Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.084057

ARTICLE

Data Mining and Uncertainty-Aware with Missing Modalities for Multimodal Sentiment Analysis

Ying Cao[1], **Penghui Zhao**[1], **Xinyu Qiao**[1], **Ningfan Zhan**[1] **and Xiaomei Zou**[2,*]

[1]School of Computer and Information Engineering, Henan University, Kaifeng, China
[2]Research Center for High Efficiency Computing Infrastructure, Zhejiang Lab, Hangzhou, China

*Corresponding Author: Xiaomei Zou. Email: zouxiaomeihy@163.com

Received: 15 April 2026; Accepted: 15 May 2026; Published: 23 July 2026

ABSTRACT: Multimodal Sentiment Analysis (MSA) integrates diverse modalities to identify emotional states, yet performance often suffers in scenarios with missing data. In this situation, despite the promising results of recent methods, the failure of part methods to fully exploit the latent valid information contained in incomplete modalities may degrade predictive performance. Besides, to address the oversight of varying contributions across modalities to sentiment understanding, the score-based weighting schemes in the exhibited methods remain overly sensitive to data fluctuations, leading to unstable and unreliable predictions. To this end, we propose a novel method, Data Mining and Uncertainty-Aware with Missing Modalities for Multimodal Sentiment Analysis (DUDF-MSA). In this method, the Common Feature Extraction (CFE) mechanism is introduced to learn common features across modalities, thereby guiding attention toward critical cues in the missing modalities. In parallel, the specific feature uncertainty-aware dynamic adjustment (SFUA) scheme is designed to, in addition to extracting modality-specific features, adaptively assess each modality's contribution by quantifying its class uncertainty based on the probability distribution of its features. According to these weights, our method can effectively mitigate the negative impact of ambiguous sentiment cues from unreliable modalities during the following feature aggregation stage. Finally, the common features and the adjusted modality-specific features are jointly learned to predict sentiment intensity. The experiments conducted on benchmark datasets indicate the superior performance of DUDF-MSA, which yielding 34.29% Acc-7 (1.062 MAE) on MOSI and 35.66% Acc-5 (0.505 MAE) on SIMS.

KEYWORDS: Multimodal sentiment analysis; missing modalities; common features extraction; specific feature uncertainty-aware dynamic adjustment

1 Introduction

Multimodal Sentiment Analysis (MSA) aims to understand and integrate sentiment cues conveyed through multiple modalities. It has attracted considerable attention in recent studies [1–3]. However, in real-world applications, missing modalities often occur due to factors such as device malfunction, data corruption, or privacy restrictions, which can substantially degrade the performance of MSA models [4,5]. Current methods for missing-modality multimodal sentiment analysis are mainly divided into two categories: reconstruction-based recovery methods and robustness-enhanced methods. On the one hand, reconstruction-based approaches adopt feature completion techniques to restore missing modal information. For instance, TFRNet [6] leverages the Transformer architecture to repair incomplete multimodal sequences, while MPLMM [7] integrates accelerated learning strategies and prompt learning mechanisms to generate missing features for sentiment analysis and emotion recognition tasks. However, such explicit

reconstruction inevitably introduces extra noise and considerable computational overhead, which greatly limits the practical deployment of relevant models. On the other hand, robustness-oriented methods avoid explicit reconstruction. For example, LNLN [8] takes text as the dominant modality and uses adversarial learning to enhance feature robustness. TF-Mamba [9] further proposes a text-enhanced Mamba framework to align complete and missing modalities with linear complexity. Overall, model enhancement paradigms reduce computation and reconstruction errors compared with recovery-based approaches. However, such robustness-enhancement methods still suffer from inherent limitations: they rely on a fixed dominant modality to resist noise and fail to adjust the contribution weights of individual modalities adaptively.

Despite their promising results, parts of existing methods fail to fully exploit the latent valid information in incomplete modalities due to the "lazy exploration" inherent in these models. This insufficient attention to critical sentiment cues can cause the model to favor incorrect categories, ultimately degrading predictive performance. Moreover, parts of existing methods may overlook the varying contributions across modalities to sentiment understanding. As illustrated on the left side of Fig. 1, the misleading information in certain modalities may confuse the sentiment cues from other modalities, leading to reduced accuracy and robustness in sentiment recognition. For modality contribution weighting, several approaches have been proposed, including sentiment score–based methods [10] and prediction outcome–based methods [11]. However, score-based models typically demand considerable computational resources for training and are highly sensitive to modality-specific data variations. As a result, they often struggle to generalize across different datasets and tend to exhibit unstable or unreliable performance, as illustrated on the right side of Fig. 1.

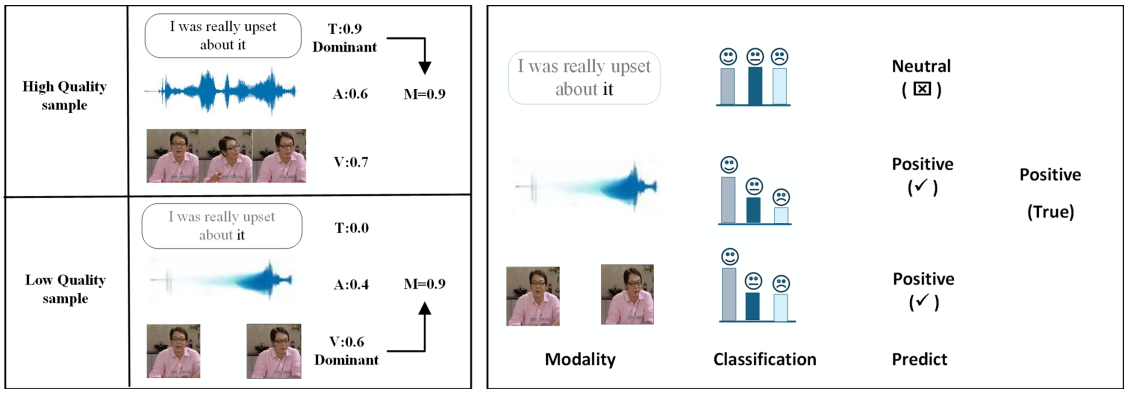


Figure 1: Problem description. The left side explains how modality dominance changes under low-quality sample conditions. The right side explains how missing data leads to unreliable model predictions.

Therefore, different from existing methods [7–11] and to thoroughly explore the information from the remaining modalities and balance the contribution of each modality during sentiment analysis under missing modality conditions, we propose the DUDF-MSA (Data Mining and Uncertainty-Aware with Missing Modalities for Multimodal Sentiment Analysis) method. Specifically, in DUDF-MSA, we first process the incomplete multimodal features X_t^s, X_a^s, and X_v^s using modality-specific encoders. We adopt Bidirectional Encoder Representations from Transformers (BERT) to encode the text modality and Transformer encoders for audio and video modalities, yielding the corresponding feature representations E_t, E_a, and E_v. To alleviate the "lazy" behavior exhibited in some recent models and enhance data mining from missing modalities, we propose the Common Feature Extraction (CFE) module, which learns common cross-modal features F_{sh} to focus on critical sentiment cues. Furthermore, we employ sentiment intensity-guided contrastive learning to capture fine-grained sentiment patterns.

Meanwhile, different with methods [8–11], which overlook the varying contributions across modality to sentiment understanding, or score-based methods, that are sensitive to modality-specific data variations, we design the Specific Feature Uncertainty-Aware Dynamic Adjustment (SFUA) scheme. It extracts modality-specific features and estimates each modality's contribution by computing class uncertainty from feature probability distributions. That is, a lower energy score implies clearer sentiment orientation and higher adaptive weight, which reduces the disturbance caused by ambiguous cues from unreliable modalities during feature aggregation.

Finally, the common features F_{sh} from the CFE module and the adjusted modality-specific features from the SFUA scheme are jointly utilized to generate the hierarchical sentiment predictions at common, specific, and final levels. The auxiliary predictions at both the common and specific levels help optimize the performance of the corresponding modules. Experiments on three public datasets validate the superior performance of our method in MSA with missing data settings. The contributions of this work are summarized as follows:

1. We introduce the Common Feature Extraction (CFE) mechanism to mitigate the model's "lazy" behavior by thoroughly exploring the potential sentiment information within each modality with missing data.
2. We propose a Specific Feature Uncertainty-Aware Dynamic Adjustment (SFUA) scheme that dynamically adjusts the weights of modality-specific features by quantifying each modality's class uncertainty from its probability distribution.
3. Experimental results on public English and Chinese datasets demonstrate that our DUDF-MSA achieves better classification accuracy and stronger robustness under various modality missing rates.

