Meta Noise Adaption Framework for Multimodal Sentiment Analysis With Feature Noise

Article *in* IEEE Transactions on Multimedia · January 2024

DOI: 10.1109/TMM.2024.3362600

Ziqi Yuan
Tsinghua University

Meta Noise Adaption Framework for Multimodal Sentiment Analysis with Feature Noise

Ziqi Yuan, Baozheng Zhang, Hua Xu*, and Kai Gao

***Abstract*—Improving the robustness of models against feature noise has emerged as one of the most crucial research topics in the field of multimodal sentiment analysis. Recent studies assume that the training instances are free of noise and develop either translation or reconstruction based method under the guidance of perfect training data for robust testing time performance. However, such an ideal assumption neglects the potential presence of the feature noise in training instances and inevitably results in degradation for the scenario where high-quality training instances are unavailable. In order to achieve robust training with noisy instances, we propose the Meta Noise Adaption (Meta-NA) learning strategy, a meta learning method accumulating the experience of dealing with various types of feature noise. Specifically, we first formulate the tasks distribution where each task is corresponding to one specific pattern of noise, and propose the feature adaption module adding on the unimodal encoder in late fusion based architecture. Through an nested online optimization between the auxiliary feature adaption module and the late fusion backbone modules, the proposed method can leverage shared knowledge across different noisy source tasks and learn how to learn from the noisy instances for robust testing performances. Extensive experiments are conducted on two benchmark multimodal sentiment analysis datasets, namely MOSI and CH-SIMS v2. The results demonstrate that our proposed method can rapidly adapt to various unseen types of feature noise and outperforms all baseline methods, particularly when the training instances are limited.**

***Index Terms*—Robust Multimodal Sentiment Analysis, Feature Noise, Late Fusion Based Architecture, Meta Learning.**

I. Introduction

With the rise of user-generated online videos, Multimodal Sentiment Analysis (MSA), which aims to analyze the speaker's sentiment through spoken words, auditory, and visual behaviors, has become increasingly relevant [1]–[3]. Numerous MSA applications have been developed with the aim of enhancing the user experience in Human-Computer Interaction (HCI) services [4]–[6]. However, the quality of data itself becomes a major concern due to the presence of background noise or transient sensor failure. As feature noise severely degrades the generalization performance of deep neural networks, learning and keeping robust for noisy instances (robust training) has become an important topic in current MSA research [7]–[9]. In this work, we address on improving the robustness of the MSA system against unknown feature noise in both training and testing instances.

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

*: Hua Xu is the corresponding author and is also with Samton (Jiangxi) Technology Development Co., Ltd., where a portion of the research was conducted in collaboration. Corresponding email: xuhua@tsinghua.edu.cn.

Our code is available at https://github.com/thuiar/Meta-NA.

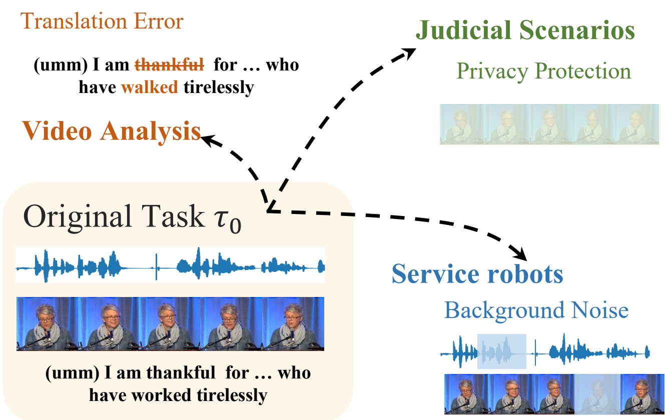


Fig. 1. An overview of multimodal sentiment analysis in real-world applications. Different application scenarios frequently give rise to varying noise patterns in multimodal sentiment analysis systems.

As illustrated in Figure 1, the fundamental challenge of achieving robust multimodal sentiment analysis lies in the fact that different application scenarios naturally give rise to varying noise patterns. For example, spoken word replacement due to the automatically speech recognition errors is commonly encountered in analyzing user-uploaded videos, while background noise in audio and visual modalities are more likely to exist in emotional service robots. In order to deal with multiple noise patterns, one simple solution is to train individual models from scratch for each type of noise [10]–[12]. However, this strategy incurs additional storage and training costs, and also requires prior knowledge of the existing feature noise, which is typically not available in practical applications. A intuitive refinement to the "one-to-one training" approach is to develop a unified model that can handle all types of noise [13], [14]. However, there are notable discrepancies between different noise patterns. For instance, feature missing in the predominant textual modality have a much greater impact than same degree feature missing in the auxiliary acoustic or visual modalities. Consequently, the optimal model parameters for one specific feature noise pattern may lead to sub-optimal results on other noise patterns.

In this work, we provide a novel meta learning perspective for the varying noise patterns challenge. Each scenario with a specific noise pattern is regarded as an individual task sampled from the general robust multimodal sentiment analysis task distribution. During the meta-training period, shared knowledge of dealing with varying types of feature noise is learned, while during the meta-testing period, the model is further fine-tuned for optimal solution of current noise pattern. In general, two levels of experiences are accumulated through the nested optimization in the proposed Meta Noise Adaption (Meta-NA) learning strategy. Firstly, inspired by the model-agnostic meta-learning, we learn a shared model initialization that serves as the fundamental solution for all noisy patterns. This shared model initialization plays a crucial role, especially in situations where limited task-specific knowledge can be obtained due to low-quality training instances during the fine-tuning in meta-testing period. Furthermore, we equip the model with the capability to learn how to denoise from the unimodal representation vectors. We develop an auxiliary noise adaptation module using a residual autoencoder to learn prior knowledge on how to alleviate the negative effects of noise on unimodal representation vectors on the late fusion architecture. The main contribution of this paper can be summarized as follows:

- Compared to previous research on robust MSA, this work addresses a more realistic scenario wherein an unknown pattern of feature noise exists in both the training and testing phases. In this context, this work presents one of the earliest efforts utilizing the meta learning perspective which treats each specific noise pattern as an individual task within the broader noisy MSA task family.
- In this paper, we propose the Meta Noise Adaption (Meta-NA) strategy, a meta-learning approach to learn shared model initialization and denoising techniques across constructed source noisy tasks, thereby enabling fast adaptation and robust performance for potential applications with unknown noise patterns.
- Extensive experiments on two benchmark MSA datasets (MOSI and SIMS v2) indicate that the propose Meta-NA strategy achieves consistent improvement for various unseen types of feature noise, especially for the cases where the training instances is limited and can be easily extended to other multimodal applications against noise.

II. Related Works

A. Robust Multimodal Sentiment Analysis

The concept of robust multimodal sentiment analysis is derived from the demand of multimedia applications where background disruptions occur unavoidably [15], [16]. Two typical forms of feature noise are considered, namely entire modality missing and fine-grained modality feature missing. For the scenario of entire modality missing, the most common solution is missing modality imputation which endeavors to approximate the missing modality from partially observed input [17], [18]. In the most recent work, Han et al. [19] utilize alignment matrices to enhance imputation performance of the missing modality. Another paradigm for entire modality missing involves modality translation, which utilizes the intermediate representation between source and target modality as a joint multimodal feature [20]–[22]. The latest effort in translation-based methods, exploiting bi-directional interplay via couple learning, is represented by the literature [23]. As for the scenario of fine-grained modality feature missing, the earliest works include literature [11], [12], which developed a low-rank regularization method based on the observation of the low-rank nature of clean data. Recently, several works have been developed with auxiliary feature reconstruction loss. Yuan et al. [10] propose a transformer-based feature reconstruction network, while Sun et al. [24] improve former low-level feature reconstruction with high-level feature attraction to achieve robust performance. However, most existing studies necessitate the coexistence of perfect instances alongside their noisy counterparts during the training phase. Specifically, missing imputation based approaches [17]–[19] and reconstruction-based approaches [10], [24] employ the perfect instances as a guiding principle for semantic reconstruction, whereas translation-based methods [20]–[23] regard the perfect instances as the target modality. To tackle the problem of feature noise in both training and testing data, we develop a meta learning method to learn how to learn from low-quality instances and keep robust to noisy testing data.

B. Multimodal Fusion Method

Fusion is a key research topic in multimodal studies, which integrates information extracted from different resources into a single compact representation [25]–[27]. According to different fusion stage, current methods can be roughly divided into early fusion, hybrid fusion, and late fusion. Early fusion methods integrate features immediately after they are extracted [28], [29]. However, due to the insufficient excavation of intra-modal interactions, early fusion architecture achieves relative low performance. In order to overcome the drawback of early fusion, hybrid fusion develops well-designed attention strategies for capturing both intra and inter modality cues [30]–[35]. Despite great improvement on aligned and clean data settings, two concerns of the hybrid fusion arise in the noisy scenarios. Firstly, the feature noise can disturb the alignment procedure [36], leading to difficulty in obtaining aligned data. Secondly, researches have verified that such sophisticated fusion methods are very sensitive to potential feature noise in both training and testing instances [15]. Late fusion method, which performs intra-modal feature extraction before the integration acts as a simple but effective method. Compared with hybrid fusion, late fusion methods achieve competitive performance with much less learnable parameters and can naturally deal with the unaligned data [37]–[39]. In this work, we choose a late fusion architecture as our backbone and further refine it to achieve robust performance against feature noise challenges.

