Noise Imitation Based Adversarial Training for Robust Multimodal Sentiment Analysis

Article *in* IEEE Transactions on Multimedia · January 2023

DOI: 10.1109/TMM.2023.3267882

Ziqi Yuan
Tsinghua University

Noise Imitation based Adversarial Training for Robust Multimodal Sentiment Analysis

Ziqi Yuan, Yihe Liu, Hua Xu, and Kai Gao

***Abstract*—As an inevitable phenomenon in real-world applications, data imperfection has emerged as one of the most critical challenges for multimodal sentiment analysis. However, existing approaches tend to overly focus on a specific type of imperfection, leading to performance degradation in real-world scenarios where multiple types of noise exist simultaneously. In this work, we formulate the imperfection with the modality feature missing at the training period and propose the noise intimation based adversarial training framework to improve the robustness against various potential imperfections at the inference period. Specifically, the proposed method first uses temporal feature erasing as the augmentation for noisy instances construction and exploits the modality interactions through the self-attention mechanism to learn multimodal representation for original-noisy instance pairs. Then, based on paired intermediate representation, a novel adversarial training strategy with semantic reconstruction supervision is proposed to learn unified joint representation between noisy and perfect data. For experiments, the proposed method is first verified with the modality feature missing, the same type of imperfection as the training period, and shows impressive performance. Moreover, we show that our approach is capable of achieving outstanding results for other types of imperfection, including modality missing, automation speech recognition error and attacks on text, highlighting the generalizability of our model. Finally, we conduct case studies on general additive distribution, which introduce background noise and blur into raw video clips, further revealing the capability of our proposed method for real-world applications.**

***Index Terms*—Robust Multimodal Sentiment Analysis, Imperfection Topology, Adversarial Training, Semantic Reconstruction.**

I. Introduction

PREVIOUS research [1]–[3] on multimodal sentiment analysis (MSA) has made impressive improvements through leveraging the synergistic effect of several modalities such as transcribed text, auditory, and visual materials. However, applying MSA system to real-world applications is still far from success due to the presence of data imperfections [4]–[6]. It has been validated that such imperfections during testing time can seriously confuse the representation learning of unprepared models and further result in degraded performances [7]–[9]. Driven by the demand from real-world applications, this work focuses on improving robustness in the MSA against the potential data imperfection challenge.

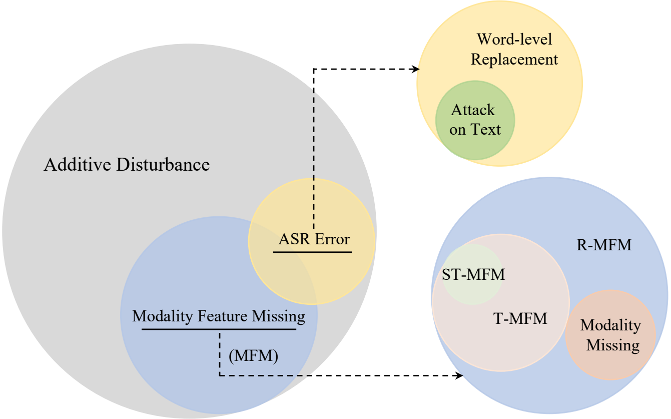


Fig. 1. Hierarchical taxonomy of data imperfection. The first level division is depicted on the left, while the fine-grained categorization of modality feature missing and ASR error are shown on the right, where R-MFM, T-MFM, and ST-MFM refer to random, temporal, and structural temporal modality feature missing respectively.

The fundamental challenge to achieving robust multimodal sentiment analysis lies in the appropriate formulation of the imperfections [10]. Figure 1 provides a hierarchical taxonomy of common data imperfections in MSA research. Specifically, *additive disturbance* encompasses all common imperfections in multimodal applications, such as background noise in audio and visual blur in raw video clips. However, few previous work can validate the proposed method on generalized additive imperfection due to the difficulty in formulation and quantitative evaluation. As a subset of additive disturbance, *ASR error* focuses on the imperfection within the predominant linguistic modality based on the fact that transcribed text is noisy and vulnerable to attack. The other subset of additive disturbance is *modality feature missing*, which models imperfections resulting from potential facial detection failure, transmission error with zero padding vectors and unknown word token for text. Several subclass imperfection is derived from the random modality feature missing. *Modality missing* refers to the scenarios that some of the modality sequences that are missing entirely. *Temporal modality feature missing* refers to the phenomenon that correlated feature missing across all modalities at random time steps, while *structural temporal modality feature missing* is a particular case of the temporal feature missing where correlated missing exists in consecutive time steps. The existing work [11]–[14] has been limited to specific imperfections, which raises concerns about whether such formulations of imperfections adequately reflect real-world imperfections. In this work, we first introduce a

This paper is founded by National Natural Science Foundation of China (Grant No. 62173195) and Beijing Academy of Artificial Intelligence (BAAI).

Hua Xu is the corresponding author, Email: xuhua@mail.tsinghua.edu.cn

Our code is available at https://github.com/Columbine21/NIAT.

Ziqi Yuan and Hua Xu is with State Key Laboratory of Intelligent Technology and Systems, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China. Yihe Liu and Kai Gao is with School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang 050018, China.

novel framework that formulates imperfection as modality feature missing and evaluate the proposed method on other heterogeneous imperfections to show generalization ability against potential real-world imperfections.

Compared to conventional MSA models, building a robust MSA model for cases where modality feature missing exists at the inference period presents three challenges. The first challenge is that potential noisy instances with the missing modality feature can be seen as a type of evasion attack [15], [16]. In such attacks, noisy samples attempt to evade a system that has been trained on clean data. The second challenge arises due to inevitable representation disparities caused by the missing components [17], [18]. In order to mitigate this issue, it is crucial to effectively reduce the distribution gap between the fused representation of the original-noisy data pairs. The third and final challenge pertains to sparse semantics in noisy data [14], [19], which requires the model's ability to recover missing semantic components with the paired perfect data. Addressing these challenges necessitates the development of a well-designed model with unified and effective multimodal representation ability for both noisy and perfect data.

To tackle the first challenge, the proposed framework employs a noise imitation-based augmentation technique to introduce instances with low-level temporal modality feature missing into the training data. This allows the proposed framework to prepare itself for potential modality feature missing while avoiding the risk of non-convergence due to low data quality. To tackle the second challenge, the proposed framework incorporates adversarial representation learning to match the distribution from the different data sources. Guided by the learned discrimination module, the fusion module is able to learn a similar representation space for both original and noisy instances, thus reducing the effect caused by data imperfection. For the last challenge, the proposed framework utilizes utterance-level feature reconstruction to guide representation learning, instead of low-level feature reconstruction, to avoid the recovery of emotionally irrelevant information. The efficacy of the proposed data augmentation, adversarial training, and reconstruction module is further validated and analyzed in the ablation and case study.

The main contributions of this work are summarized below.

- To the best of our knowledge, this work represents one of the earliest attempts to construct one unified framework capable of achieving robust performance against four distinct forms of potential data imperfections, which include (random, temporal, and structural temporal) modality feature missing, entire modality missing, ASR error, as well as attacks on text modality, simultaneously.
- In this paper, we introduce the Noise Intimating-based Adversarial Training (NIAT) framework, which integrates noise-aware adversarial training and utterance-level semantics reconstruction to narrow the representation gap between original and noisy data pairs, and further facilitate robust representation learning.
- Extensive experiments on two benchmark MSA datasets indicate that the proposed NIAT framework consistently enhances robustness against all four forms of heterogeneous data imperfection, while also demonstrating outstanding generalization ability. Furthermore, the case study comparing ideally formulated noise and real-world additional perturbations reveals that the modality feature missing is an effective formulation for real-world noise.

II. Related Works

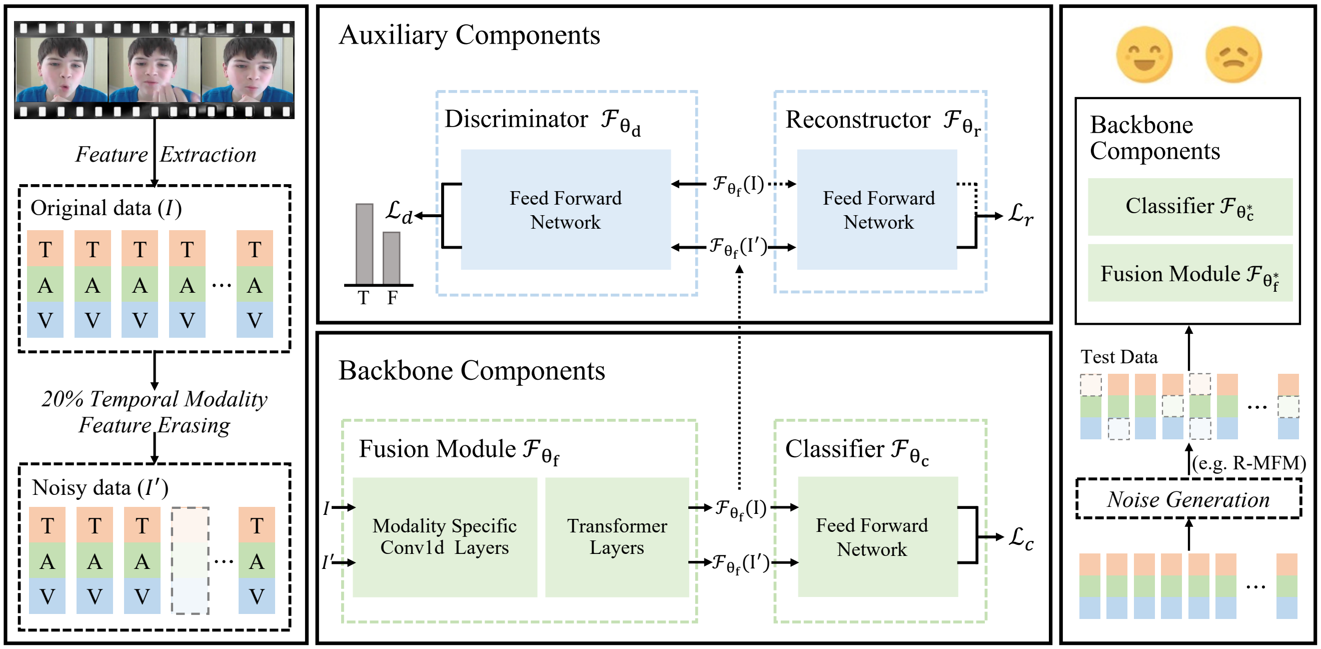
A. Robust Multimodal Sentiment Analysis

Noticing that traditional methods designed on perfect data are sensitive to imperfection, robust multimodal sentiment analysis against potential data imperfection has attracted more and more attention [7]. In the following of this subsection, we summarize recent work for each typical type of data imperfection at the inference period.

1) Modality Missing: Originating from modality ablation study in traditional MSA research, modality missing is the most concerned type of imperfection. A simple but effective method for modality missing is *missing modality imputation*. Tsai et al. [20] uses representation fission and imputes the missing modality based on learnt multimodal discriminative and modality-specific generative factors, while Ma et al. [21] imputes the representation of the missing modality by adding cluster center vectors with weights from learned Gaussian distribution. Recently, Han et al. [22] further enhances representations imputation through alignment matrices. Another paradigm involves *modality translation* [23]–[26], which utilizes the intermediate representation between source and target modality as a joint multimodal feature. However, both missing modality imputation-based methods and modality translation-based methods require knowing which of the entries or modalities are imperfect beforehand at inference periods and fail in more complicated cases where imperfection exists in frame granularity instead of modality granularity [10], [27].