2 Related Work

2.1 Multimodal Sentiment Analysis

Multimodal Sentiment Analysis (MSA) combines multiple modalities of information to enhance sentiment understanding. Current research on MSA [12–15] focuses on developing advanced fusion strategies and interaction mechanisms for better sentiment prediction accuracy. For instance, Han et al. [16] enhanced multimodal fusion by maximizing mutual information between unimodal pairs in a hierarchical manner. Zhang et al. [17] guided representation learning using sentiment-intensive cues. Despite progress, most studies assume full availability of all modalities, a condition that often fails due to missing data.

Recent studies [6–8,18] address the issue of missing modalities by learning joint multimodal representations or reconstructing incomplete information. For example, Li et al. [18] decompose modalities into sentiment-related and modality-specific components, reconstructing sentiment semantics. Zhang et al. [8] proposed a Transformer-based, language-dominant, noise-robust network to prioritize text features. However, these methods overlook the fact that the contributions of different modalities may vary under missing modality conditions and may introduce additional overhead due to reconstruction requirements. Therefore, we adopt a decision-level late fusion strategy. To fully exploit the sentiment information across modalities, we introduce the Common Feature Extraction (CFE) module to guide the model's attention toward crucial sentiment cues in incomplete modalities by learning cross-modal common features. In addition, we design a Specific Feature Uncertainty-aware Dynamic Adjustment (SFUA) scheme to evaluate the contributions of different modalities and to alleviate the negative interference caused by confusing information from unreliable modalities.

2.2 Uncertainty Estimation and Multimodal Fusion

Multimodal fusion is a core theme in multimodal learning, focused on integrating modality features into joint representations for downstream tasks. It can be classified into early, intermediate, and late fusion. While neuroscience and machine learning studies suggest intermediate fusion may enhance representation learning [19,20], late fusion remains the most widely used approach due to its simplicity in interpretation and practical implementation.

However, the reliability of current fusion methods remains significantly unexplored, limiting their application in safety-critical fields such as financial risk assessment and medical diagnosis. The motivation behind uncertainty estimation is to indicate whether the predictions made by a machine learning model are prone to error. Numerous uncertainty estimation methods have been proposed over the past decades, including Bayesian Neural Networks (BNNs) and their variants, deep ensembles, predictive confidence, Dempster-Shafer Theory, and energy scores [21]. Predictive confidence typically refers to the consistency between the predicted class probabilities and empirical accuracy, often mentioned in classification tasks. The Dempster-Shafer Theory (DST) is an extension of Bayesian theory to subjective probabilities, providing a general framework for modeling cognitive uncertainty. Energy scores are a promising method for capturing out-of-distribution (OOD) uncertainty, which arises when a machine learning model encounters inputs that differ from its training data, leading to unreliable predictions. Recent studies have extensively addressed the OOD uncertainty problem [22]. Uncertainty-based multimodal fusion methods have demonstrated significant advantages across various tasks, including clustering, classification [23,24], regression [25], object detection, and semantic segmentation. In this paper, we explore energy scores due to their theoretical interpretability and effectiveness.

3 Method

In this paper, different from existing methods [7–11] and to thoroughly explore the information from the remaining modalities and balance the contribution of each modality during sentiment analysis under missing modality conditions, we propose the DUDF-MSA (Data Mining and Uncertainty-Aware with Missing Modalities for Multimodal Sentiment Analysis) method.

3.1 Framework Description

As illustrated in Fig. 2, in our framework, we first generate multimodal features with randomly missing data, X_t^s, X_a^s, and X_v^s. After input preparation, different encoding methods are chosen for each modality: BERT for text; Transformer Encoder for video and audio, thus obtaining feature representations E_t, E_a, and E_v. Based on these representations, to sufficiently exploit useful sentiment information from incomplete modalities, we introduce a Common Feature Extraction (CFE) mechanism; in this mechanism, features extracted across modalities T_c, A_c, and V_c are aggregated based on the attention mechanism, thus forming the common features F_{sh} which guide the model's attention toward critical cues in missing modalities. Besides, contrastive learning guided by sentiment intensity is incorporated during training to capture more fine-grained sentiment cues.

Meanwhile, we design the specific feature uncertainty-aware dynamic adjustment (SFUA) scheme. In this scheme, besides extracting modality-specific features including E_t^s, E_a^s, and E_v^s from text, audio, and visual modalities, respectively, it accurately evaluates the distinct contribution of each modality by calculating the class uncertainty of each modality based on the probability distribution of its features. A lower modality uncertainty, as represented by the energy score $Energy_m$, indicates a clearer sentiment orientation, and consequently, a higher adjustment weight for this modality. Afterwards, based on these adaptively weighted, our method can further effectively mitigate the negative impact of ambiguous sentiment cues from unreliable

modalities during the following features aggregation stage. Finally, we combine the common features F_{sh} and the adjusted modality-specific features to generate the hierarchical sentiment predictions at common, specific, and final levels to obtain the analyzed multimodal sentiment result.

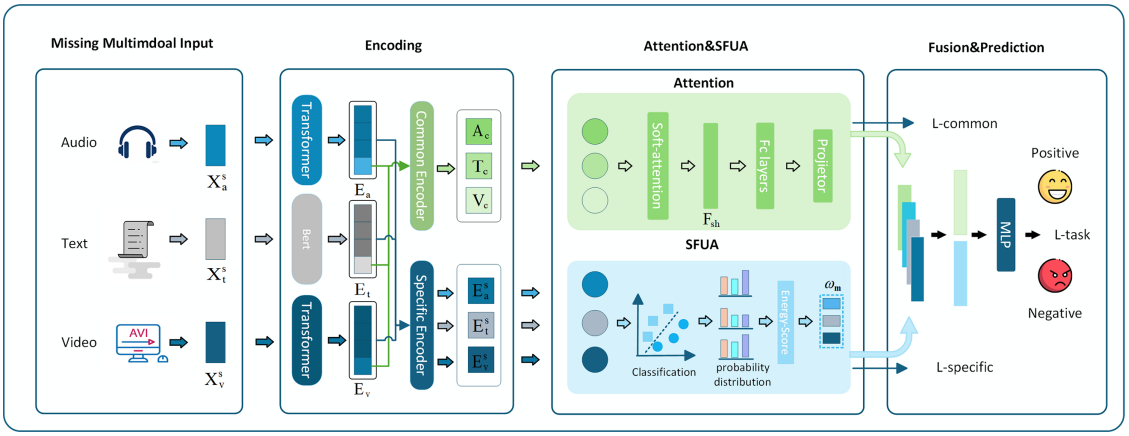


Figure 2: Overview of the proposed framework (X_t^s, X_a^s, X_v^s denote the input features from text, audio and visual modalities, respectively, where the gray regions indicate randomly missing parts. F_{sh} represents the common features, while E_t^s, E_a^s, E_v^s are the modality-specific features. L-common, L-specific, and L-task correspond to the losses associated with the common-level, specific-level and final-level predictions, respectively).

3.2 Encoding of Missing Multimodal Data

Building upon previous work, we use BERT [26], Librosa [27], and OpenFace [28] to separately encode the text, audio, and video source data. These preprocessed inputs are represented as sequences, denoted by $X_m \in \mathbb{R}^{T_m \times d_m}$, where $m \in \{t, a, v\}$ represents the modality type (t for text, a for audio, v for video), T_m is the sequence length, d_m is the dimensionality of the modality vector.

To ensure the reproducibility of the missing data generation process, we explicitly present the detailed implementation details of missing modality simulation based on previous research [8,29], including missing rate configuration, random seed strategy, and missing value replacement schemes, which are specified as follows:

3.2.1 Missing Rate Configuration

The missing rate r ranges from 0% to 100% and is configured independently for each modality. Specifically, for each modality $m \in \{t, a, v\}$, we randomly select $r \times T_m$ feature positions from the sequence X_m for masking and removal. The missing rates of different modalities are independent of one another without mutual interference. This configuration conforms to real-world scenarios, where each modality may suffer from independent data loss (e.g., audio corruption caused by device failure and text missing due to data damage).

3.2.2 Random Seed Strategy

To guarantee experimental reproducibility, the random seed is fixed to 1111 throughout the missing data generation. Meanwhile, for fair comparison, we adopt three fixed random seeds (1111, 1112, 1113) for each missing rate setting and repeat the training process three times to obtain the average result. These seeds govern the random selection of masked positions in each modal sequence, ensuring that identical missing patterns can be reproduced under consistent experimental settings.