C. Meta Learning Method

Meta-learning, which aims to improve the learning algorithm itself, has received a dramatic rise interest [40]–[42]. Several meta-learning approaches have been developed for the purpose of improving the robustness against noise [43]. For instances, Meta Loss Correction [44] learns noise transition matrix from data via the meta-learning, while Meta-weight-net [45] learns the weighting function for noisy training instances through minimizing the empirical risk of a small clean dataset.

In the field of multimodal sentiment analysis, Sun et al. [46] introduce the Adaptive Multimodal Meta-Learning to fully excavate unimodal cues via meta-training on unimodal tasks for performance improvement on clean data. Dealing with modalities noise, Ma et al. [47] design a Bayesian meta-learning method integrating missing modality reconstruction and feature regularization, while Chi et al. [48] refine the model-agnostic meta-learning approach with meta-sampling strategy. In this study, we devise the meta noise adaption strategy learning shared initialization as well as the ability to alleviate noise from the unimodal representation through nested optimization on the created noisy source tasks.

III. Preliminaries

Prior to delving into an elaborate exposition of the proposed Meta-NA framework, it is imperative to initially delineate the problem formulation and adopt a meta-learning perspective in the context of robust MSA. For easier reading, the summary of notations in our work has been shown in Table I.

TABLE I
Table of Crucial Notations.

Notations	Descriptions
τ, p_τ	individual noisy task, broader noisy tasks distribution
D_{tr}, D_{ts}	training and testing set for a given task
s	structure of the noise
r_m	degree of the noise in modality $m \in \{l, a, v\}$
$\mathbf{X}$, $\widetilde{\mathbf{X}}$	clean, noisy instances
X_m, $\widetilde{X}_m$	clean, noisy modality sequence $m \in \{l, a, v\}$
$y, \hat{y}$	sentiment intensity labels and predictions
ϕ_m	unimodal encoder for modality $m \in \{l, a, v\}$
ϕ_c	fusion and classification module
ψ_m	feature adaption module for modality $m \in \{l, a, v\}$
θ_m	learnable parameters of $\phi_m, m \in \{l, a, v\}$
θ_c	learnable parameters of ϕ_c
ω_m	learnable parameters of $\psi_m, m \in \{l, a, v\}$
f_m	unimodal feature of clean data $m \in \{l, a, v\}$
$\widetilde{f}_m$	unimodal feature of noisy data $m \in \{l, a, v\}$
$\hat{f}_m$	adapted unimodal feature of noisy data $m \in \{l, a, v\}$
α	learnable inner loop learning rate
β	fixed outer loop learning rate
E	inner loop epochs for each sampled tasks
T	outer loop epochs
η_m	weight of the disparity loss for modality $m \in \{l, a, v\}$

A. Multimodal Sentiment Analysis under Feature Noise

Traditional multimodal sentiment analysis can be regarded as a typical regression task containing training and testing data,

$$\tau \triangleq \left(D_{\text{tr}} = \{(\mathbf{X}^i_{\text{tr}}, y^i_{\text{tr}})\}^n_{i=1}, D_{\text{ts}} = \{(\mathbf{X}^j_{\text{ts}}, y^j_{\text{ts}})\}^k_{j=1} \right). \quad (1)$$

Each instance $\mathbf{X}$ consists of text, acoustic, and visual modality feature sequences, i.e. $X_m \in \mathcal{R}^{t_m \times d_m}, m \in \{l, a, v\}$, along with its sentiment intensity annotation $y \in \mathcal{R}$. In this study, we assume that feature noise can be characterized from two aspects. The first aspect models the degree of noise in each modality sequence, using the missing rate $r\%$ as an indicator. The second aspect characterizes the structure of the feature noise based on whether there is correlation between the positions of noise in the modality sequences.

Random modality feature missing refers to the scenarios where feature noise exists in unknown positions independently among each time steps. Specifically, given a preset missing rate $r\%$ for modality sequences $X_m \in \mathcal{R}^{t_m \times d_m}$, $r\% \times t_m$ time steps are randomly dropped as zero padding vector.
Structural modality feature missing refers to a special case of random modality feature missing that modality features are dropped in consecutive time steps. Specifically, given a preset missing rates $r\%$ for modality sequences $X_m \in \mathcal{R}^{t_m \times d_m}$, the starting point of the structural missing is first chosen and subsequent $r\%$ feature are dropped as zero padding vector.

In this work, to better emulate real-world scenarios, we make three basic assumptions regarding feature noise. Firstly, we assume that feature noise exists independently in textual, acoustic, and visual modality sequences, disregarding any correlation or causality of cross-modal feature noise. Secondly, we assume that each modality sequence contains same structure of feature noise, denoted as s, while the degree of the noise for each modality sequences (denoted as $r_m, m \in \{l, a, v\}$) may differ. Moreover, we assume that partial noisy instances can be manually restored for performance evaluation. Under above assumptions, given a specific type of feature noise $\epsilon = (s, (r_l\%, r_a\%, r_v\%))$, the MSA task under such feature noise can be formulated as follows,

$$\tau(\epsilon) \triangleq \left(D_{\text{tr}} = \{(\widetilde{\mathbf{X}}^i_{\text{tr}}, y^i_{\text{tr}})\}^n_{i=1}, D_{\text{ts}} = \{(\mathbf{X}^j_{\text{ts}}, \widetilde{\mathbf{X}}^j_{\text{ts}}, y^j_{\text{ts}})\}^k_{j=1} \right), \quad (2)$$

where $\widetilde{\mathbf{X}}$ is the noisy instances under feature noise ϵ, $\mathbf{X}$ refers to the manually restored clean instances, n and k refer to the train and test instance count correspondingly.

B. Meta Learning Perspective for Robust MSA

Robust MSA involving feature noise inherently forms a task distribution p_τ in which each individual task is associated with a particular type of feature noise. Instead of solving different noise pattern one by one or building one unify model, the meta learning perspective aims to found the most suitable learning algorithm parameterized by ω, which can produce the overall best learned model $\theta^*(\omega)$ on unseen task sampled from p_τ.

$$\omega = \arg\min_{\omega} \mathbb{E}_{\tau \sim p_\tau} \mathcal{L}_{\text{meta}}(\theta^*(\omega)), \quad (3)$$

$$\theta^* = \arg\min_{\theta} \sum_{(\mathbf{X}, y) \in D_{\text{tr}}} \mathcal{L}_{\text{task}}(\mathbf{X}, y; \theta, \omega), \quad (4)$$

where $\mathcal{L}_{\text{meta}}$ and $\mathcal{L}_{\text{task}}$ is the meta objective and task objective correspondingly. For robust MSA tasks, the task objective is commonly defined as the L1Loss between the sentiment prediction and the ground truth, while the meta objective can be defined as the robustness of the algorithm against noise.

IV. Methodology

In this section, we first explain the method to construct noise source tasks using original benchmark datasets in Section IV-A, and the network architecture used in Meta-NA framework in Section IV-B followed by the detailed introduction of the proposed Meta-NA approach in Section IV-C.

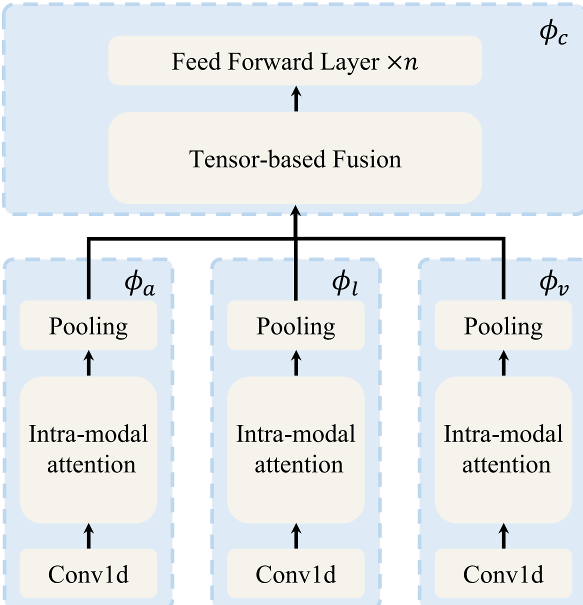


Fig. 2. Overall late fusion based backbone, which contains unimodal encoder $\phi_m, m \in \{l, a, v\}$, and fusion and classification module ϕ_c. Detailed inner structure are demonstrated in lighter blocks.

A. Source Noisy Task Construction

The construction of the source tasks contains three main steps, namely meta information sampling, instance sampling, and feature noise injection. Meta information of the source task consist of training and testing instance counts (n and k) and the pattern of feature noise ϵ. In this work, we preset the testing instance count k, and sample the training instance count range from $n_{\min}$ to $n_{\max}$ uniformly,

$$n \sim \text{Uniform}\{n_{\min}, n_{\min}+1, \cdots, n_{\max}\}. \tag{5}$$

The structure of the noise is sampled from Bernoulli distribution, i.e. with p for random modality feature missing and $1-p$ for structural modality feature missing, while the degree of feature noise is sampled from Uniform distribution,

$$s \sim \text{Bernoulli}(p), \tag{6}$$

$$r_m \sim \text{Uniform}(r_{\min}, r_{\max}), m \in \{l, a, v\}. \tag{7}$$

Given the meta information of the source task, we sample n instances from the original training set, and k instances from the original validation set, followed by instances shuffling to avoid unnecessary memorization of the position. The final step of source task construction is injecting the feature noise ϵ into all training and testing instances. Utilizing above methods, we can obtain source tasks with various training instance counts and various feature noise patterns, represented as $\{\tau_i(\epsilon_i)\}_{i=1}^{T}$.