2) ASR Error and Attack on Text Modality: ASR error and attack on transcribed spoken words, as a severe threat to text-dominant tasks, becomes another typical type of data imperfection for multimodal sentiment analysis. Literature [28] is the first work to focus on the imperfection of transcribed text. They achieve robust sentiment recognition by integrating an automatic speech recognition output with a character-level recurrent neural network. Recently, addressing sentiment word replacement caused by ASR error, Wu et al. [13] propose the sentiment word aware multimodal refinement model, which dynamically refines the erroneous sentiment words by leveraging multimodal sentiment clues. However, it should be noted that imperfection or attack on text is only a small part of real-world imperfection compared with imperfection in auditory and visual modalities.

3) Modality Feature Missing: A more general type of imperfection for MSA is the modality feature missing, which ideally formulates fine-grained multimodal perturbation as features missing in modality sequences. Under such an assumption, based on the observation of the low-rank natural from the clean data, Liang et al. [29] propose a low-rank regularization based model, and Li et al. [30] improve the model by exploiting the data dynamics across the temporal domain with lower consumption of memory resources. More recently, Yuan et al. [19] propose a transformer-based feature reconstruction


Feature Extraction
Original data (I)
T A V
20% Temporal Modality Feature Erasing
Noisy data (I')
Auxiliary Components
Discriminator $\mathcal{F}_{\theta_d}$
Reconstructor $\mathcal{F}_{\theta_r}$
Feed Forward Network
$\mathcal{L}_d$
T F
$\mathcal{F}_{\theta_f}(I)$
$\mathcal{F}_{\theta_f}(I')$
$\mathcal{L}_r$
Backbone Components
Fusion Module $\mathcal{F}_{\theta_f}$
Modality Specific Conv1d Layers
Transformer Layers
Classifier $\mathcal{F}_{\theta_c}$
$\mathcal{L}_c$
I
I'
Classifier $\mathcal{F}_{\theta_c^*}$
Fusion Module $\mathcal{F}_{\theta_f^*}$
Test Data
(e.g. R-MFM)
Noise Generation

(a) Data Processing (b) Overview Training Framework (c) Model Evaluation

Fig. 2. The overview of the proposed NIAT framework for robust MSA against data imperfection. Part (a) shows the data processing steps including feature extraction and noisy imitation based augmentation. Part (b) shows the adversarial framework consisting of backbone and auxiliary components. Part (c) shows the model robustness evaluation steps.

network to recover the missing semantics, and Sun et al. [14] combine former low-level feature reconstruction with high-level feature attraction to achieve robust performance. However, previous methods are limited by the generalization drawback that different models need to be trained for different missing degrees. This work uses temporal feature missing imitation during the training period to prepare the method for fine-grained multimodal imperfection and achieves outstanding robustness against various heterogeneous imperfections.

B. Noise-based Augmentation

Data noising is a widely used technique for data augmentation in various fields, including computer vision [31], [32] and speech recognition [33], [34]. More recently, there has been a growing interest in noise-based data augmentation methods in multimodal applications. For instance, Parthasarathy and Sundaram [35] propose a training strategy for audio-visual expression recognition, which involves introducing randomly ablated visual inputs to handle missing input modalities. Chumachenko et al. [36] utilize modality dropout strategy to improve MSA performance under incomplete data of one modality. Inspired by the existing literature, we introduce temporal modality feature erasing augmentation strategy for modality feature missing. Specifically, data of partial time steps in modality sequences are randomly erased with a preset missing rate to obtain the augmented data instead of dropping entire modality sequences.

C. Adversarial Representation Learning

Adversarial representation learning, which originates from Generative Adversarial Networks [37], is an emerging technique that enables the explicit matching of a distribution to an arbitrary prior distribution [38], [39]. More recently, adversarial learning strategy has further extended to multimodal fields such as text-to-image synthesis [40], [41]. Specifically, in the context of the MSA, researchers have utilized adversarial training to learn discriminative and generative representations for each modality, thus improving model performance [20]. Other researchers have introduced adversarial training to develop a discriminative joint embedding space for various modalities [18]. The proposed NIAT framework stands apart from previous research, as it utilizes adversarial training between original-noisy multimodal instance pairs to narrow the distribution gap for robustness against imperfection.

III. Methodology

A. Problem Statement

Robust multimodal sentiment analysis aims to predict the speakers' affective state by leveraging multimodal signals under potential data imperfection. In this work, given the extracted Bert token sequences $\mathbf{U}_t \in \mathcal{R}^{T\times 1}$, acoustic feature sequences $\mathbf{F}_a \in \mathcal{R}^{T\times d_a}$, and the visual feature sequences $\mathbf{F}_v \in \mathcal{R}^{T\times d_v}$, the feature missing is formulated as follows,

Modality Feature Missing Formulation. For text modality, the missing tokens are set to [UNK][1] to imitate the potential translation error. The missing features are set to zero padding vectors for visual and acoustic modalities.

At the training period, due to manual data collection, we assume that completed modality sequences $\mathbf{U}_t, \mathbf{F}_a, \mathbf{F}_v$ from the text, audio, and visual modalities are provided with

[1] a special token in Bert language model that refers to the unknown word.

corresponding sentiment annotation y. While during the testing period, the trained model is evaluated using instances with unknown perturbations to validate the model's effectiveness and robustness against data imperfection.

B. Noise Imitation-based Augmentation

As shown in Figure 2 (a), temporal feature erasing is utilized for augmentation to imitate the potential imperfection in inference periods. Specifically, given the clean train instances with original modality sequences $I = [\mathbf{U}_t, \mathbf{F}_a, \mathbf{F}_v]$, 20% time steps are randomly selected, and all modality features at these time steps are erased simultaneously.

$$[\mathbf{U}_t^{'}; \mathbf{F}_a^{'}; \mathbf{F}_v^{'}] = \text{Random Erasing}\left([\mathbf{U}_t; \mathbf{F}_a; \mathbf{F}_v]\right), \qquad (1)$$

where $\mathrm{I}^{'} = [\mathbf{U}_t^{'}; \mathbf{F}_a^{'}; \mathbf{F}_v^{'}]$ denotes the corresponding augmented noisy instances. The feature erasing is conducted in aligned ways to guide the model, focusing on the remaining emotion-bearing parts in the noisy instances instead of directly making use of synchronized information from other modalities.

Then, the Bert tokens sequences are converted into textual features using the pretrained Bert [42],

$$\mathbf{F}_t = \text{Bert}(\mathbf{U}_t), \mathbf{F}_t^{'} = \text{Bert}(\mathbf{U}_t^{'}) \in \mathcal{R}^{T \times d_t}. \qquad (2)$$

The obtained original-noisy pair $\mathrm{I} = [\mathbf{F}_t; \mathbf{F}_a; \mathbf{F}_v]$ and $\mathrm{I}^{'} = [\mathbf{F}_t^{'}; \mathbf{F}_a^{'}; \mathbf{F}_v^{'}]$ are passed to the NIAT training framework.

C. Backbone Components

Shown in Figure 2 (b), backbone components contain fusion and classification module.

Fusion Module. The fusion module is designed to learn the representation with the temporal and cross-modal dynamics for downstream predictions. Firstly, the original-noisy data pairs are processed with modality-specific 1D convolutional layers to obtain information of neighbour elements,

$$\mathbf{H}_m = \text{Conv1d}(\mathbf{F}_m, k_m) \in \mathcal{R}^{T \times d_m}, \qquad (3)$$

$$\mathbf{H}_m^{'} = \text{Conv1d}(\mathbf{F}_m^{'}, k_m) \in \mathcal{R}^{T \times d_m}, \qquad (4)$$

where k_m refers to the kernel sizes, and d_m is the hidden dimension for modality m. The convolved sequences are then concatenated at the time dimension to get hidden sequences representation,

$$\mathbf{H} = \text{Concat}\left([\mathbf{H}_t; \mathbf{H}_a; \mathbf{H}_v]\right) \in \mathcal{R}^{T \times d}, \qquad (5)$$

$$\mathbf{H}^{'} = \text{Concat}\left(\left[\mathbf{H}_t^{'}; \mathbf{H}_a^{'}; \mathbf{H}_v^{'}\right]\right) \in \mathcal{R}^{T \times d}, \qquad (6)$$

where $d = d_t + d_a + d_v$. Then, a self-attention mechanism based on the Transformer [43] is utilized to make full use of the complementarity of temporal modal information.

Transformer. The Transformer [43] is a stack of layers consisting of a scaled dot-product based attention module and feed-forward network. For simplicity, we denote it as Transformer$(\mathbf{Q}, \mathbf{K}, \mathbf{V})$, where $\mathbf{Q}, \mathbf{K}$ and $\mathbf{V}$ stand for query, key, and value matrices.

With the above notations, the hidden sequences of original-noisy pairs are enhanced by,

$$\overline{\mathbf{H}} = \text{Transformer}\left(\mathbf{H}, \mathbf{H}, \mathbf{H}\right) \in \mathcal{R}^{T \times d}, \qquad (7)$$

$$\overline{\mathbf{H}}^{'} = \text{Transformer}\left(\mathbf{H}^{'}, \mathbf{H}^{'}, \mathbf{H}^{'}\right) \in \mathcal{R}^{T \times d}, \qquad (8)$$

The first time step vector of the enhanced fusion sequences $\mathbf{h} = \overline{\mathbf{H}}[0,:]$ and $\mathbf{h}^{'} = \overline{\mathbf{H}}^{'}[0,:]$, which refers to the [CLS][2] in original aligned modality sequences, is utilized as the joint final multimodal representation. The overall fusion operation of the original-noisy pairs is denoted by,

$$\mathbf{h} = \mathcal{F}_{\theta_f}\left(\mathrm{I}\right), \mathbf{h}^{'} = \mathcal{F}_{\theta_f}(\mathrm{I}^{'}), \qquad (9)$$

where θ_f are the learnable parameters of the fusion module.

Classification Module. The classification module is implemented with two layers of fully-connected Feed Forward Network to predict the sentiment according to the extracted representation $\mathbf{h}$ and $\mathbf{h}^{'}$ for both original-noisy data pairs.

Feed-Forward Network. The one-layer feed-forward network is defined as,

$$\text{FFN}(\mathbf{x}) = \sigma(\mathbf{x} \cdot \mathbf{W}_f + \mathbf{b}_f), \qquad (10)$$

where σ represents the optional activation function, $\mathbf{W}_f$ and $\mathbf{b}_f$ are learnable model parameters.

According to the definition of Feed-Forward Network, the classification module is formulated as,

$$\mathcal{C}_{\theta_c}(\mathbf{h}) = \text{FFN}\left(\text{FFN}\left(\text{BN}(\mathbf{h})\right)\right) \in \mathcal{R}, \qquad (11)$$

where BN is the BatchNorm, θ_c is the learnable parameters in the classification module.

For sentiment prediction supervision, the classification loss $\mathcal{L}_c$ is calculated by the weighted sum from original and noisy data pairs,

$$\mathcal{L}_c = \frac{\text{L1}\left(y, \mathcal{C}_{\theta_c}(\mathbf{h})\right) + \alpha \cdot \text{L1}\left(y, \mathcal{C}_{\theta_c}(\mathbf{h}^{'})\right)}{1 + \alpha}, \qquad (12)$$

where L1 refers to the L1Loss operation, and α is the hyper-parameter.