3.2.3 Missing Value Replacement Process

After determining the masked positions for each modality, we adopt modality-specific replacement strategies, thereby forming multimodal inputs with missing data $X_m^s \in \mathbb{R}^{T_m \times d_m}$. The simplified procedure is illustrated in Algorithm 1:

Algorithm 1: Missing data simulation

```
1: Fix random seed = 1111
2: for each modality m do
3:     Randomly select r% positions to mask
4:     if m is audio or video then
5:     Set masked features to 0 (zero padding)
6:     else if m is text then
7:     Replace masked features with [UNK] embedding
8:     end if
9: end for
```

After forming the multimodal inputs with missing data, we introduce Transformer Encoders for encoding the audio and video modalities, while the text modality is encoded using BERT [26]. The encoded representations of the sample's text, video, and audio modalities are denoted as $E_t \in \mathbb{R}^{l_t \times d_t}$, $E_v \in \mathbb{R}^{l_v \times d_v}$, $E_a \in \mathbb{R}^{l_a \times d_a}$, respectively. Here, l_m represents the sequence length of each modality, and d_m represents the corresponding feature vector dimension, where $m \in \{t, v, a\}$.

3.3 Common Feature Extraction

To fully explore the inherent sentiment information in incomplete modalities, we introduce a Common Feature Extraction (CFE) mechanism to guide the model's attention to crucial sentiment cues by learning the common features across modalities. Simultaneously, during the training process, we employ contrastive learning guided by sentiment intensity to capture more fine-grained sentiment cues.

Notably, in our architecture, to avoid redundancy between common features and modality-specific features, the CFE module is designed to explicitly disentangle common cross-modal information and modality-specific information. CFE only captures modality-invariant sentiment patterns common across text, audio, and video, which are complementary to the modality-specific features in the subsequent SFUA module.

Based on this, the primary objective of this common feature extraction module can be described as to project information from different modalities into a shared representation space. Specifically, for the text modality, we select the [CLS] vector output by the BERT model, $E_t^c \in \mathbb{R}^{l \times d_t}$, as its common vector representation. For the video and audio modalities, we choose the output vectors from the last layer of the Transformer encoder, $E_v^c \in \mathbb{R}^{l \times d_v}$ and $E_a^c \in \mathbb{R}^{l \times d_a}$, as their common vector representations. Next, through fully connected (FC) layers and ReLU activation functions, these three vectors are unified and converted into the same dimension to obtain $T_c \in \mathbb{R}^{d_c \times 1}$, $V_c \in \mathbb{R}^{d_c \times 1}$ and $A_c \in \mathbb{R}^{d_c \times 1}$.

Based on these information, in CFE, the formed common features which contain the crucial sentiment cues of missing modalities can be defined as follows:

$$F_{sh} = \sum_{I \in \{A,T,V\}} W_I \cdot I_c \tag{1}$$

where F_{sh} denotes the common features, I_c is the feature representation of different modalities, and W_I is the fusion weight corresponding to each modality. Specifically, we use a fully connected layer (MLP) to compute

the fusion weights for each modality:

$$W_I = \text{softmax}(\text{MLP}(H_c)),\ I \in \{A, T, V\}, H_c = \{A_c, T_c, V_c\} \tag{2}$$

In this context, H_c is a set of features from three modalities. $softmax(\cdot)$ denotes the soft attention scheme. And MLP consists of two linear transformations and a ReLU activation function. In addition, to capture fine-grained sentiment information, we introduced contrastive learning guided by sentiment intensity during the training process.

As shown in Fig. 3, given samples i, j, and k, where the sentiment intensity differences from i to j and from i to k are 0.5 and 1.1, respectively, both (i, j) and (i, k) are initial negative sample pairs for i. However, the sentiment intensity difference between i and k is clearly much larger. Therefore, when calculating the contrastive loss, we assign a higher weight to the pair (i, k) to further separate i and k in the representation space compared to i and j. To achieve this, we design a weight function using the nonlinear function $|\tanh(x)|$, as shown below:

$$\omega_{(i,j)} = \begin{cases} |\tanh(D_{(i,j)} - 2\kappa)| \times 1.5, (i, j) \in \text{initial negative pairs} \\ |\tanh(D_{(i,j)})| \times 1.5, (i, j) \in \text{initial positive pairs} \end{cases} \tag{3}$$

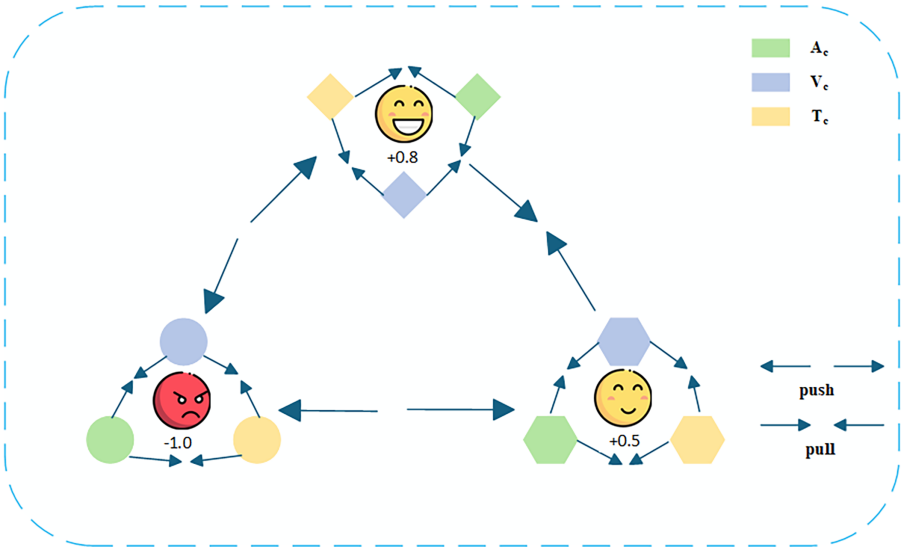


Figure 3: Contrastive learning. The numbers in the figure represent different levels of sentiment intensity.

In this process, $D_{(i,j)}$ is the selection of initial positive and negative sample pairs. κ is the sentiment intensity difference threshold, to determine whether a sample j is classified as an initial positive or negative sample of i. This process is explained in the following equation:

$$\begin{cases} D_{(i,j)} > \kappa, & (i, j) \in \text{initial negative pairs} \\ D_{(i,j)} \leq \kappa, & (i, j) \in \text{initial positive pairs} \end{cases} \tag{4}$$

and $D_{(i,j)}$ can be further described as:

$$D_{(i,j)} = |y'_i - y'_j|, j \in B \& j \neq i \tag{5}$$

In this form, y'_i and y'_j represent the sentiment intensity labels of samples i and j, respectively. B is a batch. Since the sentiment intensity ranges vary ([−3, 3] in MOSI/MOSEI and [−1, 1] in SIMS). To address the inconsistent label distribution across datasets, we adopt a strict linear normalization strategy to unify all sentiment intensity values into the interval of [−1, 1]. For MOSI and MOSEI with an original range of [−3, 3], the mapping is formulated as:

$$y' = \frac{y}{3} \tag{6}$$

This linear scaling only compresses the value scale while completely preserving the relative order and pairwise distance of sample labels. Since the transformation is bijective and reversible, no effective fine-grained sentiment information is lost during normalization. Furthermore, the normalized labels are only utilized for weighted contrastive learning, which relies on relative sentiment differences between samples; thus, the linear mapping will not introduce additional bias. The original raw labels are still adopted for the final regression prediction task. Based on the above settings, we use a unified mapping during the contrastive learning process to standardize dataset differences in label distribution. Finally, the contrastive loss we obtain is:

$$L_{cl} = -\mathbb{E}_{i\in B}\frac{\sum_{(a,p)\in P^i}\delta(a,p)}{\sum_{(a,q)\in P^i\cup N^i}\delta(a,q)} \tag{7}$$