B. Network Structure

The network structure utilized in the proposed Meta-NA contains a late fusion based backbone and the auxiliary feature adaption module $\psi_m, m \in \{l, a, v\}$ added on each modality encoder. The inner structure of the late fusion based backbone is illustrated in Figure 2, which first extracts effective unimodal representations with unimodal encoder $\phi_m(\cdot;\theta_m)$ and then perform fusion and classification with module $\phi_c(\cdot;\theta_c)$. To facilitate the detailed explanation of each module, we will first introduce the feed-forward network, which serves as a fundamental building block in the Meta-NA network structure.

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

$$\text{FFN}(x) = \sigma(W_f \cdot x + b_f), \tag{8}$$

where σ represents the optional activation function, W_f and b_f are learnable model parameters.

*1) **Unimodal Encoder Module** ϕ_m:* The modality specific encoder accumulate the emotional cues of each unimodal sequences and produce the unimodal representation for further fusion. For each modality $m \in \{l, a, v\}$, the modality sequence X_m is initially fed into 1D convolutional layer to consolidate the information pertaining to adjacent elements,

$$H_m = \text{Conv1d}(X_m) \in \mathcal{R}^{t_m \times d_m}, \tag{9}$$

where d_m is the hidden dimension for modality $m \in \{l, a, v\}$. Intra-modal multi-head attention mechanism is then applied to explore the long-time dependence in each unimodal sequence,

$$\overline{H}_m = \text{Intra-Attn}(H_m) = \text{Concat}[\overline{H}_m^{[1]}, \cdots, \overline{H}_m^{[h]}], \tag{10}$$

$$\overline{H}_m^{[i]} = \text{Sigmoid}(\frac{Q_i \cdot K_i}{\sqrt{d_m}}) \cdot V_i, \tag{11}$$

where Q_i, K_i, V_i is transformed from H_m through separate feed-forward layer, and h is the head count. Then, max-pooling is utilized for utterance level unimodal representation f_m,

$$f_m = \text{MaxPool}(\overline{H}_m) \in \mathcal{R}^{d_m}. \tag{12}$$

We denote the overall operation of the unimodal encoder as,

$$f_m = \phi_m(X_m; \theta_m), m \in \{l, a, v\}. \tag{13}$$

*2) **Fusion and Classification Module** ϕ_c:* Receiving the unimodal representation combination (f_l, f_a, f_v), tensor fusion proposed in [38] is adopted, which takes outer product of each unimodal representation to capture the cross-modal dynamics,

$$f = \begin{bmatrix} f_l \\ 1 \end{bmatrix} \otimes \begin{bmatrix} f_a \\ 1 \end{bmatrix} \otimes \begin{bmatrix} f_v \\ 1 \end{bmatrix}. \tag{14}$$

The obtained fusion representation is fed into a two layers feed-forward layer for the final sentiment prediction,

$$\hat{y} = \text{FFN}(\text{FFN}(f)). \tag{15}$$

The overall operation of above module is denoted as,

$$\hat{y} = \phi_c(f_l, f_a, f_v; \theta_c) \in \mathcal{R}. \tag{16}$$

*3) **Feature Adaption Module** ψ_m:* Inspired from shifting the classification prediction for label noise challenge, we propose the feature adaption module to mitigate the negative effects of feature noise on learned unimodal representations. Residual autoencoder is introduced. Given the extracted unimodal representation $\widetilde{f}_m$ of the noisy instances, the feature

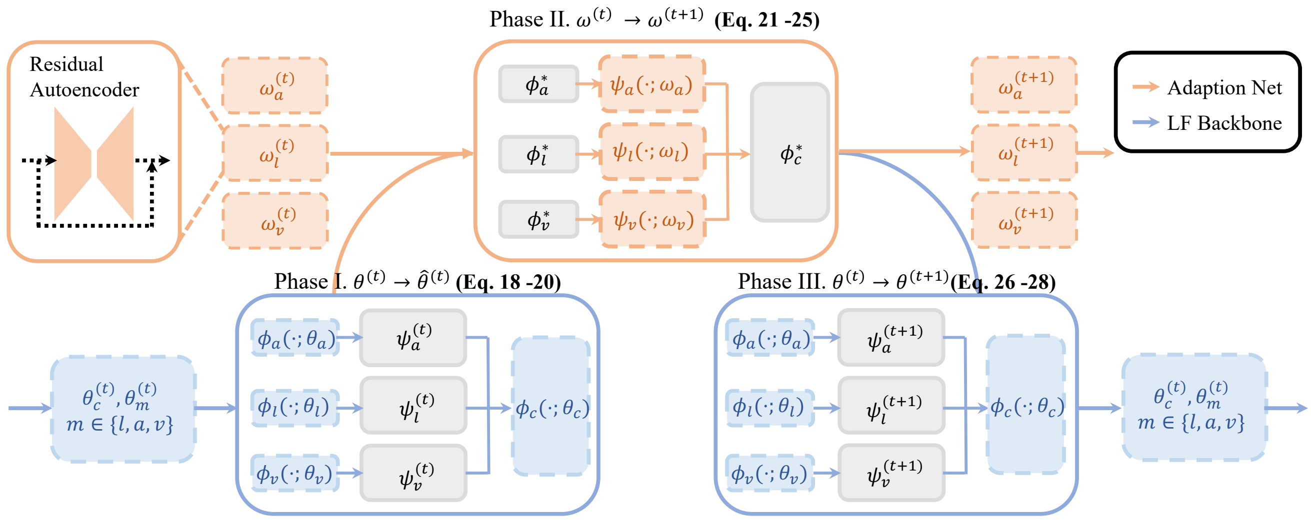


Fig. 3. Main flowchart of the proposed Meta-NA algorithm. The orange lines demonstrate the updating pipeline of the feature adaption module, the blue lines refers to the updating pipeline of the late fusion based architecture (denoted as LF Backbone). Fixed parts are illustrated in grey.

adaption module parameterized by ω_m is defined as,

$$\psi_m(\widetilde{f}_m;\omega_m) = \text{FFN}(\text{FFN}(\widetilde{f}_m)) + \widetilde{f}_m \in \mathcal{R}^{d_m}, \tag{17}$$

where the first feed-forward layer compresses the representation into $d_h = d_m/2$, while the second feed-forward layer restores the compressed representation into d_m dimension. LeakyReLU are used as activation function.

Algorithm 1 Noise Adaption Multimodal Meta-learning.

Input: Task distribution p_τ, learning rate β

Output: $\theta^{(T)}$, $\omega^{(T)}$

```
Learnable parameters initialization θ^(0), ω^(0), α^(0).
for t = 0 to T − 1 do
  Sampling the noisy source task τ_i^t(ε_i) ~ p_τ
  // Inner training phase
  for e = 0 to E − 1 do
    Pseudo-updating the late fusion based architecture
    through Eq. (18) - (20).
  end for
  // Inner evaluation phase
  Updating the feature adaption module ω^(t+1) and inner
  learning rate α^(t+1) through Eq. (21) - (25).
  // Outer training phase
  Updating the shared initialization of the late fusion
  based architecture θ^(t+1) by Eq. (26) - (28).
end for
```

C. Meta Noise Adaption Strategy

We design the meta noise adaption strategy to train the feature adaption module and the late fusion based architecture through an online optimization loop. The overall training procedure of the Meta-NA is depicted in Algorithm 1, while the intuitive illustration of the entire loop at time step t is presented in Figure 3. At time step t, for sampled source task with the training set $D_{\text{tr}} = \{(\widetilde{\mathbf{X}}_{\text{tr}}^i, y_{\text{tr}}^i)\}_{i=1}^{n_t}$ and the testing set $D_{\text{ts}} = \{(\mathbf{X}_{\text{ts}}^j, \widetilde{\mathbf{X}}_{\text{ts}}^j, y_{\text{ts}}^j)\}_{j=1}^{k_t}$, the proposed nested optimization can be divided into three main steps.

Firstly, in the inner training phase, stochastic gradient descent using the training set of task are conducted to perform pseudo-updating for the late fusion based architecture from the learned initialization $\theta^{(t)} = (\theta_l^{(t)}, \theta_a^{(t)}, \theta_v^{(t)}, \theta_c^{(t)})$, with the learned feature adaption module $\psi_m(\cdot;\omega_m^{(t)}), m \in \{l,a,v\}$ and the learned inner learning rate $\alpha^{(t)}$,

$$\hat{f}_m^i = \psi_m\left(\phi_m(\widetilde{\mathbf{X}}_{tr}^i;\theta_m^{(t)});\omega_m^{(t)}\right), m \in \{l,a,v\}, \tag{18}$$

$$\hat{y}^i = \Phi_c\left(\hat{f}_l^i, \hat{f}_a^i, \hat{f}_v^i;\theta_c^{(t)}\right), \tag{19}$$

$$\hat{\theta}^{(t)} = \theta^{(t)} - \alpha^{(t)} \circ \nabla_\theta \sum_{i=1}^{n_t} \text{L1Loss}(\hat{y}^i, y_{\text{tr}}^i), \tag{20}$$

where $\hat{\theta}^{(t)}$ refers to the task specific backbone parameters after pseudo-updating, $\alpha^{(t)}$ represents the learnable inner learning rate, and $\circ$ is utilized for element wised production. Secondly, the inner evaluation stage aims to updates the feature adaption module and inner learning rate with the testing set D_{ts}, both manually restored clean instances and the original noisy instances are feed into the unimodal encoder,

$$f_m^i = \phi_m(\mathbf{X}_{\text{ts}}^i;\hat{\theta}_m^{(t)}), m \in \{l,a,v\}, \tag{21}$$

$$\widetilde{f}_m^i = \phi_m(\widetilde{\mathbf{X}}_{\text{ts}}^i;\hat{\theta}_m^{(t)}), m \in \{l,a,v\}. \tag{22}$$

