_{Spe}^T$) converge more rapidly than those of the audio modality, with the video modality falling in between. We also observe that the VIB-related losses are slightly higher than the interaction-based ones, possibly due to inherent noise in the data. Nonetheless, all losses consistently decrease over time, confirming the effectiveness and stability of the optimization process.

Analysis of Representations' Relations. TMDC learns both modality-specific and modality-invariant representations during training. To analyze their relationships under different settings, we visualize average cosine similarities on the IEMOCAP test set in Figure 3, where $S_m = \hat{X}_{Spe}^m$ and $C_m = \hat{X}_{Com}^m$. We consider four cases: (a) the IMD stage, the IMC stage (b) with all modalities, (c) with only audio available, and (d) with audio and video available.

Across most settings, S_m are consistently more similar to each other than to C_m, suggesting they capture complementary information. This relationship remains stable under different missing conditions, demonstrating the robustness of the representations. We also observe that similarity scores decrease as more modalities are available. This indicates that richer modality input encourages more diverse representation learning. Notably, when only audio is present, the inferred text and video representations differ from the audio, showing that the IMC module generates distinct and informative approximations for missing modalities.

Conclusion

This paper presents TMDC, a Two-stage Modality Denoising and Complementation framework designed to address the challenges of noisy and missing modalities in MSA. TMDC operates in two stages: the first stage reduces noise by learning both modality-specific and modality-invariant features from complete data; the second stage enhances representations by leveraging available modalities to complement the missing ones during training on incomplete data. Extensive experiments on three benchmark datasets under seven missing modality scenarios and high-noise conditions demonstrate that TMDC consistently outperforms existing methods. While the framework achieves strong performance, it introduces some redundancy in the shared representations. Future work will aim to reduce this redundancy to further improve efficiency.

Acknowledgments

This work was supported by Sichuan Science and Technology Program (Grant No.2024YFG0006), the National Natural Science Foundation of China (Grant No.U24A20250), and the Fundamental Research Funds for the Central Universities (No.ZYGX2024Z005).

References

Alemi, A. A.; Fischer, I.; Dillon, J. V.; and Murphy, K. 2022. Deep Variational Information Bottleneck. In *International Conference on Learning Representations*.

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42: 335–359.

Fang, J.; Wang, W.; Lin, G.; and Lv, F. 2024. Sentiment-oriented Sarcasm Integration for Video Sentiment Analysis Enhancement with Sarcasm Assistance. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 5810–5819.

Gao, Z.; Jiang, X.; Xu, X.; Shen, F.; Li, Y.; and Shen, H. T. 2024. Embracing Unimodal Aleatoric Uncertainty for Robust Multimodal Fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26876–26885.

Guo, Z.; Jin, T.; and Zhao, Z. 2024. Multimodal Prompt Learning with Missing Modalities for Sentiment Analysis and Emotion Recognition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1726–1736.

Hazarika, D.; Zimmermann, R.; and Poria, S. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, 1122–1131.

He, P.; Liu, X.; Gao, J.; and Chen, W. 2021. DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION. In *International Conference on Learning Representations*.

Jaiswal, M.; and Provost, E. M. 2020. Privacy enhanced multimodal neural representations for emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 7985–7993.

Kingma, D. P. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, M.; Yang, D.; Lei, Y.; Wang, S.; Wang, S.; Su, L.; Yang, K.; Wang, Y.; Sun, M.; and Zhang, L. 2024a. A Unified Self-Distillation Framework for Multimodal Sentiment Analysis with Uncertain Missing Modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 10074–10082.

Li, M.; Yang, D.; Zhao, X.; Wang, S.; Wang, Y.; Yang, K.; Sun, M.; Kou, D.; Qian, Z.; and Zhang, L. 2024b. Correlation-Decoupled Knowledge Distillation for Multimodal Sentiment Analysis with Incomplete Modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12458–12468.

Lian, Z.; Chen, L.; Sun, L.; Liu, B.; and Tao, J. 2023. GCNet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on pattern analysis and machine intelligence*, 45(7): 8419–8432.

Liang, P. P.; Deng, Z.; Ma, M. Q.; Zou, J. Y.; Morency, L.-P.; and Salakhutdinov, R. 2024. Factorized contrastive learning: Going beyond multi-view redundancy. *Advances in Neural Information Processing Systems*, 36.