D. Auxiliary Components

As shown in Figure 2 (b), auxiliary components contain the reconstruction and discrimination modules.

Reconstruction Module. The reconstruction module aims to guide the fusion module in regenerating the missing semantics in the noisy instances by performing utterance level semantic reconstruction. Receiving the fusion representation $\mathbf{h}^{'}$ from noisy data flow, the reconstruction module is implemented by three layers feed-forward network,

$$\mathcal{R}_{\theta_r}(\mathbf{h}) = \text{FFN}\left(\text{FFN}\left(\text{FFN}\left(\text{BN}(\mathbf{h}^{'})\right)\right)\right) \in \mathcal{R}^d, \qquad (13)$$

[2] a special token in Bert language model, appended at the front of the token sequence, commonly used for sentence level representation learning

where BN is the BatchNorm, θ_r is the learnable parameters of the reconstruction module. The reconstruction loss $\mathcal{L}_r$ is designed as the L1Loss between the reconstructed fusion vector and original fusion vector from the perfect data,

$$\mathcal{L}_r = \mathrm{L1}\left(\mathbf{h}, \mathcal{R}_{\theta_r}(\mathbf{h}^{'})\right), \tag{14}$$

Discrimination Module. The noise-aware adversarial training strategy is applied to learn a unified representation by matching the fusion representation distributions of original-noisy data pairs.

The proposed framework learns robust and noise-invariant representations by confusing the discrimination module in a two-player game. The first player, the proposed discrimination module $\mathcal{D}_{\theta_d}$, is trained to distinguish original clean data from the noisy ones with the feature missing, while the second player, the fusion module $\mathcal{F}_{\theta_f}$, is trained to learn the representation that confuses $\mathcal{D}_{\theta_d}$. In the NIAT framework, $\mathcal{D}_{\theta_d}$ is a basic binary classifier formulated by,

$$\mathcal{D}_{\theta_d}(\mathbf{h}) = \sigma(\mathrm{FFN}\left(\mathrm{FFN}\left(\mathrm{FFN}\left(\mathrm{BN}(\mathbf{h})\right)\right)\right)), \tag{15}$$

where σ is the sigmoid function, BN is the BatchNorm, θ_d is the learnable parameters.

Finally, the auxiliary noise-aware adversarial training can be described by the following min-max game:

$$\begin{aligned}\min_{\theta_f}\max_{\theta_d}\mathcal{L}_d = &-\frac{1}{N}\sum_{i=1}^{N}\log\mathcal{D}_{\theta_d}(\mathcal{F}_{\theta_f}(\mathrm{I}))\\ &-\frac{1}{N}\sum_{i=1}^{N}\log(1-\mathcal{D}_{\theta_d}(\mathcal{F}_{\theta_f}(\mathrm{I}^{'}))),\end{aligned} \tag{16}$$

where N is the total instance counts.

E. Model Training

In the proposed NIAT framework, the above representation learning process is guided by three different supervisions. The average classification loss of all original-noisy data pairs $\mathcal{L}_c^{'} = -\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_c$ are the basic supervision for sentiment prediction. While the average discrimination loss $\mathcal{L}_d$ along with the average reconstruction loss $\mathcal{L}_r^{'} = -\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_r$ on all original-noisy data pairs are regarded as the auxiliary loss for robust representation learning. Integrating all objectives together, the final learning procedure is formulated as follows:

$$\min_{\theta_f,\theta_c,\theta_r}\max_{\theta_d}\mathcal{L} = ((1-\beta)\cdot\mathcal{L}_r^{'} + \beta\cdot\mathcal{L}_d) + \mathcal{L}_c^{'}, \tag{17}$$

where β is a hyper-parameter balancing the auxiliary losses of the reconstruction and the discrimination.

IV. Experimental Setups

This section describes the setups for the following experiments, including the datasets (Section IV-A), comparison baselines (Section IV-B), evaluation metrics (Section IV-C) and the Hyperparameters selection. (Section IV-D).

A. Datasets

In this paper, experiments are conducted on two benchmark MSA datasets, MOSI and MOSEI. **MOSI** [44] is a widely-used MSA dataset that consists of a collection of 2,199 video segments from 93 YouTube movie review videos. **MOSEI** [45] expands the MOSI dataset by enlarging the number of utterances and enriching the variety of samples, speakers, and topics. In addition to the larger dataset size, the average utterance length and duration of the MOSEI dataset are also increased by 53% and 74%, respectively, compared to the MOSI dataset. Detailed statistics are reported in Appendix. For both MOSI and MOSEI datasets, instances are annotated with a sentiment intensity score ranging from -3 to 3 (strongly negative to strongly positive). In all experiments, audio and visual features ($\mathbf{F}_a$, $\mathbf{F}_v$ in Section III) provided by CMU-Multimodal SDK[3] are utilized while the text is extracted with pretrained Bert tokenizer [42].

B. Baselines

The NIAT is compared with three types of baseline methods. **Traditional MSA Baseline** is the first type of baseline method. Specifically, we choose the Multimodal Transformer (MulT) [46], the Modality-Invariant and -Specific Representations (MISA) [47], the Multimodal Adaptation Gate for Bert (MAG-BERT) [48], and the Self-supervised Multi-task Multimodal sentiment analysis network Self-MM [49]. These traditional MSA methods achieve impressive performances on perfect test data through sophisticated fusion approaches.

Baseline for Modality Feature Missing is the second type of baseline method. Specifically, Temporal Tensor Fusion Network (T2FN) [29], Time Product Fusion Network (TPFN) [30], Transformer-based Feature Reconstruction Network (TFR-Net) [19] are included in this class. These methods are designed for fine-grained modality feature missing and thus can be extended to other types of data imperfection easily.

Baseline for Specific Type of Imperfection is the third type of baseline method. For modality missing, Multimodal Factorization Model (MFM) [20], SMIL [21], Modality Translation based methods (Modal-Trans) [24], [26], and MM-Aligned [22] are selected for comparison. For ASR error, the Sentiment Word Aware Multimodal Refinement model (SWRM) [13] is selected for comparison.

To ensure a fair performance comparison, the baseline methods are trained using three different strategies. “Training with clean data only” serves as the basic level comparison where the baseline methods are not equipped to handle potential inference time imperfection. The second strategy is “Training with both clean and noisy data”, which involves training the baselines with both clean and noisy data, using the same noise imitation based augmentation as the proposed NIAT framework for robust training. The third strategy, called “one-to-one training”, involves training the model directly on the same type of imperfection and missing rate as expected during inference. A detailed explanation of each baseline and training strategy can be found in the Appendix.

[3] https://github.com/A2Zadeh/CMU-MultimodalSDK

TABLE I

PERFORMANCES FOR PERFECT, RANDOM, TEMPORAL AND STRUCTURAL TEMPORAL FEATURE MISSING ON MOSI DATASET. FOR EACH TYPE OF DATA IMPERFECTION, AUILC VALUE UNDER MISSING RATES INTERVALS $\{0.0, 0.1, \cdots, 0.9, 1.0\}$ IS REPORTED. MODELS WITH * ARE TRAINED ON THE MIXTURE OF CLEAN AND NOISY DATA. EMT-DLFR (†) IS TRAINED WITH "ONE-TO-ONE" STRATEGY. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Models	Clean	Random Missing			Temporal Missing			Structural Temporal Missing		
	MAE ($\downarrow$)	MAE ($\downarrow$)	Acc-2 ($\uparrow$)	F1 ($\uparrow$)	MAE ($\downarrow$)	Acc-2 ($\uparrow$)	F1 ($\uparrow$)	MAE ($\downarrow$)	Acc-2 ($\uparrow$)	F1 ($\uparrow$)
MulT	0.819	1.533	60.89	50.27	1.532	60.83	50.70	1.620	58.56	46.14
MISA	0.790	1.113	67.23	64.51	1.136	61.15	52.68	1.278	62.30	59.37
MAG-BERT	0.857	1.133	63.01	57.98	1.135	62.61	59.30	1.471	58.37	50.51
Self-MM	0.792	1.113	67.08	64.02	**1.113**	63.97	60.07	1.294	62.30	59.00
MulT*	0.881	1.209	65.59	64.39	1.210	65.71	64.45	1.322	60.21	57.58
MISA*	0.809	1.233	67.20	64.51	1.234	67.13	65.20	1.426	60.75	56.82
MAG-BERT*	0.802	1.316	65.07	64.39	1.319	65.10	63.37	1.528	58.82	54.59
Self-MM*	0.790	1.295	67.43	65.11	1.295	67.65	65.41	1.615	60.80	56.37
T2FN*	0.890	1.211	65.60	64.76	1.211	64.45	64.63	1.303	61.51	**60.75**
TPFN*	0.896	1.195	65.23	62.67	1.196	65.23	62.67	1.267	61.41	58.58
TFR-Net*	0.980	1.204	65.83	63.25	1.201	65.99	63.55	1.265	**62.34**	59.03
EMT-DLFR†	**0.705**	**1.106**	**69.60**	**69.60**	-	-	-	-	-	-
NIAT*	0.758	1.131	68.02	66.13	1.130	**67.95**	**66.06**	**1.261**	61.99	58.70

C. Evaluation

Figure 2 (c) illustrates the general evaluation pipeline, consisting of the construction of noisy test data and the evaluation of sentiment prediction. To quantitatively evaluate the model's robustness against varying missing rates, the Area Under Indicators Line Chart (AUILC) metric proposed in literature [19] is employed. This metric is computed by taking into account the corresponding model performance $\{e_0, e_1, \cdots, e_t\}$ under the increasing missing rates sequence $\{r_0, r_1, \cdots, r_t\}$, and calculating the sum of the area between each pair of adjacent points on the line chart:

$$\sum_{i=0}^{t-1} \frac{(e_i + e_{i+1})}{2} \cdot (r_{i+1} - r_i) \tag{18}$$

For each preset missing rate in modality feature missing and other heterogeneous imperfections, the sentiment intensity prediction is formulated as a regression problem with mean absolute error (MAE) as the metric. Moreover, following the previous works [46], [47], the Acc-2 and F1-Score metrics are utilized as negative/non-negative classification criteria. For all above metrics, higher values indicate better model performance, except for MAE, where lower values are indicative of better model performance.

D. Hyperparameters Selection

For the proposed NIAT framework, the kernel sizes of the convolutional layer (k_t, k_a, k_v) are set to $(3, 3, 9)$ for MOSI, and $(3, 5, 3)$ for MOSEI, the layers of the Transformers are set to 3. Grid search on the validation set is performed for the hidden dimension and dropout rate selection. Adam [50] optimizer is utilized for all experiments. As for the weight of different losses, hyperparameters α and β are adjusted from 0 to 1 with a step length of 0.1, balancing the contribution of each module. The detailed results are discussed in Section V-A4. In addition, an early stop strategy that stops model training when the best MAE on the validation set is not updated for eight consecutive epochs is utilized to prevent the model from overfitting. All models are trained using three different random seeds for a fair comparison.

V. EXPERIMENTAL RESULTS AND ANALYSIS

This section provides a comprehensive analysis of the results obtained for different types of imperfections, including modality feature missing in Section V-A, modality missing, ASR error, attack on text modality in Section V-B, and general additive disturbance in raw video clips in Section V-C.