Then the unimodal representation of the noisy instances are passed into the noise feature adaption module to obtaining the denoised unimodal representation,

$$\hat{f}_m^i = \psi_m(\widetilde{f}_m^i;\omega_m^{(t)}), m \in \{l,a,v\}. \tag{23}$$

As for the meta objective, the disparity between the representation of the restored clean instance f_m^i and the adapted noisy representation $\hat{f}_m^i$ are utilized to provide explicit guidance for better denoise performance. In addition to the disparity, the testing performance of the noisy instances are also utilized as implicit guidance for the feature adaption module and the inner learning rate. The meta objective is defined as,

$$\mathcal{L}_{\text{meta}}^i = \text{L1Loss}(\hat{y}^i, y_{\text{ts}}^i) + \sum_{m\in\{l,a,v\}} \eta_m \|f_m^i - \hat{f}_m^i\|_1, \quad (24)$$

where $\hat{y}^i = \phi_c(\hat{f}_l^i, \hat{f}_a^i, \hat{f}_v^i; \hat{\theta}_c^{(t)})$ is the sentiment prediction using denoised unimodal representation, and $\eta_m, m \in \{l, a, v\}$ refers to the weight of the disparity loss for modality m. Under the guidance of meta objective, the feature adaption module and the inner learning rate are updated as below,

$$(\omega^{(t+1)}, \alpha^{(t+1)}) = (\omega^{(t)}, \alpha^{(t)}) - \beta \cdot \nabla_{(\omega,\alpha)} \sum_{i=1}^{k_t} \mathcal{L}_{\text{meta}}^i, \quad (25)$$

where β is the outer loop learning rate. As the last step, with the updated feature adaption module, the shared initialization of the late fusion based architecture is updated as below,

$$\hat{f}_m^i = \psi_m\left(\phi_m(\widetilde{\mathbf{X}}_{tr}^i; \theta_m^{(t)}); \omega_m^{(t+1)}\right), m \in \{l, a, v\}, \quad (26)$$

$$\hat{y}^i = \phi_c(\hat{f}_l^i, \hat{f}_a^i, \hat{f}_v^i; \theta_c^{(t)}), \quad (27)$$

$$\theta^{(t+1)} = \theta^{(t)} - \beta \cdot \nabla_\theta \text{L1Loss}(\hat{y}^i, y_{\text{tr}}^i), \quad (28)$$

where $\theta^{(t+1)} = (\theta_l^{(t+1)}, \theta_a^{(t+1)}, \theta_v^{(t+1)}, \theta_c^{(t+1)})$ refers to the updated shared initialization for the late fusion based backbone.

The updated initialization $\theta^{(t+1)}$ along with the feature adaption module $\psi_m(\cdot; \omega_m^{(t+1)}), m \in \{l, a, v\}$ and the inner learning rate $\alpha^{(t+1)}$ is then used for the next sampled task.

V. Experimental Setups

A. Datasets

In this paper, based on two considerations, we utilize the MOSI [49] and CH-SIMS v2 [50] datasets for the experiment. Cultural factors are the first level consideration, where MOSI and CH-SIMS v2 are the most popular English and Chinese MSA benchmark dataset correspondingly. The effect of the non-verbal cues becomes the second level consideration. Most instances in MOSI dataset are verified to be predominant on the textual modality, while the CH-SIMS v2 dataset extends the original CH-SIMS dataset [51] for the purpose of making non-verbal behaviours significant for the sentiment prediction. For experiments on MOSI, audio and visual features provided by CMU-Multimodal SDK[1] are utilized, while for experiments on CH-SIMS v2, audio and visual features from the SIMS v2.0 website[2] are utilized. Textual modality features are extracted using the pretrained Bert [52] on English and Chinese language for MOSI and CH-SIMS correspondingly. All experiments are conducted under unaligned setting. Detailed characteristics of these two datasets are left in Appendix.

B. Baseline Methods

In order to evaluate the effectiveness of Meta-NA, we make comparison with three levels of baseline methods.

Conventional MSA Methods are employed as basic level baselines. In particular, typical methods including the Multimodal Transformer (MulT) [53], the Modality-Invariant and -Specific Representations (MISA) [54], and the Self-supervised Multi-task Multimodal sentiment analysis network (Self-MM) [39] are first selected. Additionally, recent advancements that prioritize text as the dominant modality have also been included for comparative purposes. These include the Text Enhanced Transformer Fusion Network (TETFN) [31] and the Cross-modal Enhancement Network (CENet) [30].

Robust MSA Methods are utilized as the advance level baselines. For entire modality missing, the Missing Modality Imagination Network (MMIN) [18], and Coupled-Translation Fusion Network (CTFN) [23] representing the missing imputation and translation based method, are included. For fine-grained modality feature missing, the Temporal Tensor Fusion Network (T2FN) [11], the Time Product Fusion Network (TPFN) [12] and the Transformer-based Feature Reconstruction Network (TFR-Net) [10], which are typical low-rank regularization and reconstruction based method, are selected.

Meta-learning Methods are considered as the last type of baselines. We compare the proposed approach with the vanilla Model-Agnostic Meta-Learning (MAML) [55] strategy and Meta-SGD [56], a SGD-like meta-learner, building on the same late fusion based architecture.

C. Evaluation Metrics

For each individual noise pattern, robust MSA is formulated as a regression problem with mean absolute error (MAE) and Correlation Coefficient (Corr) as the primary metric. In order to facilitate a more intuitive comparison, binary accuracy and F1-Score metrics in the format of negative/non-negative are used as classification criteria. For all above metrics, higher values indicate better model performance, except for MAE, where lower values are indicative of better model performance.

For the purpose of quantitatively evaluate the performance for fine-grained feature noise in random and structural modality feature missing scenarios, the Area Under Indicators Line Chart (AUILC) metric proposed in [10] is adopted. 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,

$$\sum_{i=0}^{t-1} \frac{(e_i + e_{i+1})}{2} \cdot (r_{i+1} - r_i). \quad (29)$$

In the remainder of this paper, unless otherwise specified, we report the quantitative performances under missing rates interval $\{0.0, 0.1, \cdots, 1.0\}$ for fine-grained feature noise.

D. Experimental Details

All experiments are performed using the PyTorch on Tesla V100 with CUDA 11.7 and torch 1.13.1. The hyperparameters selection is provided in the Appendix. Baselines are trained from scratch for each noise pattern. For a fair comparison, we conduct experiments three times with different random seeds and report the average performance on the testing set.

[1]https://github.com/prateekvij/CMU-MultimodalDataSDK
[2]https://thuiar.github.io/sims.github.io/chsims

TABLE II
PERFORMANCE COMPARISON OF STRUCTURAL AND RANDOM MODALITY FEATURE MISSING ON MOSI AND SIMS V2. FOR EACH NOISE PATTERN, AUILC VALUES UNDER THE MISSING RATE INTERVAL $\{0.0, 0.1, \cdots, 0.9, 1.0\}$ ARE RECORDED. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Dataset	Model	Structural Modality Feature Missing				Random Modality Feature Missing			
		Acc-2 (↑)	F1 (↑)	MAE (↓)	Corr (↑)	Acc-2 (↑)	F1 (↑)	MAE (↓)	Corr (↑)
MOSI	MISA [54]	61.88	60.85	1.292	0.311	62.70	62.33	1.282	0.319
	MulT [53]	64.46	64.04	1.268	0.331	64.10	63.07	1.255	0.335
	Self-MM [39]	65.33	65.31	1.243	0.361	64.52	64.37	1.251	0.354
	TETFN [31]	64.94	63.18	1.223	0.360	65.61	63.53	1.218	0.391
	CENet [30]	65.04	63.05	1.224	0.374	65.93	65.02	1.222	0.396
	TPFN [12]	63.08	62.65	1.269	0.342	63.16	62.87	1.272	0.350
	T2FN [11]	63.22	62.60	1.296	0.330	62.80	61.96	1.292	0.337
	TFR-Net [10]	61.89	60.77	1.277	0.316	62.09	61.04	1.278	0.335
	CTFN [23]	65.00	57.34	1.324	0.273	64.17	55.12	1.363	0.258
	MMIN [18]	64.07	54.86	1.271	0.296	63.85	54.15	1.292	0.294
	MAML [55]	64.47	63.85	1.256	0.352	65.43	65.30	1.248	0.375
	Meta-SGD [56]	64.25	63.76	1.259	0.345	64.67	64.43	1.254	0.373
	Meta-NA	**65.62**	**65.47**	**1.213**	**0.391**	**67.14**	**67.06**	**1.193**	**0.416**
SIMS v2	MISA [54]	74.14	73.97	0.373	0.583	73.42	73.31	0.381	0.568
	MulT [53]	73.77	72.85	0.377	0.552	74.25	73.35	0.370	0.562
	Self-MM [39]	74.27	74.03	0.373	0.566	74.95	74.71	0.367	0.573
	TETFN [31]	69.59	67.40	0.420	0.457	71.46	69.08	0.411	0.487
	CENet [30]	72.39	71.55	0.388	0.525	71.52	70.82	0.401	0.510
	TPFN [12]	73.20	73.03	0.380	0.551	73.60	73.46	0.376	0.557
	T2FN [11]	73.35	72.52	0.386	0.538	73.08	72.85	0.375	0.551
	TFR-Net [10]	73.43	72.56	0.386	0.528	72.91	72.08	0.384	0.530
	CTFN [23]	65.66	63.55	0.408	0.504	63.40	60.47	0.418	0.474
	MMIN [18]	70.14	69.44	0.383	0.527	69.42	68.60	0.389	0.523
	MAML [55]	73.89	73.16	0.369	0.562	74.22	73.40	0.369	0.568
	Meta-SGD [56]	74.18	73.36	0.371	0.563	74.02	73.25	0.365	0.575
	Meta-NA	**75.35**	**74.86**	**0.355**	**0.587**	**75.31**	**74.95**	**0.347**	**0.601**

VI. RESULTS AND ANALYSIS

A. Performance for Fine-grained Feature Noise

In this subsection, we present the performance comparison for fine-grained random and structural modality feature missing in both training and testing instances.