In this equation, $\delta(a,p) = e^{\left[\omega_{(i,j)}*\frac{\mathrm{sim}(a,p)}{\tau}\right]}$, $\omega_{(i,j)}$ is the weight specified by Eq. (3). B is a batch. P^i and N^i are the positive and negative sample pairs for sample i, respectively. P^i and N^i are defined as follows:

$$\begin{aligned} P^i &= P^i_{inter} \cup P^i_{intra} \\ N^i &= N^i_{inter} \cup N^i_{intra} \end{aligned} \tag{8}$$

where P^i_{inter} and N^i_{inter} the positive and negative sample pairs between modalities. P^i_{intra} and N^i_{intra} are the positive and negative sample pairs within the modality. The inter-modal positive and negative sample pair selection is as follows:

$$\begin{aligned} P^i_{inter} = \Big\{ & (V^i_c, T^j_c),\ (V^i_c, A^j_c),\ (T^i_c, A^j_c) \\ & \cup (V^i_c, T^j_c),\ (T^i_c, V^j_c),\ (V^i_c, A^j_c), \\ & (A^i_c, V^j_c),\ (T^i_c, A^j_c),\ (A^i_c, T^j_c) \\ & \mid (i,j) \in \text{initial positive pairs} \Big\} \\ N^i_{inter} = \Big\{ & (V^i_c, T^k_c),\ (T^i_c, V^k_c),\ (V^i_c, A^k_c), \\ & (A^i_c, V^k_c),\ (T^i_c, A^k_c),\ (A^i_c, T^k_c) \\ & \mid (i,k) \in \text{initial negative pairs} \Big\} \end{aligned} \tag{9}$$

Besides, the intra-modal positive and negative sample pair selection can be defined as follows:

$$\begin{aligned} P^i_{intra} = \Big\{ & (T^i_c, T^j_c), (V^i_c, V^j_c), (A^i_c, A^j_c), \\ & \mid (i,j) \in \text{initial positive pairs} \Big\} \\ N^i_{intra} = \Big\{ & (T^i_c, T^k_c), (V^i_c, V^k_c), (A^i_c, A^k_c), \\ & \mid (i,k) \in \text{initial negative pairs} \Big\} \end{aligned} \tag{10}$$

In above equation, T_c^i, V_c^i, A_c^i correspond to the representations of the three different modalities of sample i, with the remaining symbols having the same meaning.

3.4 Specific Feature Uncertainty-Aware Dynamic Adjustment

In addition to the common feature representation, which strengthens the mining of latent sentiment information, we design the modality-specific branch that captures unique emotional information of every single modality and introduce the specific feature uncertainty-aware dynamic adjustment (SFUA) scheme to adaptively assess each modality's contribution by quantifying its class uncertainty based on the probability distribution of its features. Following this way, our method can further effectively mitigate the negative impact of ambiguous sentiment cues from unreliable modalities during the following feature aggregation stage.

Step 1: Specific feature extraction. Firstly, given a modality $E_m \in \mathbb{R}^{l_m \times d_m}$, $m \in \{t, v, a\}$, we use global average pooling (GAP) along the sequence length to compress E_m into $E_m^1 \in \mathbb{R}^{1 \times d_m}$. Then, two-step nonlinear transformations are applied to project E_m^1 into a new low-dimensional space:

$$E_m^s = \sigma_2(W_2\sigma_1(W_1 {E_m^1}^T)),\ m \in \{t, v, a\} \tag{11}$$

In this context, $W_1 \in \mathbb{R}^{(d_m/8)\times d_m}$, $W_2 \in \mathbb{R}^{(d_s)\times(d_m/8)}$, $E_m^s \in \mathbb{R}^{d_s \times 1}$, the σ_1 represents the ReLU function,the σ_2 represents the Sigmoid function.

Step 2: Uncertainty-aware dynamic adjustment

Unlike traditional score-based methods, which may struggle to generalize across different datasets and tend to exhibit unstable or unreliable performance, we use an energy-based uncertainty estimation to calculate the class uncertainty of each modality and output the class energy scores. And according to these scores to form the adjustment weights stably. This process can be described in Eq. (12):

$$Energy_m = -log \sum_k^K e^{f_m^{(k)} x^{(m)}}, m \in \{t, a, v\} \tag{12}$$

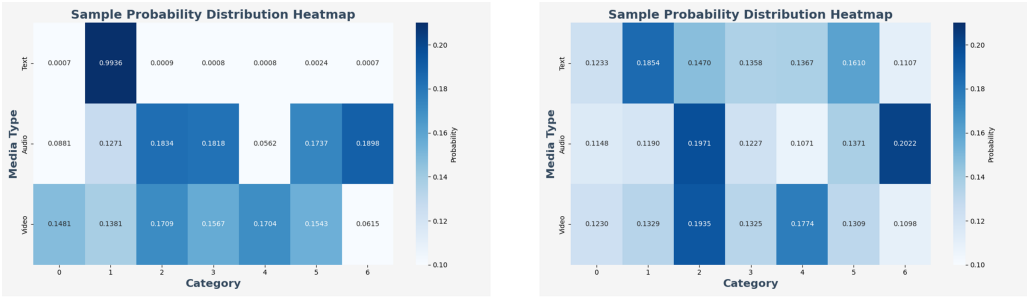
In this equation, $Energy_m$ is the energy score we calculate, $x^{(m)}$ the input single-modality feature, f_m is the classifier we designed, and $f_m^{(k)} x^m$ is the output log of the classifier f_m corresponding to the k-th class label. In this framework, we set up independent classifiers for each modality. Notably, each classifier does not require additional pre-training, which makes our model highly adaptive. We define the classification loss for each modality as follows:

$$L_{sp^m} = \sum_m \mathcal{L}_{CE}(y, f^m(x^{(m)})), m \in \{t, a, v\} \tag{13}$$

In this class, uncertainty measure for each modality, a lower energy score indicates a more concentrated probability distribution of the modality for the sample's category, implying a clearer sentiment orientation, and the less confusing information exists in this modality. Conversely, a higher energy score reflects a more dispersed distribution, suggesting that the sample lacks distinct category attributes, and may exist with more confusing sentiment category cues.

As illustrated in Fig. 4a, the predicted class probability distribution of the text modality is more concentrated, indicating that its specific features possess a clearer sentiment orientation and contain less confusing information; therefore, a higher weight should be assigned to this modality. In contrast, in Fig. 4b, the predicted probability distribution for the text modality is less concentrated than that of the audio

modality, revealing the presence of confusing information that may hinder sentiment prediction accuracy. Accordingly, a lower weight should be assigned to this modality.



(a) (b)

Figure 4: Uncertainty explanation. Based on the concentration of the category probability distribution predicted by the model. (**a**,**b**) illustrate the predicted probability distributions for different samples, respectively.

Therefore, according to the calculated energy scores, we determine the final adjustment weights:

$$\omega_m = \alpha \times Energy_m, m \in \{t, a, v\} \tag{14}$$

In this context, $\alpha < 0$ serves as a hyperparameter. Finally, the final modality-specific representations are derived according to the adjustment weights assigned to each modality:

$$F_{sp^m} = \omega_m E_m^s, m \in \{t, a, v\} \tag{15}$$

3.5 Post-Processing for Sentiment Prediction

In this part, the common and the adjusted specific features are integrated to obtain the final fused modality:

$$F^* = Concat(F_{sh}, F_{sp^t}, F_{sp^a}, F_{sp^v}) \tag{16}$$

where F_{sh} is the common features obtained by CFE; (F_{sp^t}), (F_{sp^a}), and (F_{sp^v}) are the adjusted text, audio, and video modality-specific features formed by SFUA. Then, the combined features F^* is employed to output the final sentiment results.

Moreover, to enhance the representational capability of different encoders and improve prediction accuracy, we employ a hierarchical prediction strategy comprising common-level, specific-level, and final predictions. And during the training stage, the outputs from the common-level and specific-level predictors are each supervised by the ground-truth sentiment labels. This auxiliary supervision helps guide the learning of individual modules, enabling them to achieve better representation quality and overall performance.

3.6 Overall Learning Objective

The entire network is trained end-to-end by integrating common features and modality-specific features. The two branches have different optimization objectives and non-shared parameters. Meanwhile, we introduce hierarchical prediction to provide separate supervision for the two types of features, which enhances the discriminability of different encoders and ensures no redundancy between common features

and modality-specific features. We define the overall learning objective as a combination of multiple loss components corresponding to different sub-modules. Specifically, the overall training loss is formulated as:

$$L_{overall} = L_{sh} + L_{sp^m} + L_{task} + \gamma L_{cl} \tag{17}$$

In this context, L_{sh} and L_{sp^m} are the losses associated with the common features extraction and specific feature learning, L_{sh} is demonstrated as:

$$L_{sh} = \frac{1}{N_b} \sum_{i=0}^{N_b} |y_i - \hat{y}_i| \tag{18}$$

where N_b is the number of samples in batch B, $\hat{y}_i$ is the true sentiment intensity and y_i is the predicted sentiment intensity value. Furthermore, the L_{task} and L_{cl} denote the output loss and the contrastive loss. Here, L_{task} can be described as:

$$L_{task} = \frac{1}{N_b} \sum_{i=0}^{N_b} |y_i - \hat{y}_i| \tag{19}$$

4 Experiments and Analysis

4.1 Datasets and Metrics

We evaluated our method on three widely used MSA datasets: the English datasets MOSI [30] and MOSEI [31], and the Chinese dataset SIMS [32]. More detailed information about the datasets can be found in Appendix A.