A. Results for Modality Feature Missing

1) Quantitative Result: Quantitative experiments are conducted for the clean, random, temporal, and structural temporal modality feature missing scenarios. For each missing scenario, the quantitative experiment evaluates the overall performance under the missing rates interval $\{0.0, 0.1, \cdots, 1.0\}$. Table I and Table II present the experimental results on MOSI and MOSEI datasets, respectively. According to the results, observations can be concluded from two aspects.

Model Comparison Aspect. According to Section IV-B, baselines are divided into traditional MSA models trained on the clean data (G-I), improved traditional MSA models with noise-based augmentation (G-II), baselines for modality feature missing (G-III), and EMT-DLFR trained with "one-to-one training" strategy (ideally training one specific model for each missing rate). Firstly, from the comparison between G-I and G-II, we can observe that with the help of noise imitation-based augmentation, traditional MSA models can improve their robustness significantly against data imperfection while maintaining competitive performance on perfect data. Secondly, when compared with G-II and G-III models, the NIAT shows the overall best performance against three types of modality feature missing, especially on MOSEI dataset which contains a larger dataset size and longer average video duration. Besides, the proposed method achieves competitive performance with the EMT-DLFR trained by "one-to-one training" strategy and even better Acc-2 and F1-score on MOSEI. Such results reveal

TABLE II

PERFORMANCES FOR PERFECT, RANDOM, TEMPORAL AND STRUCTURAL TEMPORAL FEATURE MISSING ON MOSEI DATASET. FOR EACH TYPE OF DATA IMPERFECTION, AUILC VALUE UNDER MISSING RATES INTERVALS $\{0.0, 0.1, \cdots, 0.9, 1.0\}$ IS REPORTED. MODELS WITH * ARE TRAINED ON THE MIXTURE OF CLEAN AND NOISY DATA. EMT-DLFR (†) IS TRAINED WITH "ONE-TO-ONE" STRATEGY. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Models	Clean	Random Missing			Temporal Missing			Structural Temporal Missing		
	MAE(↓)	MAE(↓)	Acc-2(↑)	F1(↑)	MAE(↓)	Acc-2(↑)	F1(↑)	MAE(↓)	Acc-2(↑)	F1(↑)
MulT	0.560	0.760	71.84	69.49	0.760	71.88	69.52	0.801	66.86	65.31
MISA	0.572	0.748	75.05	71.02	0.747	75.03	70.99	0.779	73.31	68.24
MAG-BERT	0.541	0.708	76.55	73.04	0.710	76.29	72.88	0.741	74.19	70.41
Self_MM	0.578	0.745	64.03	61.18	0.744	64.16	61.50	0.783	55.58	51.86
MulT*	0.559	0.715	68.67	68.89	0.715	68.52	68.70	0.763	61.35	61.70
MISA*	0.571	0.721	73.75	73.09	0.720	73.77	73.09	0.766	71.71	69.69
MAG-BERT*	0.536	0.697	74.33	74.11	0.698	73.48	73.55	**0.723**	70.17	69.97
Self-MM*	0.574	0.722	70.39	70.30	0.723	70.40	70.35	0.762	65.43	65.35
T2FN*	0.580	0.723	73.27	71.63	0.722	73.31	71.66	0.760	67.72	66.24
TPFN*	0.590	0.725	73.78	72.84	0.724	73.71	72.78	0.758	69.73	68.98
TFR-Net*	0.593	0.725	73.39	71.44	0.724	73.40	71.44	0.756	71.28	67.74
EMT-DLFR†	**0.527**	**0.665**	76.40	75.20	-	-	-	-	-	-
NIAT*	0.554	0.690	**77.81**	**75.24**	**0.690**	**77.79**	**75.22**	0.735	**75.29**	**71.26**

that the proposed NIAT is capable of dealing with various potential modality feature missing, and might be further improved when the missing rate and type is known beforehand through "one-to-one training" strategy. Lastly, the competitive result on clean data, along with the outstanding robustness against various types of modality feature missing, validates the proposed NIAT method a unified MSA framework balancing robustness and generalization ability.

Imperfection Comparison Aspect. Firstly, the apparent performance gap between clean data and all missing scenarios shows that the perturbation is an inevitable threat to real-world MSA applications. Secondly, among three missing scenarios, all models perform similarly in the cases of random and temporal modality feature missing (Most model performance changes within 1%). In comparison, they perform significantly worse on the structural temporal modality missing compared to the former two scenarios. The result shows that the structural temporal modality missing is more challenging since the consecutive sequence missing can prevent models from recovering missing semantics from the nearby modality signal.

2) Qualitative Result: The diagram presented in Figure 3 displays the performance curves of the NIAT, G-II, and G-III baselines for temporal and structural temporal modality feature missing scenarios. The graph indicates that the proposed NIAT method surpasses all G-II and G-III baselines on the MOSEI dataset, across all missing rates intervals. On the MOSI dataset, the NIAT performs best in the low-level missing rate interval (0% to 50%), but its performance deteriorates for higher missing rate intervals. Furthermore, it is worth noting that on MOSI dataset which contains fewer training data, traditional MSA method with noisy augmentation (G-II) is advantageous for lower missing rate intervals due to its stronger adaptability of sophisticated fusion strategy. Conversely, the baselines for modality feature missing (G-III) demonstrate better performance for higher missing rate intervals. Comparing two distinct sub-types of modality feature missing, structural temporal modality feature missing results in more rapid degradation at low-level missing rate interval revealing that it is a more severe threat for robust MSA.

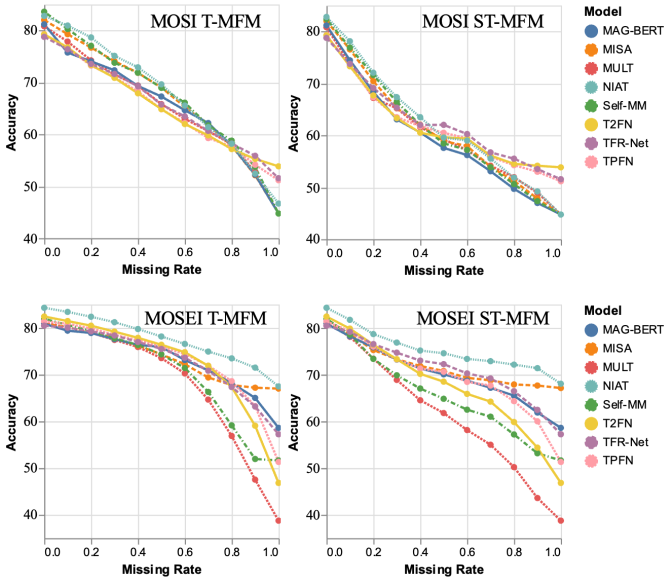


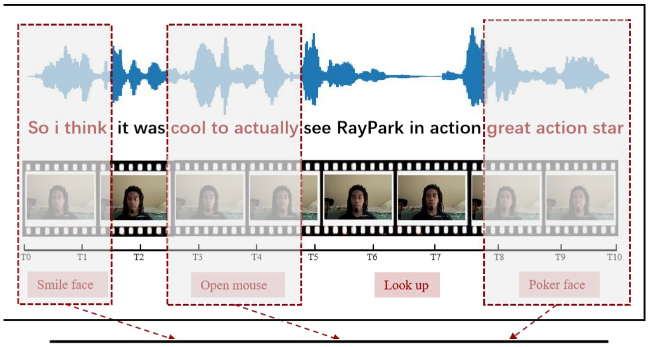
Fig. 3. Qualitative comparison between the proposed NIAT method with the improved MSA models with noise-based augmentation (G-II), as well as the baselines for modality feature missing (G-III), with respect to both temporal and structural temporal feature missing on MOSI and MOSEI datasets.

3) Ablation Study: Table III presents the results of the ablation study on the MOSI and MOSEI datasets for temporal modality feature missing. Firstly, we removed all data augmentation and auxiliary modules from the proposed NIAT model, denoted as **w/o aug**. This led to a reduction of 4.46% in average binary accuracy on MOSI dataset and 9.44% on MOSEI dataset, indicating the significance of the noise imitation-based augmentation. Further ablation studies were conducted on both backbone and auxiliary components. For backbone components, we replaced the transformer-based fusion module with a simple LSTM-based late fusion method, denoted as **w/o fus**. For auxiliary components, we ablated the discrimination module, denoted as **w/o dis**, the reconstruction

TABLE III
ABLATION STUDY RESULTS FOR TEMPORAL MODALITY FEATURE MISSING. THE PROVIDED RESULT IS REPORTED IN MOSI / MOSEI FORMAT. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

	MAE (↓)	Acc-2 (↑)	F1 (↑)
w/o aug	1.180 / 0.735	64.92 / 70.45	61.06 / 70.20
w/o fus	1.217 / 0.723	65.35 / 73.16	63.67 / 72.43
w/o dis, rec	1.184 / 0.727	65.13 / 72.23	62.54 / 71.78
w/o dis	1.157 / 0.717	67.18 / 73.87	65.18 / 73.03
w/o rec	1.172 / 0.711	66.22 / 75.47	64.76 / 73.69
NIAT	**1.130 / 0.690**	**67.95 / 77.79**	**66.06 / 75.22**

module, denoted as **w/o rec**, and both of them, denoted as **w/o dis, rec**, for comparison. It can be observed that the removal of both discrimination and reconstruction modules results in the largest performance gap, underscoring the efficacy of adversarial training and semantic reconstruction. Conversely, the removal of each auxiliary module individually exerts minimal impact on results. This indicates that the discrimination and reconstruction modules are complementary, as the removal of either can be mitigated when the other component remains, and can be mutually enhanced. Moreover, a noteworthy degradation is also observed in the w/o fus scenario, which can be attributed to the significance of an expressive fusion strategy for model performance at lower missing rate intervals, as illustrated in Section V-A2.



	# B-1 Missing	# B-2 Missing	#B-3 Missing
w/p-a	1.23 (1.17)	1.26 (1.14)	1.19 (1.21)
w/o-d	2.11 (0.29)	2.11 (0.29)	2.20 (0.20)
w/o-r	2.24 (0.06)	2.02 (0.38)	2.11 (0.29)
NIAT	2.36 (**0.04**)	2.12 (**0.28**)	2.27 (**0.13**)
Label		2.40	

Fig. 4. Case study results for structural temporal modality feature missing. Human annotation is shown in the last line of the table. The sentiment intensity prediction along with its absolute error is recorded for comparison.

As shown in Figure 4, we further conduct a case study under structural temporal modality feature missing for intuitive demonstration. The first case (#B-1) refers to the situation where the missing block does not contain relevant sentiment factors. In contrast, the second case (#B-2) and the third case (#B-3) refer to situations where some of the crucial sentiment factors, such as “cool” and “great” are missing. It can be found that the discrimination module, which is designed to narrow the distribution gap, contributes similarly to all situations (#B-1, #B-2, #B-3), while the reconstruction module, which aims to regenerate semantic factors, affects the instances with sentiment factor missing (#B-2, #B-3) more seriously.

4) Hyper-parameter Analysis: The NIAT framework uses three different losses to supervise the representation learning during training. Under such circumstances, balancing different losses becomes the fundamental problem as the preset weights of different losses significantly affect the model performance.