1) Quantitative Results for Balanced Setting: We first compare the Meta-NA with baselines under the scenarios where the degree of the feature noise is the same across different modalities, i.e. $r_l\% = r_a\% = r_v\%$. Accordingly, experiments under the missing rates interval $\{0.0, 0.1, \cdots, 1.0\}$ for both structural and random modality feature missing are conducted. We record the AUILC values on MOSI and SIMS v2 in Table II. Observations can be summarized from two aspects.

Model Comparison Aspect. Firstly, it can be found that the proposed Meta-NA outperforms all baseline methods in terms of all metrics, for both structural and random feature missing on MOSI and SIMS v2 datasets. Notably, the proposed approach achieves an average improvement of 2.3% and 3.4% on primary MAE metrics for structural and random modality feature missing, respectively. Secondly, it is evident that models with late fusion-based architecture, such as Self-MM, T2FN, and TPFN, perform better than models with hybrid fusion utilizing an sophisticated attention mechanism. Such result verifies that the late fusion based architecture with less learnable parameter is more robust to feature noise avoiding overfitting on noisy training instances. Meanwhile, text-focused baselines (TETFN, CENet) perform well on the MOSI

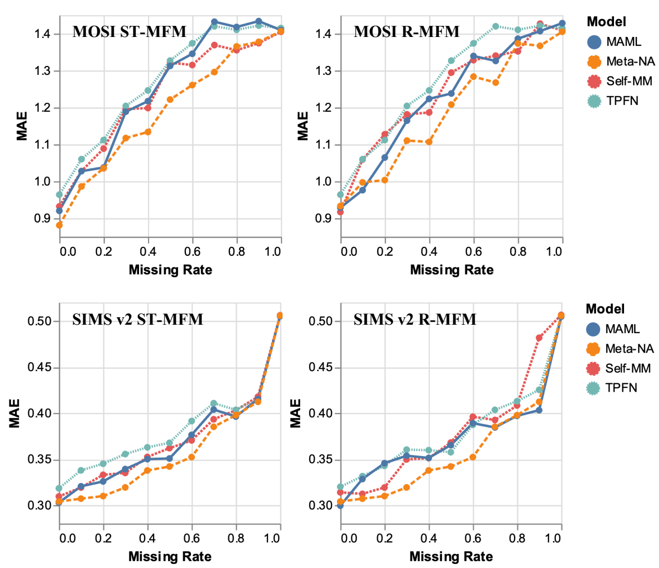


Fig. 4. Qualitative comparison between the proposed Meta-NA method with three representative baseline methods for both structural and random modality feature missing on MOSI and CH-SIMS v2 datasets.

but struggle with the SIMS v2 dataset. This outcome further validates the significance of non-verbal cues in effectively sentiment predicting under potential feature noise. Furthermore, it is worth noticing that all meta learning strategies, namely vanilla MAML and Meta-SGD, demonstrate competitive re-

TABLE III
PERFORMANCE COMPARISON OF UNBALANCED MODALITY FEATURE MISSING ON MOSI AND SIMS V2 DATASETS. FOR EACH NOISE STRUCTURE, THE MISSING RATE $\in [0.2, 0.4]$ ARE PROVIDED. THE MIXED FEATURE MISSING (50%) IS A COMPROMISE OF PURE STRUCTURAL AND RANDOM FEATURE MISSING, WHERE EACH INSTANCE MAY CONTAIN EITHER RANDOM OR STRUCTURAL TYPE OF NOISE WITH PROBABILITY 50% EACH. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Dataset	Model	Structural Feature Missing			Random Feature Missing			Mixed Feature Missing (50%)		
		Acc-2 (↑)	F1 (↑)	MAE (↓)	Acc-2 (↑)	F1 (↑)	MAE (↓)	Acc-2 (↑)	F1 (↑)	MAE (↓)
MOSI	Self-MM [39]	68.26	68.38	1.174	68.34	68.33	1.237	70.17	70.21	1.173
	TPFN [12]	67.04	67.08	1.214	67.73	67.86	1.209	68.90	68.97	1.198
	MAML [55]	70.54	70.61	1.149	70.89	70.90	1.119	70.43	70.08	1.138
	Meta-NA	**72.09**	**72.08**	**1.112**	**72.56**	**72.54**	**1.084**	**72.92**	**72.92**	**1.092**
SIMS v2	Self-MM [39]	76.66	76.78	0.344	77.53	77.62	0.339	77.27	77.38	0.341
	TPFN [12]	74.24	74.30	0.364	74.63	74.76	0.356	74.73	74.81	0.365
	MAML [55]	76.34	76.42	0.340	77.02	77.14	0.329	77.05	77.11	0.335
	Meta-NA	**78.02**	**78.02**	**0.322**	**78.53**	**78.62**	**0.313**	**77.85**	**77.91**	**0.325**

sults compared to conventional MSA methods and robust MSA methods, indicating the efficacy of meta learning perspective in addressing feature noise in both training and testing instances. **Imperfection Comparison Aspect.** From the experimental result, it can be observed that in general, structural modality feature missing is more challenging than random modality feature missing, especially on the MOSI dataset. Such phenomenon can be result from the lack of non-verbal cues and shorter acoustic and visual sequence lengths.

2) Qualitative Results for Balanced Setting: The diagram illustrated in Figure 4 illustrates the MAE curves of the proposed Meta-NA approach and typical baseline methods for scenarios where features missing exist in either the structural or random mode. In general, the proposed Meta-NA effectively mitigates the degrading trend in model performance as the degree of feature noise increased and achieves superior performance in most cases. Furthermore, in contrast to the MOSI dataset, the MAE curves on the SIMS v2 dataset displayed a higher degree of smoothness as the missing rate increased. It is reasonable because due to the more expressive non-verbal behaviours of the SIMS v2 dataset, which offers better modality complementarity. Even when some crucial textual information was absent, the model could still accurately assess the speaker's overall sentiment intensity by relying on partial information from other modalities.

3) Results for Unbalanced Setting: In this subsection, we consider a more general unbalance experimental setup compared to previous experiments. We presume that both the training and testing instances encompass a certain level of low-level feature noise across all modalities. Specifically, within each instance, the degree of feature absence ranges from 20% to 40% in all modality sequences. Under these circumstances, we evaluate the performance of the proposed Meta-NA in comparison to three representative baseline methods on the MOSI and SIMS v2 datasets. The evaluation is conducted under three different scenarios: structural feature missing, random feature missing, and mixed feature missing. In the mixed feature missing scenario, each instance has an equal probability of containing either structural or random feature missing, with a probability of 50%. The experimental results, as presented in Table III, demonstrate that models trained with meta-

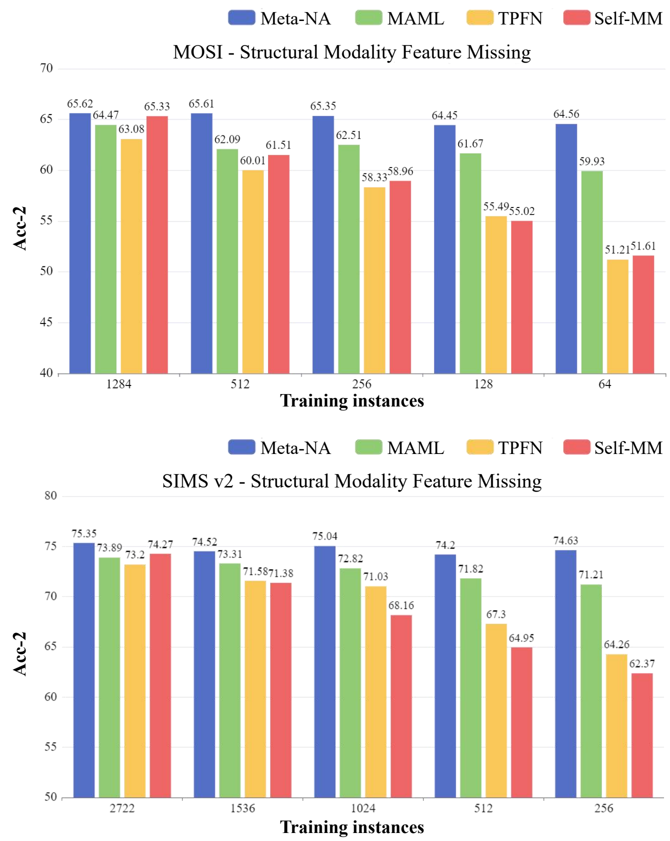


Fig. 5. Performance comparison between the proposed Meta-NA and three typical baselines selected from conventional MSA methods, robust MSA methods, and meta-learning methods under the situation where limited training instances of target noise pattern is provided.

learning techniques achieve superior performance compared to conventional and robust MSA methods. This reveals the effectiveness of meta-learning methods in generalizing to low-level and unbalanced modality noise settings. Additionally, the proposed Meta-NA framework, utilizing the meta noise adaption strategy, achieves the best performance. Specifically, it demonstrates an average improvement of 2.6% and 1.2% on the Acc-2 criteria for above three noise structures on the MOSI dataset and SIMS v2 dataset, respectively. Besides,

TABLE IV
PERFORMANCE COMPARISON OF MODALITY MISSING ON MOSI AND CH-SIMS V2.0 DATASETS. THE BEST RESULTS ARE EMPHASIZED IN BOLD.