To ensure a fair comparison and evaluation, we follow the same missing modality setting as LNLN [8], where training data is randomly dropped with a certain probability. During testing, the missing rate r is varied from 0 to 0.9 in increments of 0.1, notably, r = 1.0 is excluded as it removes all modalities, making the task ill-posed. Based on the results for each missing rate, we calculated the average value from the missing rate of 0 to 0.9 as the overall performance of the model under different noise levels.

Based on the research by Zhang et al. [8], we report the 5-level (Acc-5) and 7-level (Acc-7) accuracies for MOSI and MOSEI, as well as the 3-level (Acc-3) and 5-level (Acc-5) accuracies for SIMS. Additionally, all datasets report the 2-level accuracy (Acc-2), Mean Absolute Error (MAE), Pearson Correlation (Corr), and F1 score (F1).

For Acc-2 and F1 on MOSI and MOSEI, the results are provided for two settings: Negative vs. Positive (to the left of "/") and Negative vs. Non-negative (to the right of "/"). Except for MAE, higher values indicate better performance.

Comprehensive details of the experimental setup are provided in Appendix B.

4.2 State-of-the-Art Comparison

We compare DUDF-MSA with seven typical multimodal sentiment analysis methods: MISA [33], Self-MM [29], CENET [34], TFR-Net [6], ALMT [17], GLGIS [35], and LNLN [8]. These baselines include traditional multimodal fusion approaches, modality completion and reconstruction methods, as well as recent robust models designed for missing and noisy scenarios. The proposed method combines common feature learning and uncertainty-aware dynamic weighting, which outperforms existing approaches and yields stronger robustness under various modality missing conditions. Detailed baseline settings and descriptions are provided in Appendix C.

As presented in Table 1, our method consistently outperforms existing approaches on most evaluation metrics across both the MOSI and MOSEI English datasets (note that lower MAE values are better). Specifically, on the MOSI dataset, when compared with GLGIS [35], which is our baseline, our method achieves improvements of 0.95% in multi-class accuracy (Acc-7) and 0.9% in the stability metric (Corr), while reducing MAE by 2.5%. On the MOSI dataset, when compared to the text-dominant LNLN [8], our approach further yields gains of 1.04% in Acc-7 and 1.16% in F1. On the MOSEI dataset, our method surpasses LNLN [8] by 1.2% on Acc-7, 1.81% on F1, while reducing MAE by 0.9%. These results demonstrate the robustness and effectiveness of our method in multimodal sentiment analysis, even under challenging conditions with missing data.

Table 1: Overall performance comparison on the MOSI and MOSEI datasets under missing modality settings (The best results are shown in bold. Note: The smaller MAE indicates the better performance).

Method	MOSI						MOSEI					
	Acc-7 (%)	Acc-5 (%)	Acc-2 (%)	F1 (%)	MAE	Corr	Acc-7 (%)	Acc-5 (%)	Acc-2 (%)	F1 (%)	MAE	Corr
MISA [33]	29.85	33.08	71.49/70.33	71.28/70.00	1.085	0.524	40.84	39.39	71.27/75.82	63.85/68.73	0.780	0.503
Self-MM [29]	29.55	34.67	70.51/69.26	66.60/67.54	1.070	0.512	44.70	45.38	73.89/77.42	68.92/72.31	0.695	0.498
CENET [34]	30.38	33.62	71.46/67.73	68.41/64.85	1.080	0.504	47.18	47.83	74.67/77.34	70.68/74.08	0.685	0.535
TFR-Net [6]	29.54	34.67	68.15/66.35	61.73/60.06	1.200	0.459	46.83	34.67	73.62/77.23	68.80/71.99	0.697	0.489
ALMT [17]	30.30	33.42	70.40/68.39	**72.57/71.80**	1.083	0.498	40.92	41.64	76.64/77.54	77.14/78.03	0.674	0.481
GLGIS [35]	33.34	37.40	**72.39/71.21**	72.36/71.63	1.087	0.516	47.43	48.36	**78.15**/79.53	**79.51**/80.58	0.668	**0.578**
LNLN [8]	33.25	37.10	71.14/69.92	71.12/70.01	1.064	**0.527**	46.47	47.31	77.22/78.04	77.69/78.98	0.676	0.577
DUDF-MSA	**34.29**	**38.44**	72.30/71.16	72.26/71.27	**1.062**	0.525	**47.67**	**48.41**	**78.15/79.73**	79.50/**80.96**	**0.667**	0.576

Moreover, as shown in Table 2, our method also achieves state-of-the-art performance on most metrics for the Chinese dataset SIMS. Compared to the static fusion model GLGIS [35], our method improves by 1.01% on Acc-5, 2.0% on Corr, and Moreover, compared to the text modality-dominated LNLN [8], our method improves by 1.62% on Acc-5 and 2.3% on Corr and the MAE is reduced by 1.3%. These results demonstrate that our model performs well in terms of accuracy and robustness across different types of datasets, proving the generalization capability of our model. Detailed experimental results for various missing rates are presented in Appendix D.

Table 2: Overall performance comparison on the SIMS datasets under missing modality settings (The best results are shown in bold. Note: The smaller MAE indicates the better performance).

Method	Acc-5 (%)	Acc-3 (%)	Acc-2 (%)	F1 (%)	MAE	Corr
MISA [33]	31.53	56.87	72.71	66.30	0.539	0.348
Self-MM [29]	32.28	56.75	72.81	68.43	0.508	0.376
CENET [34]	22.29	53.17	68.13	57.90	0.589	0.107
TFR-Net [6]	26.52	52.89	68.13	58.70	0.661	0.169
ALMT [17]	20.00	45.36	69.66	72.76	0.561	0.364
GLGIS [35]	34.65	**57.75**	73.06	76.01	0.519	0.387
LNLN [8]	34.04	57.32	73.19	**76.66**	0.518	0.384
DUDF-MSA	**35.66**	57.71	**73.21**	75.08	**0.505**	**0.407**

4.3 Robustness Comparison

To visually compare the robustness across different methods at various missing rates, Fig. 5 plots the performance curves of five state-of-the-art baseline methods: MISA [33], CENET [34], ALMT [17], GLGIS [35], and LNLN [8]. Fig. 5a–c show the F1 curves on MOSI, MOSEI, and SIMS, respectively, while Fig. 5d–f depict the corresponding MAE curves. It can be observed that our method consistently achieves superior performance across all datasets, maintaining higher F1 scores and lower MAE values compared with other approaches. These results further demonstrate the robustness of our method under varying missing rates (note that lower MAE values indicate better performance).

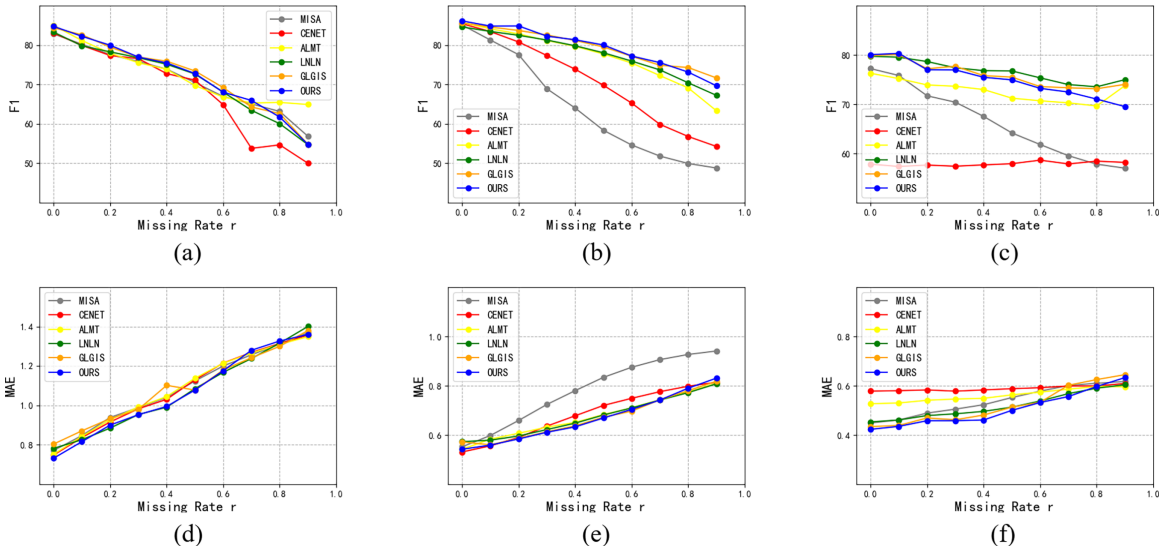


Figure 5: Performance curves at various missing rates ((**a**–**c**) show the F1 scores on MOSI, MOSEI, and SIMS, respectively, while (**d**–**f**) depict the corresponding MAE results on the same datasets).