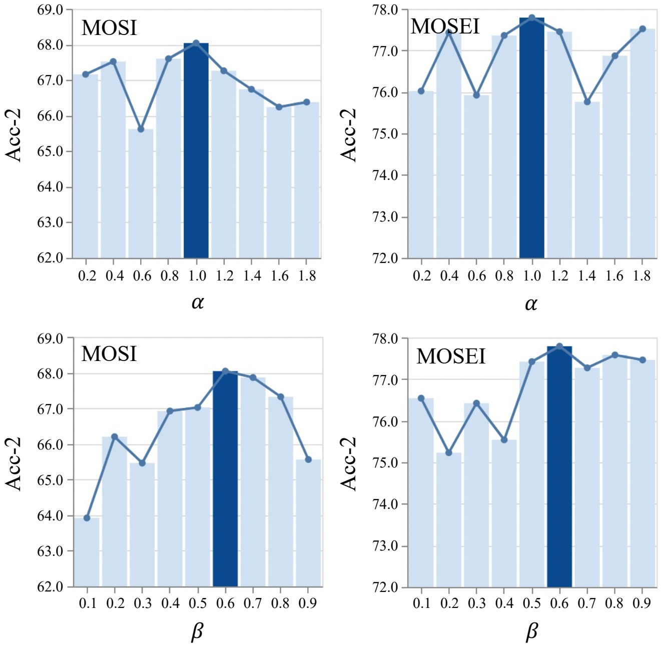


Fig. 5. Performances for different alpha and beta hyper-parameters on MOSI and MOSEI datasets. The best model performances are marked in dark mode.

Balance Between Original and Noise Data. As indicated by Equation 12, the hyperparameter α determines the trade-off between classification losses from perfect data flow and noisy data flow. Our analysis presented in Figure 5 reveals that the model achieves the best performance when $\alpha = 1.0$, indicating that the classification loss from both perfect data flow and noisy data flow contributes equally to the proposed NIAT framework. The efficacy of noise imitation-based augmentation is highlighted again with the hyperparameter analysis.

Balance Between Discrimination and Reconstruction. The hyperparameter β determines the trade-off between the discrimination loss and the reconstruction loss. According to Equation 17, a larger value of β corresponds to a higher weight assigned to the discrimination loss. Based on the results of our tuning experiments depicted in Figure 5, where β ranges from 0.1 to 0.9, we found that $\beta = 0.6$ yielded the best model performance. The relatively higher model performance with larger β indicates that the discrimination loss is crucial for effectively learning representations of corrupted data.

B. Results for Other Heterogeneous Imperfection

Though the NIAT is trained for temporal feature missing, it is also evaluated with other heterogeneous imperfections. For

TABLE IV
PERFORMANCES FOR MODALITY MISSING ON MOSEI DATASET. MODEL WITH * ARE TRAINED ON THE MIXTURE OF CLEAN AND NOISY DATA. MODEL RESULT WITH † IS DIRECTLY EXCERPTED FROM ORIGINAL PAPER.

Models	Clean	Text Missing			Audio Missing			Visual Missing		
	MAE(↓)	MAE(↓)	Acc-2(↑)	F1(↑)	MAE(↓)	Acc-2(↑)	F1(↑)	MAE(↓)	Acc-2(↑)	F1(↑)
T2FN*	0.580	0.851	51.93	47.85	0.583	82.25	82.30	0.599	82.11	82.13
TPFN*	0.590	0.828	62.12	**59.93**	0.614	79.61	79.95	0.593	80.66	80.84
TFR-Net*	0.593	0.867	64.91	58.99	0.589	81.34	81.28	0.626	79.91	80.02
MFM†	-	0.821	62.00	-	0.658	79.10	-	0.658	79.20	-
SMIL†	-	0.820	63.10	-	0.684	78.50	-	0.680	78.30	-
Modal-Trans†	-	0.817	65.10	-	0.643	79.90	-	0.645	79.60	-
MM-Aligned†	-	**0.811**	66.20	-	0.635	81.00	-	0.637	80.80	-
NIAT*	**0.554**	0.836	**70.91**	58.99	**0.556**	**84.42**	**84.23**	**0.562**	**83.99**	**83.96**

TABLE V
MODEL PERFORMANCES FOR AUTOMATIC SPEECH RECOGNITION ERROR ON MOSI DATASET.

Model	MAE (↓)	Acc-2 (↑)	F1 (↑)
MulT*	1.013	74.15	74.18
MISA*	0.946	75.63	75.51
MAG-BERT*	1.119	68.85	68.76
Self_MM*	0.923	76.15	76.18
T2FN*	1.022	72.69	72.32
TPFN*	1.038	71.82	71.41
TFR-Net*	1.084	72.55	72.39
SWRM	0.894	76.45	76.48
NIAT	**0.887**	**77.21**	**77.01**

TABLE VI
MODEL PERFORMANCES FOR SENTIMENT-WORDS DELETION (DEL) AND ANTONYM REPLACEMENT (ANT) ON MOSI DATASET. RESULTS ARE RECORD IN THE FORMAT OF (PERFECT / DEL / ANT).

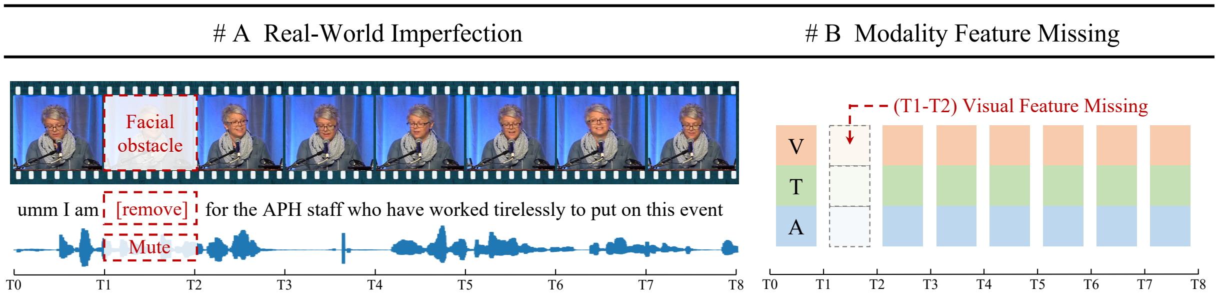
Models	MAE (↓) PER / DEL / ANT	Acc-2 (↑) PER / DEL / ANT
MulT*	0.881 / 1.114 / 1.344	80.80 / 71.82 / 62.24
MISA*	0.809 / 1.062 / 1.326	81.80 / 72.50 / 62.05
MAG-BERT*	0.802 / 1.162 / **1.272**	80.23 / 69.10 / **63.70**
Self-MM*	0.790 / 1.067 / 1.327	80.81 / 72.25 / 62.73
T2FN*	0.890 / 1.111 / 1.320	79.16 / 70.94 / 62.34
TPFN*	0.896 / 1.146 / 1.352	79.30 / 69.44 / 62.05
TFR-Net*	0.980 / 1.136 / 1.326	78.77 / 70.02 / 62.39
SWRM	0.945 / 1.122 / 1.294	80.14 / 71.07 / 61.13
NIAT	**0.758 / 1.026** / 1.368	**81.82 / 73.08** / 62.49

the results in this subsection, we directly record the evaluation indicators of the model on the constructed noise test set without using the AUILC value.

1) Modality Missing: In this setup, one of the text, audio, or visual modality sequences is completely removed during inference. Table IV presents a comparison of the NIAT model with two groups of baselines on the MOSEI dataset. The first group (G-I) consists of baselines trained on both clean and noisy data settings for modality feature missing, while the second group (G-II) comprises several baselines for modality missing imperfections. We observe that G-I outperforms G-II on audio and visual modality missing but performs worse on text modality missing. This phenomenon indicates that G-I achieved robust results over-reliance on text modality, while G-II fails to exploit the effectiveness of the text modality for visual or audio modality missing scenarios. The proposed NIAT model achieves the best overall and more balanced performances for different modality-missing scenarios. Despite the outstanding results of NIAT, its performance with text modality missing degrades much more than that of the non-verbal modalities, highlighting the dominant position of the text modality in the MSA task.

2) ASR Error: In this setup, we replace the provided text on the MOSI test set with the transcribed text from one of the state-of-the-art ASR systems [51] to evaluate the model robustness against potential ASR error. Specifically, the word error rate of the ASR system is about 35% on the MOSI test set. Detailed error cases are shown in Appendix. We compare the NIAT model with traditional MSA baselines trained on clean and noisy data, baselines for modality feature missing and the Sentiment Word Aware Multimodal Refinement model. The result is recorded in Table V. It can be observed that baselines for the modality feature missing (T2FN, TPFN, TFR-Net) show the worst performances. SWRM, which is improved from Self-MM for ASR error, performs better than all traditional MSA baselines trained on clean and noisy data, and the proposed NIAT achieves the best result for all metrics. These results indicate that the NIAT, designed for the modality feature missing, can still be effective in real-world applications, even in the presence of potential ASR errors.

3) Sentiment Word Erasing and Antonym Replacement: In addition to the potential ASR errors, attacks on transcribed text pose another threat to the MSA system. Building on previous research [52], we explore two types of textual attack on sentiment words, i.e. deletion (DEL) and antonym replacement (ANT). The detailed experimental setting of this section can be found in Appendix. Table VI reports the experimental results. The proposed NIAT outperforms all baselines for sentiment word deletion, while the MAG-BERT achieves the best result for sentiment word antonym replacement. Despite the effectiveness of the proposed NIAT method for sentiment word deletion, there is still an obvious performance gap between the perfect situation and attacks on sentiment words. Thus, capturing emotion-related non-verbal cues becomes a crucial


# A Real-World Imperfection
# B Modality Feature Missing
Facial obstacle
umm I am [remove] for the APH staff who have worked tirelessly to put on this event
Mute
T0 T1 T2 T3 T4 T5 T6 T7 T8
(T1-T2) Visual Feature Missing
V
T
A
T0 T1 T2 T3 T4 T5 T6 T7 T8

Imperfection	# Real World Imperfection MAG-BERT	 NIAT	# Modality Feature Missing MAG-BERT	 NIAT
Text (T1-T2)	1.406 (↓ 0.603)	0.937 (↓ **0.469**)	0.097 (↓ 1.912)	0.740 (↓ **0.666**)
Audio (T1-T2)	1.992 (↓ 0.017)	1.402 (↓ **0.004**)	1.338 (↓ 0.406)	1.406 (- **0.000**)
Visual (T1-T2)	1.996 (↓ 0.013)	1.402 (↓ **0.004**)	1.419 (↓ 0.590)	1.401 (↓ **0.005**)
Multimodal (T1-T2)	0.996 (↓ 1.013)	0.973 (↓ **0.433**)	0.193 (↓ 1.816)	0.739 (↓ **0.667**)
Clean Data	2.009	1.406	2.009	1.406

Fig. 6. Result of a case study on the impact of real-world imperfections and modality features missing from time step T1 to time step T2 for MAG-BERT and the proposed NIAT model. The corresponding changes in sentiment prediction are indicated in parentheses.

step for MSA applications in defending against sentiment word antonym replacement attacks.