Dataset	Model	Text Modality Missing			Acoustic Modality Missing			Visual Modality Missing		
		Acc-2 (↑)	F1 (↑)	MAE (↓)	Acc-2 (↑)	F1 (↑)	MAE (↓)	Acc-2 (↑)	F1 (↑)	MAE (↓)
MOSI	MISA [54]	54.72	54.33	1.472	78.00	78.11	1.007	77.29	77.35	0.999
	MulT [53]	58.89	58.80	1.384	79.22	79.18	0.923	78.51	78.50	0.953
	Self-MM [39]	**60.82**	**60.93**	**1.330**	79.42	79.45	0.963	79.17	79.20	0.983
	TETFN [31]	52.18	43.04	1.453	80.34	80.33	0.916	79.78	79.80	0.941
	CENet [30]	51.98	49.49	1.423	80.18	80.21	1.017	80.59	80.66	1.023
	TPFN [12]	54.07	51.50	1.426	77.54	77.55	0.923	78.05	78.10	0.934
	T2FN [11]	57.47	57.35	1.393	80.34	80.27	0.906	79.47	79.45	0.900
	TFR-Net [10]	54.11	52.72	1.410	78.10	78.16	0.946	79.47	79.52	0.905
	CTFN [23]	54.53	45.56	1.417	79.46	79.45	0.940	80.21	80.19	0.926
	MMIN [18]	53.20	51.94	1.579	80.28	80.30	0.935	80.23	80.27	0.934
	MAML [55]	58.03	56.93	1.399	80.95	80.88	0.896	80.28	80.26	0.893
	Meta-SGD [56]	55.54	54.77	1.400	81.61	81.39	**0.871**	80.34	80.22	0.878
	Meta-NA	59.50	59.63	1.357	**81.71**	**81.53**	0.886	**81.00**	**80.78**	**0.874**
SIMS v2	MISA [54]	70.79	70.73	0.422	81.36	81.47	0.305	76.76	76.71	0.345
	MulT [53]	72.99	72.77	0.388	79.37	79.32	0.323	76.31	76.33	0.337
	Self-MM [39]	73.39	73.16	0.393	81.16	81.26	0.309	76.90	76.91	0.346
	TETFN [30]	72.69	71.91	0.447	76.85	76.84	0.358	76.18	76.12	0.359
	CENet [30]	66.38	61.31	0.468	77.66	77.66	0.337	76.34	76.43	0.341
	TPFN [12]	71.37	70.98	0.407	78.34	78.43	0.326	76.98	**76.96**	0.347
	T2FN [11]	72.24	72.07	0.393	79.69	79.76	0.318	74.86	74.69	0.368
	TFR-Net [10]	**74.98**	**74.71**	**0.376**	74.66	74.64	0.364	72.60	72.66	0.372
	CTFN [23]	62.11	62.25	0.477	81.73	81.71	0.311	71.86	71.86	0.382
	MMIN [18]	72.14	71.77	0.390	80.79	80.85	0.317	76.69	76.70	0.340
	MAML [55]	72.54	72.53	0.381	80.82	80.91	0.300	76.08	76.14	0.337
	Meta-SGD [56]	72.21	72.22	0.385	81.98	82.05	0.297	76.92	76.88	0.338
	Meta-NA	71.44	71.45	0.383	**82.50**	**82.56**	**0.287**	**77.02**	76.95	**0.334**

more experimental results of unbalanced settings under higher noise interval (40% to 60%) are recorded in Appendix.

B. Results for Entire Modality Missing

In this setup, one of the textual, acoustic, or visual modality sequences is completely removed in both training and testing instances. Table IV presents the comparison of the Meta-NA approach with the baseline methods on the MOSI and SIMS v2 dataset. For entire acoustic and visual modality missing scenarios, the proposed Meta-NA achieves the overall best performance. Despite the overall advantages, it can be observed that the Meta-SGD method shows similar performance on these scenarios. Such results reveals that the benefits for entire modality missing is mainly contributed to the shared prior knowledge acquiring from the tasks distribution, while the feature adaption module which aims to migrate the negative effects of noise in unimodal representation is not efficient due to the difficulty in recovering unimodal representations under the condition of completely missing modality scenarios. Meanwhile, for entire textual modality missing, Self-MM and TFR-Net performs best on the MOSI and SIMS v2 dataset respectively. This phenomenon can potentially be attributed to the notable disparity between the entity text modality missing and other common noise patterns, consequently resulting in poor performance of shared prior knowledge acquisition from the common tasks distribution under entity text modality missing. Although the proposed method achieves slightly worse performance than some baselines, the difference on primary MAE indicator is not significant. These results validate the effectiveness of the Meta-NA for dealing entire modality missing utilizing meta learning perspective.

C. Fast Adaption Analysis

Recognizing that one of the key advantages of the meta-learning approach is its ability to quickly adapt to unseen task under few-shot setting, in addition to the quantitative experiment using all training instances, we also evaluate the model's robustness under fine-grained modality feature noise in such settings. In this regard, we retained the same experimental conditions as Section VI-A1 except for using only partial training instances instead of entire training set in each individual task. Specifically, for the MOSI dataset, we conducted experiments on $\{1284, 512, 256, 128, 64\}$ selected training instances, whereas for the SIMS v2 dataset, we selected $\{2722, 1536, 1024, 512, 256\}$ instances as the training set. The trend of AUILC value changes is illustrated in Figure 5. Our findings indicate that for both the MOSI and SIMS v2 datasets, the performance of conventional MSA methods decreases most rapidly as the number of training instances decreases (shows 21% degradation on MOSI with 64 training instances, and 16% degradation on SIMS v2 with 256 training instances), while the meta-learning methods demonstrate the most stable performance. By leveraging the proposed meta noise adaption strategy, the stability of the meta learning approach can be further improved. Specifically, the proposed Meta-NA approach only shows 1.6% degradation on MOSI

TABLE V
CONVERGENCE ANALYSIS OF THE PROPOSED META-NA APPROACH ON THE MOSI DATASET. THE PERFORMANCE ON SIMILAR NOISY SOURCE TASKS ARE RECORDED ALONG WITH THE CORRESPONDING TASK DETAILS.

N-Task	Task Settings			Acc-2
	n-Sup	n	$(r_t\%, r_a\%, r_v\%)$	Train / Test
9	152	ST	(14%, 68%, 34%)	57.53 / 56.56
21	356	RD	(18%, 72%, 6%)	66.28 / 62.81
55	456	ST	(19%, 7%, 76%)	67.51 / 69.42
82	136	RD	(13%, 30%, 77%)	76.15 / 73.95
101	401	ST	(14%, 32%, 37%)	80.83 / 78.69
180	297	ST	(11%, 25%, 24%)	80.07 / 78.51
238	154	RD	(13%, 13%, 50%)	81.51 / 82.64
292	336	RD	(12%, 29%, 67%)	84.06 / 83.76
371	400	ST	(16%, 56%, 36%)	84.90 / 80.33
396	209	ST	(10%, 40%, 21%)	86.00 / 86.67

with 64 training instances, and 1.0% degradation on SIMS v2 with 256 training instances. Such above performance underscores the efficiency of the Meta-NA approach for applications where only a few training instances are available.

D. Convergence Analysis

In this subsection, we present the empirical convergence analysis of the proposed Meta-NA approach. We record the binary accuracy metrics of the proposed Meta-NA during the inner training phase and inner evaluation phase, along with the corresponding details of the source task. To demonstrate the learning process of the Meta-NA from source tasks more intuitively, we select source tasks that share a similar missing rate for each modality sequence (approximately 15% for $r_t\%$, 30% for $r_a\%$, 50% for $r_v\%$). Experimental results are provided in Table V. According to the results, it is evident that the proposed Meta-NA improves its performances on dealing with instances with similar noise patterns as the seen noisy source tasks increase. These results validate the convergence of the proposed Meta-NA approach and further demonstrate its ability to efficiently accumulate prior knowledge of tackling noisy instances through the constructed source tasks.

E. Ablation Studies

In this subsection, we perform ablation studies to investigate the contribution of the selected meta learning strategy and fusion strategy. Experiments are conducted for both two types of modality feature missing on both datasets under the same experimental conditions as Section VI-A1. The results of the MOSI dataset are presented in Table VI, while the results of the SIMS v2 dataset are shown in Table VII.

1) Analysis on Meta Learning Strategy: Firstly, we ablate the entire meta learning strategy, and train the late fusion backbone illustrated in Figure 2 from scratch for each noise pattern, denoted as **- Meta Learning**. It can be observed that the removal of the entire meta learning strategy degrades the Meta-NA into a conventional late fusion base methods and further leads to an average decrease of 3.47% in the Acc-2 metric on both datasets. The sharp performance drop indicates

TABLE VI
ABLATION STUDY RESULTS ON MOSI DATASET. THE PROVIDED RESULT IS REPORTED IN STRUCTURAL / RANDOM MODALITY FEATURE MISSING FORMAT. THE BEST RESULTS ARE EMPHASIZED IN BOLD.