4.4 Ablation Study

4.4.1 Effectiveness of Main Components

In our framework, the Common Features Extraction (CFE) and Specific Features Uncertainty-aware Dynamic Adjustment (SFUA) modules are the two key components responsible for constructing the common features and the adjusted specific features, respectively. To verify their effectiveness, we conducted ablation experiments on the MOSI and SIMS datasets by individually removing each module from our method. The corresponding results are reported in Table 3.

In this table, "w/o CFE" denotes the removal of the CFE module, and "w/o attention" represents the removal of the fusion weights generation scheme in CFE, as defined in Eq. (2); in this case, the features from each modality are aggregated averagely. Additionally, "w/o SFUA" denotes the removal of the specific feature uncertainty-aware dynamic adjustment (step 2) scheme in SFUA module, and "w/o spe" means removing specific features from the overall framework.

1. Effectiveness of CFE for Common Features

In the "w/o CFE" setting, removing the CFE, which is the common feature extractor, leads to performance degradation. Specifically:

Table 3: Effectiveness of combination of common and specific features (The best results are shown in bold, and the smaller MAE indicate the better performance).

Method	MOSI			SIMS		
	Acc-7 (%)	F1 (%)	MAE	Acc-5 (%)	F1 (%)	MAE
w/o CFE	33.58	71.70	1.088	34.68	73.54	0.502
w/o attention	33.53	71.43	1.086	34.72	75.03	0.508
w/o Spe	33.42	71.73	1.069	33.46	74.16	**0.500**
w/o SFUA	33.81	71.60	1.078	35.09	74.51	0.501
w/o CFE & SFUA	33.00	72.07	1.094	34.89	75.02	0.511
DUDF-MSA	**34.29**	**72.30**	**1.062**	**35.66**	**75.08**	0.505

On the MOSI dataset, the Acc-7 decreases by 0.71%, and the MAE increases by 1.6%. Moreover, on the SIMS dataset, the Acc-5 drops by 0.98%. The performance degradations demonstrate the crucial role of the CFE in sentiment accurate analysis by capturing inherent sentiment cues from missing modalities.

To further analyze the contribution of our CFE in enhancing the model's data mining ability, we visualize the confusion matrices on the MOSI dataset under different missing rates (0%, 0.5%, and 0.9%). As illustrated in Fig. 6a–c, which are the results without CFE, with the missing rate increases, the model tends to bias its predictions toward specific incorrect categories. These increasingly concentrated predicted results indicate the "lazy" behavior of the model with missing data, where the model over relies on partial modalities and ignores inherent informative cues.

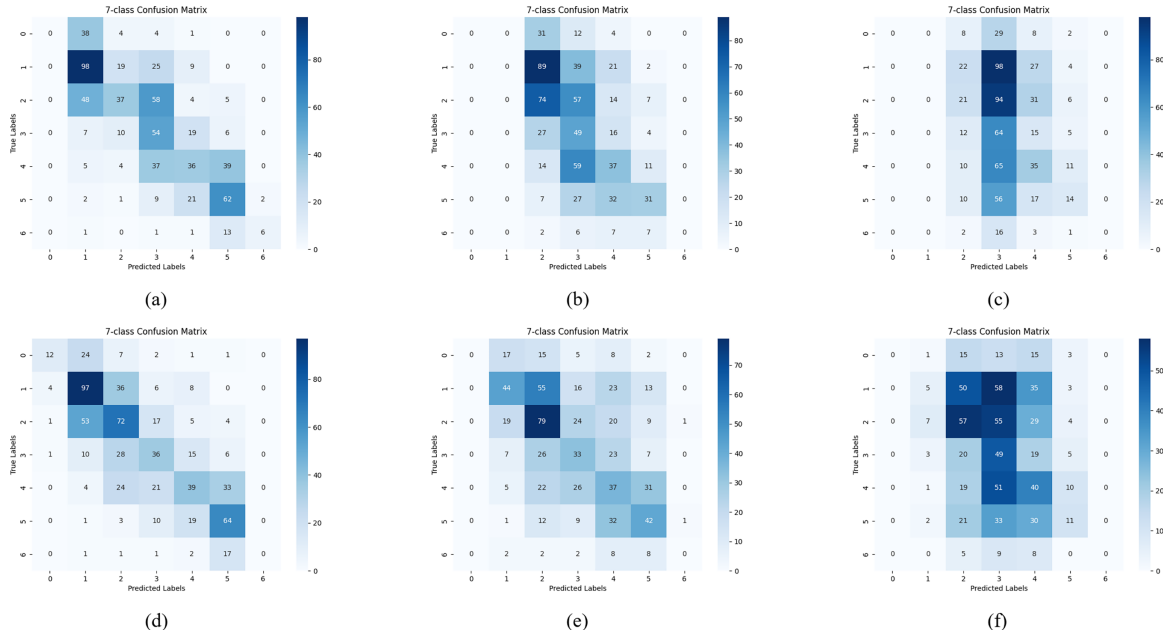


Figure 6: Confusion matrices of the MOSI dataset under different missing rates (The upper row illustrates the confusion matrices after removing the common features, while the lower row presents those of the complete model, both under varying missing rates). (**a**–**c**) illustrate the confusion matrices on the MOSI dataset without common features at missing rates of 0%, 0.5% and 0.9%, respectively. (**d**–**f**) illustrate the confusion matrices on the MOSI dataset with common features at missing rates of 0%, 0.5% and 0.9%, respectively.

Conversely, Fig. 6d–f display the confusion matrices of our entire method on the MOSI dataset for the same missing rates (0%, 0.5%, and 0.9%, respectively). Compared with the results without CFE, our model maintains more diverse and relatively balanced category distributions. Combined with the quantitative results, these findings confirm that the proposed CFE effectively improves the model's mining of valid information under conditions of missing modality.

2. Effectiveness of SFUA for Specific Feature Fusion

To examine the contribution of SFUA module, we conduct two ablation settings by removing the specific features ("w/o Spe") and the uncertainty-aware weight adjustment (Ste p2) scheme ("w/o SFUA") from our framework, respectively.

As shown in Table 3, when remove the step 2 from SFUA ("w/o SFUA"), the Acc-7 on the MOSI dataset decreases by 0.48%, and the Acc-5 on the SIMS dataset drops by 0.57%. This demonstrates the importance of effectively measuring the contributions of each modalities, as the confusing or noisy class information in certain modalities can otherwise hinder the accuracy of the final prediction.

Furthermore, when the specific features are completely removed ("w/o Spe"), the Acc-7 on the MOSI dataset further decreases by 0.39%, and the Acc-5 on the SIMS dataset drops by 1.63%. These performance degradations highlight the necessity of molding modality-specific representations to overall sentiment prediction.

3. Effectiveness of the Combination

As exhibited in Table 3, jointly learning both common and the adjusted specific features achieves the best overall performance. This result confirm the complementary relationship between the common and specific components in our framework, demonstrating the coherent integration of Common Features Extraction (CEF) and Specific Feature Uncertainty-Aware Dynamic Adjustment (SFUA) enables more accurate multimodal sentiment prediction.

4.4.2 Effectiveness of Contrastive Learning and Hierarchical Prediction

During the training stage, contrastive learning is integrated into the CFE module to capture fine-grained sentiment information. At the same time, hierarchical prediction is utilised to compute the losses for the CFE, SFUA, and final prediction, allowing for separate and coordinated optimisation to achieve better performance.