C. Case Study for Real-world Imperfection

As described in Section I, most previous researches have developed and evaluated on one specific noise formulation, such as modality feature missing, disregarding the potential disparity between the ideally formulated noise and real-world imperfections. This section endeavors to bridge such disparity by comparing the impact of real-world imperfections with that of modality feature missing. Utilizing the existing MSA toolkit [53], [54], we first introduce real-world imperfections into the video clip. Specifically, we conduct four types of imperfections, including the removal of corresponding text, the utilization of mute mode for audio, the masking of the speaker's face for visual, and a combination of all three imperfections (more types of imperfections are shown in the Appendix). These imperfections are implemented from time step T1 to time step T2, which encompasses rich emotional cues from text, audio, and visual modalities. The experimental results are shown in Figure 6. Generally, for both MAG-BERT and the proposed NIAT framework, the imperfections in raw videos and modality feature missing exhibit a similar effect, i.e. resulting in a more neutral sentiment prediction. Moreover, the proposed NIAT framework which performs more robustly in the case of modality feature missing also achieves more robust results (lower prediction changes) on real-world imperfection. Such results reveal that the modality feature missing is a simple yet effective simulation for most real-world imperfection. However, it can also be found that for certain cases, such as face obstacle from T1 to T2, the MAG-BERT displays an apparent performance gap between real-world imperfection and modality feature missing simulation.

VI. Discussion and Conclusion

In this study, we highlight the existence of multiple types of potential data imperfections in real-world applications. To address this issue, we propose a unified framework called noise imitation based adversarial training (NIAT). This framework first utilizes a temporal feature erasing strategy to introduce noisy instances, and combines adversarial training, and semantic reconstruction techniques to guide robust representation learning for both original and noisy data pairs. Our experiments demonstrate that the proposed NIAT model shows an overall better results compared with existing methods under three different modality feature missing scenarios. Moreover, our framework also exhibits outstanding performance on perfect data as well as on other heterogeneous imperfections such as modality missing, ASR errors, and attacks on textual modality, which shows the impressive generalization ability of our proposed NIAT framework. Lastly, our study highlights the potential discrepancy between ideally formulated noise and real-world imperfections. Through a case study, we reveal that modality feature missing is a simple yet effective simulation for real-world imperfections.

Nonetheless, quantitatively evaluating the proposed method on additive disturbance directly presents a challenge. As a substitute, this paper suggests future research should not evaluate their proposed method on one specific type of data imperfection only, for the purpose of reducing the potential drawbacks on overfitting on such type of noise. Moreover, there is an urgent need for an open-source benchmark test dataset containing as many potential imperfection situations as possible. We believe such work will benefit the researchers for convenient and fair comparisons. In the future, we plan to extend the proposed NIAT framework to other multimodal classification tasks and further investigate how to defend the MSA model against real-world imperfections.

References

[1] W. Han, H. Chen, and S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 9180–9192.

[2] D. Wang, S. Liu, Q. Wang, Y. Tian, L. He, and X. Gao, "Cross-modal enhancement network for multimodal sentiment analysis," *IEEE Transactions on Multimedia*, pp. 1–13, 2022.

[3] S. Mai, Y. Zeng and H. Hu, "Multimodal Information Bottleneck: Learning Minimal Sufficient Unimodal and Multimodal Representations," in *IEEE Transactions on Multimedia*, doi: 10.1109/TMM.2022.3171679.

[4] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," *Image and Vision Computing*, vol. 65, pp. 3–14, 2017.

[5] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[6] P. P. Liang, Y. Lyu, X. Fan, Z. Wu, Y. Cheng, J. Wu, L. Chen, P. Wu, M. A. Lee, Y. Zhu *et al.*, "Multibench: Multiscale benchmarks for multimodal representation learning," *arXiv preprint arXiv:2107.07502*, 2021.

[7] M. Ma, J. Ren, L. Zhao, D. Testuggine, and X. Peng, "Are multimodal transformers robust to missing modality?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 177–18 186.

[8] D. Hazarika, Y. Li, B. Cheng, S. Zhao, R. Zimmermann, and S. Poria, "Analyzing modality robustness in multimodal sentiment analysis," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, pp. 685–696.

[9] H. Chi, M. Yang, J. Zhu, G. Wang, and G. Wang, "Missing modality meets meta sampling (m3s): An efficient universal approach for multimodal sentiment analysis with missing modality," in *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, 2022, pp. 121–130.

[10] P. P. Liang, A. Zadeh, and L.-P. Morency, "Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions," *arXiv preprint arXiv:2209.03430*, 2022.

[11] Z. Lian, L. Chen, L. Sun, B. Liu, and J. Tao, "Gcnet: Graph completion network for incomplete multimodal learning in conversation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[12] J. Zhao, R. Li, and Q. Jin, "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)*, 2021, pp. 2608–2618.

[13] Y. Wu, Y. Zhao, H. Yang, S. Chen, B. Qin, X. Cao, and W. Zhao, "Sentiment word aware multimodal refinement for multimodal sentiment analysis with asr errors," in *Findings of the Association for Computational Linguistics: ACL 2022*, 2022, pp. 1397–1406.

[14] L. Sun, Z. Lian, B. Liu, and J. Tao, "Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis," *arXiv preprint arXiv:2208.07589*, 2022.

[15] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "A survey on adversarial attacks and defences," *CAAI Transactions on Intelligence Technology*, vol. 6, no. 1, pp. 25–45, 2021.

[16] K. Yang, W.-Y. Lin, M. Barman, F. Condessa, and Z. Kolter, "Defending multimodal fusion models against single-source adversaries," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3340–3349.

[17] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," *arXiv preprint arXiv:1511.05644*, 2015.

[18] S. Mai, H. Hu, and S. Xing, "Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 164–172.

[19] Z. Yuan, W. Li, H. Xu, and W. Yu, "Transformer-based feature reconstruction network for robust multimodal sentiment analysis," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4400–4407.

[20] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=rygqqsA9KX

[21] M. Ma, J. Ren, L. Zhao, S. Tulyakov, C. Wu, and X. Peng, "Smil: Multimodal learning with severely missing modality," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 2302–2310.

[22] W. Han, H. Chen, M.-Y. Kan, and S. Poria, "Mm-align: Learning optimal transport-based alignment dynamics for fast and accurate inference on missing modality sequences," *arXiv preprint arXiv:2210.12798*, 2022.

[23] H. Pham, P. P. Liang, T. Manzini, L.-P. Morency, and B. Póczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6892–6899.

[24] Z. Wang, Z. Wan, and X. Wan, "Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis," in *Proceedings of The Web Conference 2020*, 2020, pp. 2514–2520.

[25] H. Pham, T. Manzini, P. P. Liang, and B. Poczos, "Seq2seq2sentiment: Multimodal sequence to sequence models for sentiment analysis," *arXiv preprint arXiv:1807.03915*, 2018.

[26] J. Tang, K. Li, X. Jin, A. Cichocki, Q. Zhao, and W. Kong, "Ctfn: Hierarchical learning for multimodal sentiment analysis using coupled-translation fusion network," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 5301–5311.

[27] J. Zeng, J. Zhou and T. Liu, "Robust Multimodal Sentiment Analysis Via Tag Encoding of Uncertain Missing Modalities," in *IEEE Transactions on Multimedia*, 2022, doi: 10.1109/TMM.2022.3207572.

[28] E. Lakomkin, M. A. Zamani, C. Weber, S. Magg, and S. Wermter, "Incorporating end-to-end speech recognition models for sentiment analysis," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 7976–7982.

[29] P. P. Liang, Z. Liu, Y.-H. H. Tsai, Q. Zhao, R. Salakhutdinov, and L.-P. Morency, "Learning representations from imperfect time series data via tensor rank regularization," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 1569–1576.

[30] B. Li, C. Li, F. Duan, N. Zheng, and Q. Zhao, "Tpfn: Applying outer product along time to multimodal sentiment analysis fusion on incomplete data," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*. Springer, 2020, pp. 431–447.

[31] H. Bao, L. Dong, and F. Wei, "Beit: Bert pre-training of image transformers," *arXiv preprint arXiv:2106.08254*, 2021.

[32] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," *arXiv preprint arXiv:2111.06377*, 2021.

[33] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[34] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[35] S. Parthasarathy and S. Sundaram, "Training strategies to handle missing modalities for audio-visual expression recognition," in *Companion Publication of the 2020 International Conference on Multimodal Interaction*, 2020, pp. 400–404.

[36] K. Chumachenko, A. Iosifidis, and M. Gabbouj, "Self-attention fusion for audiovisual emotion recognition with incomplete data," *arXiv preprint arXiv:2201.11095*, 2022.

[37] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[38] T. Pang, X. Yang, Y. Dong, H. Su, and J. Zhu, "Bag of tricks for adversarial training," *arXiv preprint arXiv:2010.00467*, 2020.

[39] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[40] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *International Conference on Machine Learning*. PMLR, 2016, pp. 1060–1069.

[41] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5907–5915.

[42] J. Devlin, M.-W. Chang, K. Lee, and K. N. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter*

of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2018, pp. 4171–4186.

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[44] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.

[45] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "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)*, 2018, pp. 2236–2246.

[46] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, vol. 2019, 2019, pp. 6558–6569.

[47] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: modality-invariant and -specific representations for multimodal sentiment analysis," *CoRR*, vol. abs/2005.03545, 2020. [Online]. Available: https://arxiv.org/abs/2005.03545

[48] W. Rahman, M. K. Hasan, S. Lee, A. B. Zadeh, C. Mao, L.-P. Morency, and E. Hoque, "Integrating multimodal information in large pretrained transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 2359–2369.

[49] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, 2021, pp. 10 790–10 797.

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

[51] J. Grosman, "Fine-tuned XLSR-53 large model for speech recognition in English, "https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english, 2021.

[52] S. Balakrishnan, Y. Fang, and X. Zhu, "Exploring robustness of prefix tuning in noisy data: A case study in financial sentiment analysis," *arXiv preprint arXiv:2211.05584*, 2022.

[53] H. Mao, Z. Yuan, H. Xu, W. Yu, Y. Liu, and K. Gao, "M-sena: An integrated platform for multimodal sentiment analysis," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2022, pp. 204–213.

[54] H. Mao, B. Zhang, H. Xu, Z. Yuan, and Y. Liu, "Robust-msa: Understanding the impact of modality noise on multimodal sentiment analysis," *arXiv preprint arXiv:2211.13484*, 2022.

Appendix A
Other relevant Experimental Setups

A. Dataset Statistics

TABLE I
Dataset statistics for MOSI and MOSEI. *Avg-T.* refers to the average time of each utterances (s), *Avg-W.* refers to the average word count per utterance, and *Ins.* refers to the total instance count.

Item	MOSI			MOSEI		
	#Tr	#Dv	#Ts	#Tr	#Dv	#Ts
Avg-T.	3.64	3.54	4.54	6.70	6.96	7.16
Avg-W.	11.5	10.8	13.2	18.2	18.5	18.7
Ins.	1,284	229	686	16,216	1,835	4,625

MOSI. The MOSI dataset [1] is one of the most popular benchmark datasets for MSA. It comprises 2199 short monologue video clips taken from 93 Youtube movie review videos. Human annotators label each sample with a sentiment score from -3 (strongly negative) to 3 (strongly positive).

MOSEI. The MOSEI dataset [2] expands its data with a higher number of utterances, greater variety in samples, speakers, and topics over MOSI. The dataset contains 23,453 annotated video segments (utterances), from 5,000 videos, 1,000 distinct speakers and 250 different topics.

B. Compared Baselines

MulT. The Multimodal Transformer [3] extends transformer architecture fusion the source modality into the target modality using directional pairwise cross-attention mechanism.

MISA. The Modality-Invariant and -Specific Representations [4] is made up of a combination of losses including similarity loss, orthogonal loss, reconstruction loss and prediction loss to learn modality-invariant and modality-specific representation.