	Acc-2 (↑)	F1 (↑)	MAE (↓)
- Meta Learning	63.55/63.67	63.43/63.27	1.279/1.265
- Learnable α	65.35/66.79	65.14/66.84	1.228/1.209
+ fus Con.	64.40/65.62	64.10/65.53	1.232/1.219
+ fus Add.	64.53/66.00	64.05/65.82	1.229/1.218
+ fus Mul.	65.08/66.44	64.96/66.19	1.239/1.228
+ fus Cma.	64.65/65.75	64.14/65.51	1.242/1.227
Meta-NA	**65.62/67.14**	**65.47/67.06**	**1.213/1.193**

TABLE VII
ABLATION STUDY RESULTS ON SIMS V2 DATASET. THE PROVIDED RESULT IS REPORTED IN STRUCTURAL / RANDOM MODALITY FEATURE MISSING FORMAT. THE BEST RESULTS ARE EMPHASIZED IN BOLD.

	Acc-2 (↑)	F1 (↑)	MAE (↓)
- Meta Learning	73.01/73.44	72.41/72.72	0.385/0.379
- Learnable α	74.75/74.91	73.82/74.00	0.372/0.366
+ fus Con.	74.21/74.72	73.63/74.35	0.365/0.361
+ fus Add.	74.99/74.98	74.25/74.36	0.362/0.358
+ fus Mul.	75.12/75.22	74.85/74.87	0.357/0.352
+ fus Cma.	74.30/74.85	73.68/74.34	0.363/0.357
Meta-NA	**75.35/75.31**	**74.86/74.95**	**0.355/0.347**

the significance of the usage of meta learning strategy. Furthermore, we ablate the learnable inner learning rate, denoted as **- Learnable** α. This results in an average performance decrease of 0.56% in the Acc-2 metric.

2) Analysis on Fusion Strategy: We also provide a comprehensively comparison between the selected tensor based fusion and the direct concatenation (denoted as **+ fus Con**), addition (denoted as **+ fus Add**), multiplication based fusion strategy (denoted as **+ fus Mul**), and multimodal attention based fusion utilized in literature [54] (denoted as **+ fus Cma**) respectively. The selected tensor fusion strategy performs the best, while the concatenation show the worst performance. This is reasonable because the tensor-based fusion method can effectively extract interactive information among different modality representations when compared with + fus Con, + fus Add, and + fus Mul. Additionally, the lower performances observed for + fus Cma substantiate the statement that employing a more potent fusion strategy does not invariably result in enhanced performance for robust MSA tasks.

F. Crucial Hyper-parameter Selection

The meta objective in Equation 24 consists of the disparity loss for each modality as well as the sentiment regression loss on noisy testing instances. It is acknowledged that the weights of the disparity loss $\eta_m, m \in \{l, a, v\}$ act as a crucial hyper-parameter for the model performance. As a result, a grid search on η_m under the random modality feature missing scenario is performed on MOSI dataset. The results are illustrated in Figure 6. In general, modifying the weight of disparity

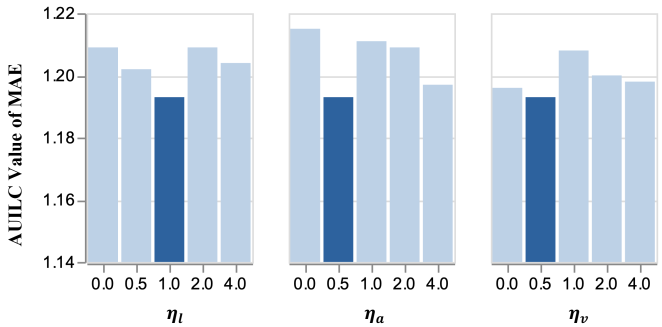
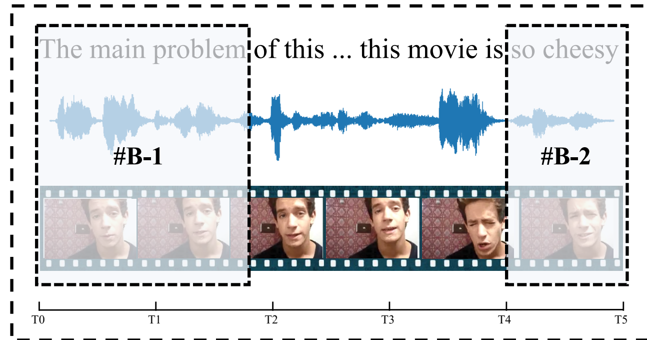


Fig. 6. Crucial Hyper-parameter selection for random feature missing on the MOSI dataset. The best model performances are marked in dark mode.

loss for each modality can result in an absolute performance fluctuations within the range of 0.02 in the MAE criteria, and the best hyperparameter combination is $(1.0, 0.5, 2.0)$ for (η_l, η_a, η_v) correspondingly. Besides, the model performance is more sensitive to the changes in the η_a compared to the changes in η_l and η_v. Such phenomenon can be results from the potential larger instances disparity in acoustic modality.

G. Case Study

In this subsection, we present a case study to intuitively demonstrate the effectiveness of the noise adaptation module in structural modality features missing scenario. As depicted in Figure 7, we manually construct noisy testing instances by removing modality sequences from T0 to T2, referred to as #B-1 and from T4 to T5, referred to as #B-2. We compare the Mean Absolute Error (MAE) criteria and L1 distance between noisy and clean unimodal representations, with and without the learned noise adaptation module. The results show that, in both cases #B-1 and #B-2, the learned noise adaptation module can effectively reduce the disparity between the noisy and manually restored clean unimodal representations, and improve sentiment prediction performance for the noisy testing instances. Furthermore, it is worth noting that, in case #B-2, where the disparity between the noisy and clean instance representations is smaller compared to case #B-1, the model achieves better prediction performance.



	# B-1 Missing		# B-2 Missing	
	- Adapt	+ Adapt	- Adapt	+ Adapt
Feat L Dis.	1.074	1.072	0.641	0.564
Feat A Dis.	0.037	0.026	0.023	0.011
Feat V Dis.	0.036	0.007	0.036	0.008
MAE (↓)	1.136	**1.066**	0.653	**0.356**

Fig. 7. Case Studies for the efficient of the feature adaption module. **Feat** m **Dis.** refers to the L1 distances between the noisy unimodal representation and the clean unimodal representation. **- Adapt** denotes removing the feature adaption module, **+ Adapt** denotes using the feature adaption module.

VII. Discussion and Conclusion

In this study, we emphasize that feature noise can exist in both training and testing instances. To deal with above challenges, we propose the Meta Noise Adaptation (Meta-NA) approach from a novel meta learning perspective. As a compromise between training from scratch and utilizing the unified model for an unseen noise pattern, the meta learning paradigm first acquires shared knowledge for all potential types of feature noise during the meta training period, and further refines using instances with the target noise pattern during the meta testing period. Experiments are classified into two groups. The first group shows the performance comparison for fine-grained feature noise. Under the assumption that the feature noise in each modality shares the same degree, quantitative results demonstrate that the proposed Meta-NA approach outperforms all existing methods under both structural and random modality feature missing. Moreover, the superiority of the Meta-NA is further reflected in general unbalanced scenarios as well as the few-shots scenarios. The second group evaluates the performances for the entire modality missing. The proposed method achieves the best results on audio and visual modality missing scenarios and achieves competitive performance on the text modality missing scenario.

It is worth noting that the proposed Meta-NA can be seamlessly extended to other multimodal tasks and considered as a general approach to enhancing the robustness of multimedia applications. For future researches, we aspire to authenticate the proposed meta learning methodology on other multimedia applications and establish a paradigm for enhancing the robustness of multimodal models against feature noise.

Acknowledgement

This work is founded by National Natural Science Foundation of China (Grant No. 62173195), National Science and Technology Major Project towards the new generation of broadband wireless mobile communication networks of Jiangxi Province (03 and 5G Major Project of Jiangxi Province, Grant No. 20232ABC03A02), High-level Scientific and Technological Innovation Talents "Double Hundred Plan" of Nanchang City in 2022 (Grant No. Hongke Zi (2022) 321-16), and Natural Science Foundation of Hebei Province, China (Grant No. F2022208006).

References

[1] 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.
[2] 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.

[3] R. Kaur and S. Kautish, "Multimodal sentiment analysis: A survey and comparison," *Research Anthology on Implementing Sentiment Analysis Across Multiple Disciplines*, pp. 1846–1870, 2022.

[4] S. H.-W. Chuah and J. Yu, "The future of service: The power of emotion in human-robot interaction," *Journal of Retailing and Consumer Services*, vol. 61, p. 102551, 2021.

[5] J. A. Rincon, A. Costa, P. Novais, V. Julian, and C. Carrascosa, "A new emotional robot assistant that facilitates human interaction and persuasion," *Knowledge and Information Systems*, vol. 60, pp. 363–383, 2019.

[6] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, "Emotional chatting machine: Emotional conversation generation with internal and external memory," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[7] 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.

[8] 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."

[9] X. Guo, A. Kot, and A. W.-K. Kong, "Pace-adaptive and noise-resistant contrastive learning for multimodal feature fusion," *IEEE Transactions on Multimedia*, 2023.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] Z. Yuan, Y. Liu, H. Xu, and K. Gao, "Noise imitation based adversarial training for robust multimodal sentiment analysis," *IEEE Transactions on Multimedia*, 2023.

[15] 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.

[16] 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.

[17] 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

[18] 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.

[19] 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," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 10 498–10 511.

[20] 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.

[21] H. Pham, T. Manzini, P. P. Liang, and B. Poczós, "Seq2seq2sentiment: Multimodal sequence to sequence models for sentiment analysis," in *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, 2018, pp. 53–63.