To evaluate the effectiveness of these two designs, we conduct ablation experiments on the MOSI and SIMS datasets by removing contrastive learning and hierarchical prediction, respectively. As shown in Table 4, "w/o CL" denotes the removal of the contrastive learning mechanism; "w/o L-common" represents removing the loss from the CFE module, while "w/o L-specific" denotes removing the loss from the SFUA component.

Specifically, after removing contrastive learning ("w/o CL"), the Acc-7 on the MOSI dataset decreases by 0.9%, and the Acc-5 on the SIMS dataset drops by 1.87%. This notable degradation in multi-class accuracy highlights the importance of integrating contrastive learning to capture more fine-grained sentiment representations.

Furthermore, when the common-level loss is removed ("w/o L-common"), the Acc-7 on MOSI decreases by 0.76%, and the MAE increases by 1.0%. Similarly, removing the specific-level loss ("w/o L-specific") results in a 0.93% decrease in Acc-7 and a 1.2% increase in MAE. These results verify that hierarchical prediction effectively enhances the training of both modules, promotes mutual feature refinement, and ultimately contributes to improving the overall model accuracy.

Table 4: Effects of different regularization (The best results are shown in bold. Note: The smaller MAE indicate the better performance).

Method	MOSI			SIMS		
	Acc-7 (%)	F1 (%)	MAE	Acc-5 (%)	F1 (%)	MAE
w/o CL	33.39	72.04	1.087	33.79	73.48	0.507
w/o L-common	33.53	71.65	1.072	35.50	**75.57**	0.508
w/o L-specific	33.36	71.87	1.074	35.28	75.12	0.509
DUDF-MSA	**34.29**	**72.30**	**1.062**	**35.66**	75.08	**0.505**

4.5 Sensitivity Analysis of Hyperparameter α

To determine the optimal hyperparameter α for the uncertainty-aware weighting strategy, we conduct a sensitivity analysis. Since $\alpha < 0$, we test a fine-grained set of values: -0.05, -0.1, -0.3, -0.5, -0.7, -0.9, and -1.0. All experiments are conducted under the same modality missing setting, and we report Acc-7/Acc-5 and MAE on the MOSI and SIMS datasets.

As shown in Table 5, our model achieves the best accuracy and the lowest MAE when $\alpha = -0.1$. When $\alpha = -0.05$, the weighting strength is insufficient, leading to slightly inferior performance compared with $\alpha = -0.1$. As $|\alpha|$ increases (i.e., $\alpha \leq -0.3$), the performance declines continuously, indicating that excessive suppression causes the loss of useful information. The performance curve is stable around $\alpha = -0.1$, which verifies that this value is robust and reasonable. Therefore, we uniformly adopt $\alpha = -0.1$ for all experiments in this paper.

Table 5: Sensitivity analysis of hyperparameter α (The best results are shown in bold. Note: The smaller MAE indicate the better performance).

α	MOSI			SIMS		
	Acc-7 (%)	F1 (%)	MAE	Acc-5 (%)	F1 (%)	MAE
−0.05	34.15	72.12	1.065	35.52	74.96	0.506
−0.1	**34.29**	**72.30**	**1.062**	**35.66**	**75.08**	**0.505**
−0.3	34.12	72.08	1.069	35.41	74.87	0.507
−0.5	33.95	71.84	1.075	35.18	74.63	0.509
−0.7	33.73	71.51	1.081	34.95	74.31	0.511

5 Conclusion

This paper proposes a novel framework, DUDF-MSA, to address the multimodal sentiment analysis (MSA) task under conditions of randomly missing data. In DUDF-MSA, to fully mine and exploit the inherent sentiment cues within each modality, we first introduce a Common Feature Extraction (CFE) mechanism that guides the model's attention toward critical cues in incomplete modalities by learning modality-invariant common features. During this stage, contrastive learning is incorporated into the module training to enhance the extraction of fine-grained sentiment features. Subsequently, we design a Specific Feature Uncertainty-Aware Dynamic Adjustment (SFUA) scheme to accurately evaluate the contribution of each modality. SFUA quantifies the class uncertainty of modality-specific features based on their probability distributions and adaptively determines aggregation weights, thereby reducing the negative influence of ambiguous sentiment cues from unreliable modalities during feature aggregation. Finally, the common features and the adaptively

weighted modality-specific features are jointly learned to predict sentiment intensity. Extensive experiments on three public benchmark datasets demonstrate that DUDF-MSA consistently outperforms state-of-the-art methods in handling varying degrees of data incompleteness.

Even with the promising performance of the proposed method, it still has certain limitations. First, the introduction of contrastive learning incurs extra computational cost and training overhead. Second, the hyperparameter α in the uncertainty weighting strategy is determined empirically, and the model exhibits obvious sensitivity to this parameter. For future work, we will conduct further research from three perspectives: developing a more efficient contrastive learning scheme to reduce computational consumption, designing an adaptive mechanism to automatically optimize the hyperparameter α and improve model robustness, and extending the current framework to handle more challenging missing cases, including full modality missing and hardware-caused data corruption.

Acknowledgement: Not applicable.

Funding Statement: This work is supported by the Key Research and Development Program of Zhejiang Province under Grand No. 2024C01026 and Key Research and Promotion Projects of Henan Province under Grand NO. 262102320010.

Author Contributions: Ying Cao contributed to methodology, validation, software, investigation, writing—original draft, supervision, and project administration. Penghui Zhao contributed to methodology, validation, software, writing—original draft, and visualization. Xinyu Qiao contributed to visualization and resources. Ningfan Zhan contributed to formal analysis and data curation. Xiaomei Zou contributed to data curation, resources, funding acquisition, validation, and visualization. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data used in this study are publicly available benchmark datasets: [MOSI] "The data that support the findings of this study are openly available in GitHub at https://github.com/A2Zadeh/CMU-MultimodalSDK". [MOSEI] "The data that support the findings of this study are openly available in GitHub at https://github.com/A2Zadeh/CMU-MultimodalSDK". [SIMS] "The data that support the findings of this study are openly available in GitHub at https://github.com/thuiar/MMSA".

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A Dateset

The MOSI [30] dataset is a widely used dataset comprising three modalities (text, video, and audio). It is collected from 93 YouTube videos, where a speaker expresses their opinion about a movie. It is divided into a training set with 1284 samples, a validation set with 229 samples, and a test set with 686 samples. Each sample has an emotion score ranging from –3 (representing strongly negative emotions) to 3 (representing strongly positive emotions).

The MOSEI [31] dataset is a larger version of MOSI, comprising 22,856 annotated video segments spanning 250 distinct topics. The samples are divided into 16,326 segments for training, 1871 segments for validation, and 4659 segments for testing. Each segment is labelled with a score ranging from −3 (strongly negative) to 3 (strongly positive).

The SIMS [32] dataset is a Chinese multimodal sentiment dataset comprising 2281 video segments from various movies and TV shows. It is divided into 1368 samples for training, 456 samples for validation, and 457 samples for testing. Each sample has a multimodal label and three unimodal labels, with sentiment scores ranging from −1 (strongly negative) to +1 (strongly positive).

Appendix B Baselines

We compare DUDF-MSA against several MSA baselines. All experiments are conducted under the same dataset settings for fairness. We reproduce GLGIS and LNLN in our environment, while the results for other baselines are reported from [8]. Below are brief descriptions of them.

MISA [33] models the shared and specific features of different modalities to improve the robustness of sentiment prediction.

Self-MM [29] generates unimodal labels and performs multi-task training to capture both consistent and divergent representations.

CENET [34] is a cross-modal enhancement network that adaptively aligns and fuses multimodal signals through a cross-attention interaction mechanism.

TFRNet [6] is a transformer-based model that combines fusion-reconstruction strategies to improve sentiment prediction performance under incomplete modality conditions.

ALMT [17] leverages linguistic features to suppress irrelevant or conflicting information from visual and audio inputs, and learns complementary multimodal representations.

GLGIS [35] models the common and specific features of different modalities, while incorporating contrastive learning guided by sentiment intensity to improve the robustness of sentiment prediction.

LNLN [8] treats the language modality as the dominant modality and introduces dominant modality correction and dominant-modality-based multimodal learning to enhance robustness in noisy and missing modality scenarios.

Appendix C Implementation Details

The experiments on the MOSI and SIMS datasets were conducted using a single NVIDIA RTX 3080 Ti GPU, while those on the MOSEI dataset were performed using an NVIDIA RTX 4060 Ti. The training mode adapted was complete training, without any additional pre-training. Both the video and audio Transformer encoder consists of two layers. The AdamW optimizer was used for parameter optimization.