Liang, P. P.; Zadeh, A.; and Morency, L.-P. 2022. Foundations and Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions. *arXiv preprint arXiv:2209.03430*.

Lin, R.; and Hu, H. 2023. Dynamically shifting multimodal representations via hybrid-modal attention for multimodal sentiment analysis. *IEEE Transactions on Multimedia*.

Liu, A.; Tan, Z.; Wan, J.; Liang, Y.; Lei, Z.; Guo, G.; and Li, S. Z. 2021. Face anti-spoofing via adversarial cross-modality translation. *IEEE Transactions on Information Forensics and Security*, 16: 2759–2772.

Luo, W.; Xu, M.; and Lai, H. 2023. Multimodal reconstruct and align net for missing modality problem in sentiment analysis. In *International conference on multimedia modeling*, 411–422. Springer.

Mai, S.; Sun, Y.; Xiong, A.; Zeng, Y.; and Hu, H. 2023. Multimodal Boosting: Addressing Noisy Modalities and Identifying Modality Contribution. *IEEE Transactions on Multimedia*.

Mai, S.; Zeng, Y.; and Hu, H. 2022. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia*, 25: 4121–4134.

Pham, H.; Liang, P. P.; Manzini, T.; Morency, L.-P.; and Póczos, B. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 6892–6899.

Schneider, S.; Baevski, A.; Collobert, R.; and Auli, M. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Tsai, Y.-H. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.-P.; and Salakhutdinov, R. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, 6558. NIH Public Access.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, D.; Liu, S.; Wang, Q.; Tian, Y.; He, L.; and Gao, X. 2022. Cross-modal enhancement network for multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 25: 4909–4921.

Wang, Y.; Cui, Z.; and Li, Y. 2023. Distribution-consistent modal recovering for incomplete multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22025–22034.

Wang, Y.; Li, Y.; and Cui, Z. 2023. Incomplete multimodality-diffused emotion recognition. *Advances in Neural Information Processing Systems*, 36: 17117–17128.

Wei, S.; Luo, C.; and Luo, Y. 2023. MMANet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20039–20049.

Xu, W.; Jiang, H.; and Liang, X. 2024. Leveraging Knowledge of Modality Experts for Incomplete Multimodal Learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 438–446.

Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L.-P. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

Zadeh, A. B.; Liang, P. P.; Poria, S.; Cambria, E.; and Morency, L.-P. 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)*, 2236–2246.

Zeng, Y.; Mai, S.; Yan, W.; and Hu, H. 2023. Multimodal reaction: Information modulation for cross-modal representation learning. *IEEE Transactions on Multimedia*.

Zhang, K.; Zhang, Z.; Li, Z.; and Qiao, Y. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10): 1499–1503.

Zhao, J.; Li, R.; and Jin, Q. 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)*, 2608–2618.

Zhao, Z.; Liu, Q.; and Wang, S. 2021. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE Transactions on Image Processing*, 30: 6544–6556.

Zhu, A.; Hu, M.; Wang, X.; Yang, J.; Tang, Y.; and Ren, F. 2024. KEBR: Knowledge Enhanced Self-Supervised Balanced Representation for Multimodal Sentiment Analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 5732–5741.

Zhuang, Y.; Liu, M.; Zhang, Y.; Deng, J.; and Ren, F. 2025a. TMDC: A Two-Stage Modality Denoising and Complementation Framework for Multimodal Sentiment Analysis with Missing and Noisy Modalities. arXiv:2511.10325.

Zhuang, Y.; Zhang, Y.; Deng, J.; and Ren, F. 2025b. R3DG: Retrieve, Rank, and Reconstruction with Different Granularities for Multimodal Sentiment Analysis. *Research*, 8: 0729.

Zhuang, Y.; Zhang, Y.; Hu, Z.; Zhang, X.; Deng, J.; and Ren, F. 2024. GLoMo: Global-Local Modal Fusion for Multimodal Sentiment Analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 1800–1809.

Zong, D.; Ding, C.; Li, B.; Li, J.; Zheng, K.; and Zhou, Q. 2023. AcFormer: An Aligned and Compact Transformer for Multimodal Sentiment Analysis. In *Proceedings of the 31st ACM International Conference on Multimedia*, 833–842.

Zuo, H.; Liu, R.; Zhao, J.; Gao, G.; and Li, H. 2023. Exploiting modality-invariant feature for robust multimodal emotion recognition with missing modalities. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.