[22] 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.

[23] 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.

[24] L. Sun, Z. Lian, B. Liu, and J. Tao, "Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis," *IEEE Transactions on Affective Computing*, 2023.

[25] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal intelligence: Representation learning, information fusion, and applications," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 478–493, 2020.

[26] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," *Information Fusion*, 2022.

[27] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information fusion*, vol. 37, pp. 98–125, 2017.

[28] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 2017, pp. 873–883.

[29] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proceedings of the 13th international conference on multimodal interfaces*, 2011, pp. 169–176.

[30] 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*, 2022.

[31] D. Wang, X. Guo, Y. Tian, J. Liu, L. He, and X. Luo, "Tetfn: A text enhanced transformer fusion network for multimodal sentiment analysis," *Pattern Recognition*, vol. 136, p. 109259, 2023.

[32] W. Han, H. Chen, and S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[33] 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.

[34] 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.

[35] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[36] 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.

[37] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. B. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2247–2256.

[38] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1103–1114.

[39] 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.

[40] M. Huisman, J. N. Van Rijn, and A. Plaat, "A survey of deep meta-learning," *Artificial Intelligence Review*, vol. 54, no. 6, pp. 4483–4541, 2021.

[41] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 9, pp. 5149–5169, 2021.

[42] R. Vilalta and Y. Drissi, "A perspective view and survey of meta-learning," *Artificial intelligence review*, vol. 18, pp. 77–95, 2002.

[43] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning from noisy labels with deep neural networks: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[44] Z. Wang, G. Hu, and Q. Hu, "Training noise-robust deep neural networks via meta-learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4524–4533.
[45] J. Shu, Q. Xie, L. Yi, Q. Zhao, S. Zhou, Z. Xu, and D. Meng, "Meta-weight-net: Learning an explicit mapping for sample weighting," *Advances in neural information processing systems*, vol. 32, 2019.
[46] Y. Sun, S. Mai, and H. Hu, "Learning to learn better unimodal representations via adaptive multimodal meta-learning," *IEEE Transactions on Affective Computing*, 2022.
[47] 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.
[48] 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.
[49] 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.
[50] Y. Liu, Z. Yuan, H. Mao, Z. Liang, W. Yang, Y. Qiu, T. Cheng, X. Li, H. Xu, and K. Gao, "Make acoustic and visual cues matter: Ch-sims v2. 0 dataset and av-mixup consistent module," in *Proceedings of the 2022 International Conference on Multimodal Interaction*, 2022, pp. 247–258.
[51] W. Yu, H. Xu, F. Meng, Y. Zhu, Y. Ma, J. Wu, J. Zou, and K. Yang, "Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 3718–3727.
[52] 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.
[53] 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.
[54] 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
[55] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.
[56] Z. Li, F. Zhou, F. Chen, and H. Li, "Meta-sgd: Learning to learn quickly for few-shot learning," *arXiv preprint arXiv:1707.09835*, 2017.

APPENDIX A
OTHER RELEVANT EXPERIMENTAL SETUPS

A. Hyper-parameters Selection

The hyperparameters selection is provided in the Table VIII.

B. Dataset Introduction and Statistics

MOSI. The MOSI dataset [49] is widely recognized as one of the most popular datasets for Multimodal Sentiment Analysis (MSA). It comprises 2199 monologue video clips from 93 YouTube movie review videos. The annotations in the MOSI range from -3 (strongly negative) to 3 (strongly positive).

SIMS v2. The SIMS v2 dataset [50] is a popular Chinese MSA benchmark dataset. It has doubled the size of the original CH-SIMS dataset, making it more comprehensive and diverse. Human annotators label each sample with a sentiment score from -1 (strongly negative) to 1 (strongly positive).

Detailed characteristics of the MOSI and the SIMS v2 dataset is shown in Table IX.

TABLE VIII
CRITICAL HYPER-PARAMETER SETTINGS IN THE EXPERIMENT.

Hyper-parameters	MOSI	SIMS v2
Train Ins. $(n_{\min}, n_{\max})$	(64, 512)	(64, 1024)
Test Ins. m	128	256
Noise Str. p	0.5	0.5
Noise Deg. $(r_{\min}, r_{\max})$	(0%, 90%)	(0%, 90%)
Text weight η_t	1.0	1.0
Audio weight η_a	0.5	0.5
Visual weight η_v	0.5	2.0
Total tasks T	400	400
Inner Epochs E	2	2
Inner learning rate α	1e-3	1e-3
Outer learning rate β	5e-4	3e-4

TABLE IX
DETAILED CHARACTERISTICS OF THE USED DATASETS.

	MOSI	SIMS v2
Language	English	Chinese
Train Ins.	1284	2722
Valid Ins.	229	647
Test Ins.	686	1034
Feature Dims.	(768, 5, 20)	(768, 25, 177)
Sequence Lens.	(50, 375, 500)	(50, 925, 232)

C. Detailed Introduction of the Baseline Methods

MulT. The Multimodal Transformer [53] extends transformer architecture fusion the source modality into the target modality using directional pairwise cross-attention mechanism.

MISA. The Modality-Invariant and -Specific Representations [54] 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.

Self-MM. The Self-supervised Multi-task Multimodal sentiment analysis network [39] first generates the pseudo unimodal sentiment labels and then adopts them to train the model in a multi-task learning manner.

TETFN. The Text Enhanced Transformer Fusion Network [31] learns text-oriented pairwise cross-modal mappings and generates labels for each modality to learn consistency and differentiated information.

CENet. The Cross-modal Enhancement Network (CENet) [30] enriches text representations by integrating visual and acoustic information into the pretrained language model, thereby improving its performance.

MMIN. The Missing Modality Imagination Network (MMIN) [18] learns robust joint multimodal representations by the Cascade Residual Auto-encoder and Cycle Consistency Learning. By leveraging the available modality(s), the network can predict the representation of the missing modality(s) effectively.

CTFN. The Coupled-Translation Fusion Network [23] enhances bi-directional cross-modality inter-correlation through

TABLE X
PERFORMANCE COMPARISON OF UNBALANCED MODALITY FEATURE MISSING ON MOSI AND SIMS V2 DATASETS. FOR EACH NOISE STRUCTURAL, THE MISSING RATE $\in [0.4, 0.6]$ ARE PROVIDED. THE MIXED FEATURE MISSING (50%) IS A COMPROMISE OF PURE STRUCTURAL AND RANDOM FEATURE MISSING, WHERE EACH INSTANCE MAY CONTAIN EITHER RANDOM OR STRUCTURAL TYPE OF NOISE WITH PROBABILITY 50% EACH. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Dataset	Model	Structural Feature Missing			Random Feature Missing			Mixed Feature Missing (50%)		
		Acc-2 (↑)	F1 (↑)	MAE (↓)	Acc-2 (↑)	F1 (↑)	MAE (↓)	Acc-2 (↑)	F1 (↑)	MAE (↓)
MOSI	Self-MM [39]	64.43	64.60	1.276	64.58	64.77	1.279	64.58	64.75	1.281
	TPFN [12]	63.31	63.48	1.298	61.89	61.99	1.326	59.50	59.18	1.358
	MAML [55]	64.68	64.75	1.266	65.09	64.70	1.260	64.84	64.64	1.243
	Meta-NA	**65.60**	**65.74**	**1.229**	**66.21**	**66.27**	**1.219**	**65.80**	**65.90**	**1.222**
SIMS v2	Self-MM [39]	74.43	74.51	0.371	73.95	74.07	0.368	74.79	74.86	0.372
	TPFN [12]	72.66	72.65	0.388	74.24	74.32	0.384	72.18	72.30	0.390
	MAML [55]	74.99	75.02	0.365	74.28	74.24	0.356	74.85	74.88	0.358
	Meta-NA	**75.98**	**75.99**	**0.345**	**76.34**	**76.34**	**0.337**	**74.89**	**75.02**	**0.346**

couple learning, and establishes a hierarchical architecture to exploit multiple bi-directional translations.

T2FN. The Temporal Tensor Fusion Network (T2FN) [11] is a regularization technique that relies on tensor rank minimization to handle imperfect data.

TPFN. The Time Product Fusion Network (TPFN) [12] is an enhanced version of T2FN that incorporates high-order statistics from both modalities and temporal dynamics to address the challenges posed by imperfect data.

TFR-Net. The Transformer-based Feature Reconstruction Network [10] enhances model robustness by leveraging a proposed reconstruction framework to recover missing semantics.

MAML. The Model-Agnostic Meta-Learning (MAML) [55] introduces a meta-learning approach that focuses on learning adaptable model parameters through gradient descent, regardless of the specific model architecture.

Meta-SGD. The Meta-SGD [56] is a powerful meta-learner that shares similarities with stochastic gradient descent (SGD) and exhibits ease-of-training. It has the capability to initialize and adapt any differentiable learner in a single step, demonstrating its effectiveness across both supervised learning and reinforcement learning tasks.

APPENDIX B
SUPPLEMENTARY EXPERIMENTS

A. Supplementary Results for Unbalanced Setting

In addition to the low-level noise interval experiments under unbalanced setup discussed in the main content of the manuscript, we have also conducted experiments to investigate the impact of a high-level noise degree interval. Specifically, within each instance, the degree of feature absence ranges from 40% to 60% in all modality sequences. The experimental results, as summarized in Table X, further emphasize the superior performance of models trained using meta-learning techniques. Notably, the proposed Meta-NA framework achieved an average improvement of 1.5% and 1.3% on the Acc-2 criterion for the MOSI dataset and SIMS v2 dataset, respectively, across the three noise structures.