MAG-BERT. The Multimodal Adaptation Gate for Bert [5] incorporates aligned nonverbal information into the text representation within BERT pretrained model.

Self-MM. The Self-supervised Multi-task Multimodal sentiment analysis network [6] first generates the pseudo unimodal sentiment labels and then adopts them to train the model in a multi-task learning manner.

T2FN. The Temporal Tensor Fusion Network [14] is a regularization method based on tensor rank minimization for the imperfect data.

TPFN. The Time Product Fusion Network [15] is an improvement of T2FN, which takes the high-order statistics over both modalities and temporal dynamics into account for the imperfect data.

TFR-Net. The Transformer-based Feature Reconstruction Network [16] improve model robustness by recovering the missing semantics under the guidance of the proposed reconstruction framework.

EMT-DLFR. The Efficient Multimodal Transformer with Dual-Level Feature Restoration [17] combines both implicit low level feature reconstruction and explicit high-level feature attraction to realize robust representation learning.

MFM. The Multimodal Factorization Model [7] factorizes multimodal representations into discriminative factors and modality-specific generative factors at training and imputes the missing modality based on these factors at test time.

SMIL. Literature [8] imputes the representation of the missing modality by linearly adding clustered center vectors with weights from learned Gaussian distribution.

Modal-Trans. The Modality translation based methods [10], [12] build a cyclic sequence-to-sequence model and learns bidirectional reconstruction at training and utilize the intermediate state as the joint fusion representation for classification.

MM-Align. Literature [9] applies optimal transport and denoising training to imitate some indirect but informative clues for the paired modality sequences for modality missing.

SWRM. The Sentiment Word aware multimodal Refinement Model [13] improve the Self-MM model through dynamically refine the erroneous sentiment words by leveraging multimodal sentiment clues.

C. Training Strategies

This section compares three different training strategies. “Training with clean data only” trains model with original train instances, while “Training with both clean and noisy data” prepares model with augmented noisy data. Both “Training with clean data only” and “Training with both clean and noisy data” strategies train unified model once for all potential test imperfection. We show the pseudo code for “Training with both clean and noisy data.” in Algorithm 1.

Algorithm 1 Training with both clean and noisy data.

```
Input: train set $\mathrm{T} = \{\mathrm{I}_1, \mathrm{I}_2, \cdots, \mathrm{I}_n\}, \mathrm{I}_i = [\mathbf{U}_t; \mathbf{U}_a; \mathbf{U}_v]$
Initialize model parameters $M(\theta; x)$
Construct augmented data with 20% temporal modality feature missing $\mathrm{T}^{'} = \{\mathrm{I}_1^{'}, \mathrm{I}_2^{'}, \cdots, \mathrm{I}_n^{'}\}$.
Train model $M(\hat{\theta}; x)$ with the data $T$ and $T^{'}$.
for $r \in$ Missing rate Interval $[0.0, 0.1, \cdots, 1.0]$ do
    Construct Test set $D(r)$ with Missing rate $r$.
    Evaluate model $M(\hat{\theta}; x)$ on constructed test set $D(r)$
end for
Calculate the AUILC value.
```

“One-to-one training” refers to strategy where each model is trained for each missing rates. The pseudo code is shown in Algorithm 2.

Algorithm 2 One-to-one Training Strategy.

```
Input: train set $\mathrm{T} = \{\mathrm{I}_1, \mathrm{I}_2, \cdots, \mathrm{I}_n\}, \mathrm{I}_i = [\mathbf{U}_t; \mathbf{U}_a; \mathbf{U}_v]$
for $r \in$ Missing rate Interval $[0.0, 0.1, \cdots, 1.0]$ do
    Initialize model parameters $M(\theta; x)$
    Construct augmented data with $r$% temporal modality feature missing $\mathrm{T}^{'} = \{\mathrm{I}_1^{'}, \mathrm{I}_2^{'}, \cdots, \mathrm{I}_n^{'}\}$.
    Train model $M_r(\hat{\theta}; x)$ with the data $T$ and $T^{'}$.
    Construct Test set $D(r)$ with Missing rate $r$%.
    Evaluate model $M_r(\hat{\theta}; x)$ on constructed test set $D(r)$
end for
Calculate the AUILC value.
```

TABLE II
ASR ERROR CASES FROM THE MOSI DATASET. FOR EACH EXAMPLE. WE SHOW GROUND-TRUTH TRANSCRIPTION, AND ASR WORD ERROR RATE FOR THIS EXAMPLE.

Ground-truth text	ASR transcription	Sentiment	WER
That is not to say that the ending was not good because it really was	That is to say that the ending was not good because it really	Positive	0.15
The sequences of the beginning of the film which features columbus explaining some of his rules um to surviving a zobie apocalipse are really well done	The sequences at the enginning of the file which feature columbus explaining some of a rules to surviving a zamipocalyps are really well done	Positive	0.27
But it was really really awesome	But it was really really alf some	Positive	0.33
Just like look away what he can do he can do lots of stuff	Just like look at what he can do he can do lots of stuff suffeir	negative	0.14
And he I don't I think he maybe he got mad when HAH i don't konw	He I don't I think he maybe he got mad when I don't know	negative	0.25
And that was really boring	And that was re bory	Negative	0.4

D. Modality Feature Missing Modeling

For modality feature missing, the test data was constructed according to the combination of missing scenario s and missing rate $r\%$. According to the temporal unit of the perturbation, three different missing scenarios are considered in this work.

Random modality feature missing. refers to the scenarios where missing exists in unknown positions independently among each modality. Specifically, given a preset missing rate $r\%$, three modality specific missing mask $\mathrm{M}_k \in \mathcal{R}^T, k \in \{t, a, v\}$ is generated, where $\mathrm{M}_k^{(i)} = 0$ means the ith feature in modality k sequence is missing.

Temporal modality feature missing. refers to a special case of random modality feature missing where modality features are dropped synchronously at certain time steps. Given a preset missing rates, one shared missing mask $\mathrm{M} \in \mathcal{R}^T$ is utilized for all modality, where $\mathrm{M}^{(i)} = 0$ means the ith feature in all modality sequences is missing synchronously.

Structural temporal modality feature missing. refers to a special case of temporal modality feature missing that modality features are perturbed synchronously in consecutive time steps. Specifically, given a preset missing rates $r\%$, the starting point i of the block missing is first chosen and $r\%$ feature for all modalities is dropped by setting $(\mathrm{M}_k^{(i)}, \cdots, \mathrm{M}_k^{(i+t)}) = 0, k \in \{t, a, v\}$, where $t = T \times r\%$.

```python

def random(t_seq, a_seq, v_seq, m_r):
    """ t_seq: bert token seq.
        a_seq, v_seq: sequences for a and v.
        m_r: missing rate.
    """
    #
    token, i_mask = t_seq[:,0,:], t_seq[:,1,:]

    # Missing Mask Construction.
    m_t = (uniform(size=i_mask.shape)>m_r)*i_mask
    m_a = (uniform(size=i_mask.shape)>m_r)*i_mask
    m_v = (uniform(size=i_mask.shape)>m_r)*i_mask

    # Feature Erase.
    t_m = (100*ones_like(token)) * (i_mask-m_t) \
        + m_t * token # [unk] token: 100
    t_m = concatenate(t_m, text[:,1:,:]), axis=1)
    a_m = missing_mask_a * a_seq
    v_m = missing_mask_v * v_seq

    return t_m, a_m, v_m
```

Listing 1. An example code for Random Modality Feature Missing.

The Python implementation of the random modality feature missing is provided in Listing 1. The temporal and structural temporal modality feature missing can be implemented with minor modifications. It should be noticed that for all experiments in Section V-A, each missing scenario s and missing rate $r\%$ pair is tested with five different missing positions using different random seeds to reduce the effect of randomness for model evaluation.

E. Modality Missing Modeling

The modality missing is also constructed similarly to the random modality feature missing, where $(1.0, 0.0, 0.0), (0.0, 1.0, 0.0), (0.0, 0.0, 1.0)$ is used as a missing rate combination of text, audio, and visual modality.

F. ASR Error Modeling

The pre-trained Wav2Vec 2.0 model[1] is utilized for automatic speech recognition. Initially, we acquired the transcribed text for MOSI data. Table II depicts the cases of error in the ASR system. It can be observed that certain instances of directly transcribed text by ASR display a phenomenon of sentiment shift. For instance, the translation of "But it was really really awesome" to "But it was really really alf some" changes the textual sentiment from positive to neutral. In such cases, it becomes crucial to capture non-verbal cues for sentiment prediction. In order to remain consistent with the original paper, the SWRM [13] is trained on transcribed text, while the other baselines are trained on instances of clean and augmented noisy data, using the provided text. For equitable comparisons, all the methods are evaluated using the transcribed text, along with the provided audio and visual sequences, on the same test set.

G. Attacks on Sentiment Word Modeling

For attack on sentiment word, we first perform the sentiment word detection according to the preset sentiment word table (shown in Table III). Then the sentiment words in provided transcribe text are replaced with its antonyms or [UNK] token on MOSI dataset. As a statistics, over 77% of test instances on MOSI dataset are changed. Model performances are evaluated with the attacked transcribe text for comparison.

[1] https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english

TABLE III
ANTONYM REPLACEMENT ON TEST SET OF MOSI DATASET.

Antonym replacement						
boring-interesting	bored-interesting	happy-sad	ok-bad	love-hate	hated-loved	dull-interesting
horrible-lovely	funny-serious	enjoy-hate	enjoyed-hated	like-hate	cool-raw	plucky-cowardly
pretty-ugly	good-bad	fond-hate	above-below	after-before	all-none	alone-together
always-sometimes	answer-ask	answer-question	back-front	bad-good	badly-well	beautiful-ugly
before-after	begin-end	best-worst	better-worse	big-small	black-white	borrow-lend
both-neither	break-mend	busy-free	buy-sell	certainly-perhaps	cheap-expensive	clean-dirty
clever-foolish	cloudy-bright	cold-hot	come-go	cool-warm	danger-safety	dark-bright
day-night	dead-alive	death-life	die-live	down-up	dry-wet	early-late
easy-difficult	empty-fill	empty-full	entrance-exit	fall-rise	far-near	fast-slow
fine-cloudy	finish-begin	first-last	foreign-home	forget-remember	freeze-melt	give-take
glad-sad	good-bad	great-little	happy-unhappy	hard-easy	hard-soft	hate-love
here-there	high-low	hold-drop	holiday-weekday	ill-healthy	in-out	innocent-guilty
inside-outside	kill-save	laugh-cry	leave-arrive	leave-stay	left-right	light-dark
light-heavy	like-hate	like-unlike	lose-find	lose-win	many-few	miss-catch
miss-hit	more-less	most-least	move-stop	much-little	neattidy-messy	never-ever
next-last	nobody-everybody	nothing-everything	now-then	old-new	old-young	open-closed
pain-pleasure	pass-fail	poor-rich	pull-push	punish-reward	rainy-dry	right-left
right-wrong	safe-dangerous	same-different	serious-silly	short-long	short-tall	shy-social
sleep-wake	small-big	smooth-rough	start-reach	strong-weak	take-bring	take-give
takeon-takeoff	teach-learn	thin-fat	thin-thick	town-country	true-false	war-peace
warm-cool	whole-part	win-fail	wide-narrow	with-without	yes-no	

APPENDIX B
SUPPLEMENTARY EXPERIMENTS

A. *Comparison of Augmentation Strategy*

As introduced in Methodology Section, the proposed NIAT utilizes 20% temporal modality feature erasing (#T, 20%) for the augmentation. For further analysis, the performance comparison under different noise instance construction strategies and missing rates are recorded in Table IV.

Missing Construction Strategy Comparison. Following the modality feature missing modeling for evaluation, we compared temporal modality feature erasing (#T) with random modality feature erasing (#R) and structural temporal modality feature missing (#ST) like augmentations. It can be observed that the #T strategy achieves the best performances, while the #ST strategy performs the worst under all evaluation scenarios.

Missing Rate Comparison. We compare the model performances under 10% and 40% missing construction using #T (temporal modality feature erasing) strategy. The results show simulating the noise with too low or too high missing rate does harm to the overall model performances.

For the overall performance on three evaluation scenarios, (#T, 20%) is utilized in the NIAT framework for noise construction.

TABLE IV
COMPARISON OF DIFFERENT AUGMENTATION STRATEGIES ON MOSEI, WHERE #R, #T, #ST, REFERS TO THE RANDOM, TEMPORAL, AND STRUCTURAL MISSING SCENARIO-LIKE AUGMENTATION STRATEGY. RESULTS ARE SHOWN IN ACC-2 / F1 SCORE FORMAT. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

	Random	Temporal	Structural Temporal
(#R, 20%)	73.70 / 71.93	73.68 / 71.91	68.83 / 65.52
(#ST, 20%)	66.56 / 66.08	66.46 / 65.96	60.65 / 58.93
(#T, 10%)	72.63 / 71.93	72.60 / 71.91	66.72 / 65.96
(#T, 40%)	75.95 / 74.61	75.96 / 74.62	72.88 / 70.46
(#T, 20%)	**77.81 / 75.24**	**77.79 / 75.22**	**75.29 / 71.26**

B. *Extended Case Studies for Noise Simulation*

In the appendix, we further shows a detailed version of case study to discuss whether simulating real-world imperfection by modality feature missing is reasonable. The Experimental results are shown in Figure 1. In the extended version, we further consider the additive disturbance (visual blurry and auditory additive white noise) in real-world applications. In general, from the horizontal comparison of table at bottom right, both the raw data missing, additive disturbance show similar effect on MAG-BERT and NIAT prediction. Such results including additive disturbances further reveal that the modality feature missing is a simple but effective simulation for most real-world imperfection.

REFERENCES

[1] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, “Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages,” *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.

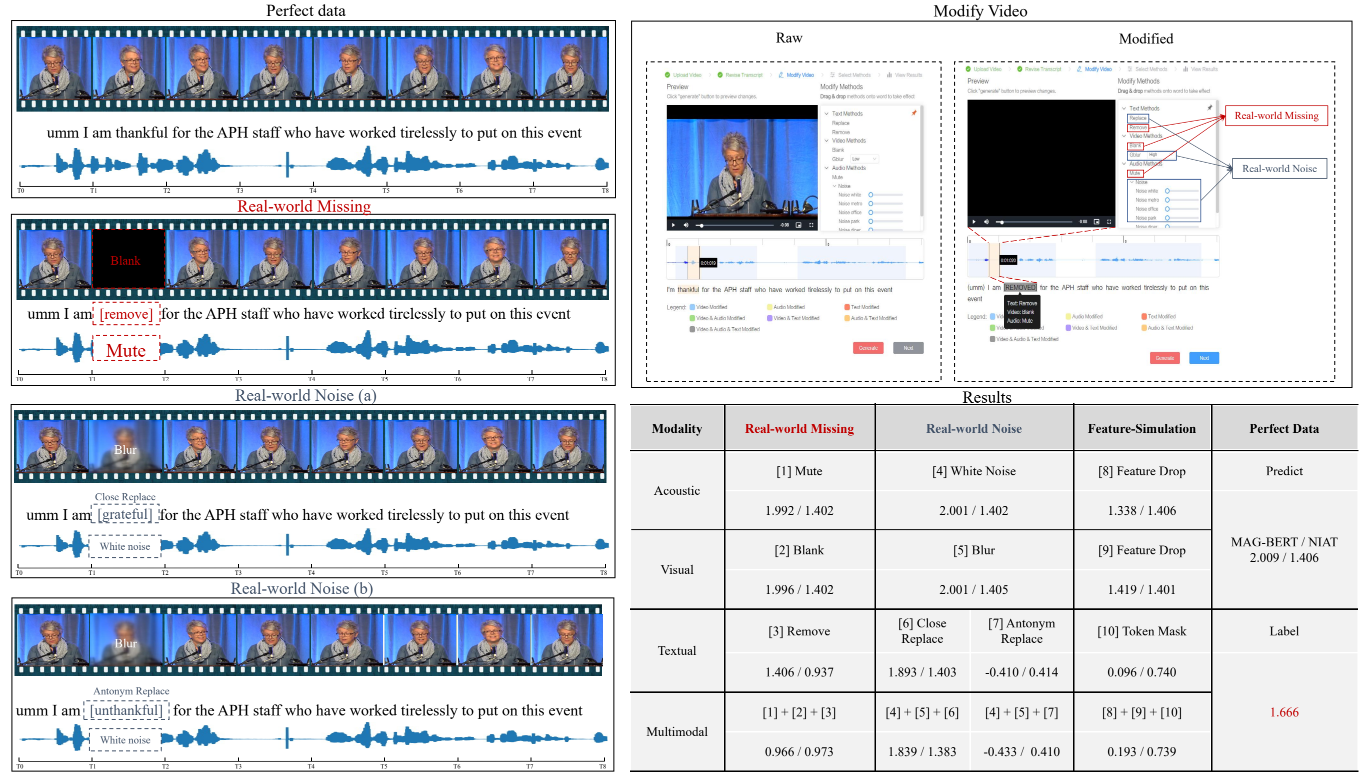
[2] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, “Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2236–2246.

[3] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, “Multimodal transformer for unaligned multimodal language sequences,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, vol. 2019, 2019, pp. 6558–6569.

[4] D. Hazarika, R. Zimmermann, and S. Poria, “MISA: modality-invariant and -specific representations for multimodal sentiment analysis,” *CoRR*, vol. abs/2005.03545, 2020. [Online]. Available: https://arxiv.org/abs/2005.03545

[5] W. Rahman, M. K. Hasan, S. Lee, A. B. Zadeh, C. Mao, L.-P. Morency, and E. Hoque, “Integrating multimodal information in large pretrained transformers,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 2359–2369.

[6] W. Yu, H. Xu, Z. Yuan, and J. Wu, “Learning modality-specific representations with self-supervised multi-task learning for multimodal


Perfect data
umm I am thankful for the APH staff who have worked tirelessly to put on this event
T0
T1
T2
T3
T4
T5
T6
T7
T8
Real-world Missing
Blank
umm I am [remove] for the APH staff who have worked tirelessly to put on this event
Mute
Real-world Noise (a)
Blur
Close Replace
umm I am [grateful] for the APH staff who have worked tirelessly to put on this event
White noise
Real-world Noise (b)
Blur
Antonym Replace
umm I am [unthankful] for the APH staff who have worked tirelessly to put on this event
White noise
Modify Video
Raw
Modified
Upload Video
Revise Transcript
Modify Video
Select Methods
View Results
Preview
Click "generate" button to preview changes.
Modify Methods
Drag & drop methods onto word to take effect
Text Methods
Replace
Remove
Video Methods
Blank
Gblur
Low
High
Audio Methods
Mute
Noise
Noise white
Noise metro
Noise office
Noise park
Noise diner
0:01:019
0:01:020
I'm thankful for the APH staff who have worked tirelessly to put on this event
(umm) I am [REMOVED] for the APH staff who have worked tirelessly to put on this event
Text: Remove
Video: Blank
Audio: Mute
Legend:
Video Modified
Audio Modified
Text Modified
Video & Audio Modified
Video & Text Modified
Audio & Text Modified
Video & Audio & Text Modified
Generate
Next
Real-world Missing
Real-world Noise
Results

Modality	Real-world Missing	Real-world Noise		Feature-Simulation	Perfect Data
Acoustic	[1] Mute	[4] White Noise		[8] Feature Drop	Predict
	1.992 / 1.402	2.001 / 1.402		1.338 / 1.406	MAG-BERT / NIAT 2.009 / 1.406
Visual	[2] Blank	[5] Blur		[9] Feature Drop	
	1.996 / 1.402	2.001 / 1.405		1.419 / 1.401	
Textual	[3] Remove	[6] Close Replace	[7] Antonym Replace	[10] Token Mask	Label
	1.406 / 0.937	1.893 / 1.403	-0.410 / 0.414	0.096 / 0.740	1.666
Multimodal	[1] + [2] + [3]	[4] + [5] + [6]	[4] + [5] + [7]	[8] + [9] + [10]	
	0.966 / 0.973	1.839 / 1.383	-0.433 / 0.410	0.193 / 0.739	

Fig. 1. A detailed version for comparison between real-world imperfection and modality feature missing simulation. Real-world imperfections are illustrated on the left side, including video frame missing, visual blurry, audio signal missing, and additive white noise in auditory. At top right, we demonstrate how to construct noisy instances using the Robust-MSA platform. Results are recorded at bottom right in format MAG-Bert / NIAT model.

sentiment analysis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, 2021, pp. 10 790–10 797.

[7] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," *arXiv preprint arXiv:1806.06176*, 2018.

[8] M. Ma, J. Ren, L. Zhao, S. Tulyakov, C. Wu, and X. Peng, "Smil: Multimodal learning with severely missing modality," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 2302–2310.

[9] W. Han, H. Chen, M.-Y. Kan, and S. Poria, "Mm-align: Learning optimal transport-based alignment dynamics for fast and accurate inference on missing modality sequences," *arXiv preprint arXiv:2210.12798*, 2022.

[10] Z. Wang, Z. Wan, and X. Wan, "Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis," in *Proceedings of The Web Conference 2020*, 2020, pp. 2514–2520.

[11] H. Pham, T. Manzini, P. P. Liang, and B. Poczos, "Seq2seq2sentiment: Multimodal sequence to sequence models for sentiment analysis," *arXiv preprint arXiv:1807.03915*, 2018.

[12] J. Tang, K. Li, X. Jin, A. Cichocki, Q. Zhao, and W. Kong, "Ctfn: Hierarchical learning for multimodal sentiment analysis using coupled-translation fusion network," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 5301–5311.

[13] Y. Wu, Y. Zhao, H. Yang, S. Chen, B. Qin, X. Cao, and W. Zhao, "Sentiment word aware multimodal refinement for multimodal sentiment analysis with asr errors," in *Findings of the Association for Computational Linguistics: ACL 2022*, 2022, pp. 1397–1406.

[14] P. P. Liang, Z. Liu, Y.-H. H. Tsai, Q. Zhao, R. Salakhutdinov, and L.-P. Morency, "Learning representations from imperfect time series data via tensor rank regularization," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 1569–1576.

[15] B. Li, C. Li, F. Duan, N. Zheng, and Q. Zhao, "Tpfn: Applying outer product along time to multimodal sentiment analysis fusion on incomplete data," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*. Springer, 2020, pp. 431–447.

[16] Z. Yuan, W. Li, H. Xu, and W. Yu, "Transformer-based feature reconstruction network for robust multimodal sentiment analysis," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4400–4407.

[17] L. Sun, Z. Lian, B. Liu, and J. Tao, "Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis," *arXiv preprint arXiv:2208.07589*, 2022.