To ensure consistent and fair comparisons across all methods, each experiment was repeated three times with fixed random seeds of 1111, 1112, and 1113. The key hyperparameters used in our experiments are summarized in Table A1.

Table A1: Parameter settings. Hyper-parameters of DUDF-MSA for the multimodal sentiment analysis.

Para	MOSI	MOSEI	SIMS
Seed	1111, 1112, 1113	1111, 1112, 1113	1111, 1112, 1113
Batch-size	64	128	64
Bert lr	5e−5	5e−5	5e−5
Audio Encoder lr	1e−3	5e−4	5e−4
Visual Encoder lr	5e−3	5e−4	5e−4
Others lr	1e−2	25e−4	5e−4
d_c & d_s	64	128	64

Appendix D Details of Robust Comparison

We conduct comprehensive experiments under different missing rate conditions to further assess model robustness and examine the impact of discarding different modalities. The detailed results on the MOSI, MOSEI, and SIMS datasets are presented in Tables A2–A4, respectively.

Table A2: Details of robust comparison on MOSI with different random missing rates (The best results are shown in bold. Note: The smaller MAE indicates the better performance).

Method	Acc-7 (%)	Acc-5 (%)	Acc-2 (%)	F1 (%)	MAE	Corr
Random Missing Rate r = 0						
MISA [33]	43.05	48.30	82.78/81.24	82.83/81.23	0.771	0.777
Self-MM [29]	42.81	**52.38**	**85.22/83.24**	**85.19/83.26**	0.781	**0.790**
CENET [34]	43.20	50.39	83.08/81.49	83.06/81.48	0.748	0.785
TFR-Net [6]	40.82	47.91	83.64/81.68	83.57/81.61	0.805	0.760
ALMT [17]	42.37	48.49	84.91/82.75	85.01/82.94	0.752	0.768
GLGIS [35]	44.18	50.34	84.21/82.18	84.38/82.32	0.804	0.771
LNLN [8]	42.17	48.01	82.22/80.81	82.29/80.83	0.782	0.778
DUDF-MSA	**45.22**	51.6	84.71/81.87	84.75/82.03	**0.732**	0.773
Random Missing Rate r = 0.1						
MISA [33]	40.28	46.21	80.18/79.01	80.21/78.97	0.847	0.721
Self-MM [29]	40.33	**49.03**	81.40/80.03	81.19/80.03	0.812	**0.728**
CENET [34]	40.13	46.60	80.08/78.38	79.91/78.20	0.837	0.719
TFR-Net [6]	38.63	45.82	79.27/77.99	78.70/77.61	0.872	0.705
ALMT [17]	39.84	45.48	80.90/78.67	81.15/79.08	0.843	0.703
GLGIS [35]	40.90	47.54	**82.42/80.09**	**82.61/81.12**	0.871	0.707
LNLN [8]	40.14	45.96	79.73/78.60	79.81/78.75	0.826	**0.728**
DUDF-MSA	**41.45**	47.42	82.16/79.64	82.22/80.03	**0.816**	0.720
Random Missing Rate r = 0.2						
MISA [33]	36.25	41.55	77.54/76.34	77.58/76.30	0.939	0.654
Self-MM [29]	36.64	43.98	78.15/76.48	77.76/76.51	0.901	0.660
CENET [34]	38.00	42.32	77.49/74.64	77.35/74.28	0.916	0.654
TFR-Net [6]	34.70	40.13	74.70/73.52	73.57/72.70	0.987	0.622
ALMT [17]	35.33	40.33	77.64/75.70	77.94/76.24	0.927	0.645
GLGIS [35]	37.72	43.42	**79.52/77.51**	79.47/**77.62**	0.933	0.659
LNLN [8]	38.63	43.97	78.20/77.40	78.27/77.49	**0.885**	**0.675**
DUDF-MSA	**38.92**	**44.56**	79.27/77.45	**79.96**/77.50	0.900	0.666
Random Missing Rate r = 0.3						
MISA [33]	34.60	38.97	75.76/74.54	75.82/74.51	0.989	0.618
Self-MM [29]	34.89	40.67	76.37/74.98	75.68/74.94	0.967	0.614
CENET [34]	34.74	38.97	76.83/72.01	76.56/71.30	0.983	0.605
TFR-Net [6]	32.55	38.34	72.36/71.28	70.12/69.58	1.065	0.572
ALMT [17]	33.04	37.17	75.15/72.94	75.51/73.66	0.992	0.596
GLGIS [35]	36.38	42.10	**77.74**/75.19	**76.95**/76.23	0.980	0.614

(Continued)

Table A2 (continued)

Method	Acc-7 (%)	Acc-5 (%)	Acc-2 (%)	F1 (%)	MAE	Corr
LNLN [8]	37.27	41.69	76.73/75.46	76.79/75.55	0.954	**0.620**
DUDF-MSA	**37.90**	**43.30**	77.15/**76.85**	**76.95/76.88**	**0.953**	0.614
Random Missing Rate r = 0.4						
MISA [33]	32.65	35.37	73.88/72.59	73.88/72.49	1.041	0.585
Self-MM [29]	31.20	36.30	73.17/71.96	71.74/71.75	1.027	0.579
CENET [34]	32.26	36.15	73.38/71.53	72.75/70.26	1.031	0.574
TFR-Net [6]	30.17	35.76	68.75/67.74	64.71/64.41	1.142	0.537
ALMT [17]	31.44	35.03	73.12/71.14	73.85/72.47	1.045	0.560
GLGIS [35]	36.33	40.55	**76.40/74.97**	**75.90/75.43**	1.102	**0.594**
LNLN [8]	36.10	40.23	75.10/73.97	75.13/73.99	**0.991**	0.593
DUDF-MSA	**36.93**	**41.31**	75.45/74.66	75.45/74.68	0.995	0.589
Random Missing Rate r = 0.5						
MISA [33]	28.14	30.61	70.53/69.34	70.50/69.20	1.124	0.519
Self-MM [29]	26.97	31.39	67.43/67.54	64.27/66.81	1.129	0.503
CENET [34]	28.33	30.90	72.46/66.08	71.10/63.50	1.130	0.496
TFR-Net [6]	25.85	30.71	64.83/63.02	58.04/56.64	1.270	0.443
ALMT [17]	28.42	31.25	68.24/65.94	69.74/68.54	1.138	0.485
GLGIS [35]	33.77	37.39	**73.43/72.11**	**73.43/73.51**	**1.078**	0.539
LNLN [8]	34.36	37.80	72.61/71.67	72.66/71.80	1.085	0.521
DUDF-MSA	**34.84**	**39.16**	72.82/71.53	72.77/71.44	**1.078**	**0.558**
Random Missing Rate r = 0.6						
MISA [33]	24.68	27.12	66.97/65.84	66.94/65.69	1.200	0.441
Self-MM [29]	24.34	27.31	63.47/63.36	58.94/62.07	1.209	0.425
CENET [34]	24.54	26.53	67.58/61.47	64.87/57.86	1.215	0.415
TFR-Net [6]	24.05	28.33	61.64/59.47	52.44/50.53	1.371	0.363
ALMT [17]	25.41	27.36	64.53/62.15	66.81/65.87	1.214	0.407
GLGIS [35]	29.74	33.13	**69.00/67.43**	**69.21/68.27**	1.180	**0.443**
LNLN [8]	31.39	33.97	68.09/66.18	68.07/66.28	**1.168**	0.441
DUDF-MSA	**31.63**	**34.81**	68.14/67.15	68.03/67.08	1.176	0.438
Random Missing Rate r = 0.7						
MISA [33]	21.14	23.27	65.09/63.89	65.07/63.74	1.257	0.381
Self-MM [29]	20.70	23.81	61.74/61.46	55.11/58.97	1.271	0.339
CENET [34]	22.35	23.57	63.82/59.43	53.79/54.22	1.269	0.335
TFR-Net [6]	23.71	26.92	59.91/57.34	48.41/45.48	1.454	0.276
ALMT [17]	23.71	24.97	61.84/59.67	65.30/65.19	1.266	0.336
GLGIS [35]	27.40	**30.34**	64.98/63.95	64.28/63.66	1.242	0.335
LNLN [8]	27.16	29.45	63.57/62.49	63.44/62.65	**1.258**	0.368
DUDF-MSA	**27.70**	30.27	**65.96/64.67**	**65.95/64.60**	1.280	**0.386**

(Continued)

Table A2 (continued)

Method	Acc-7 (%)	Acc-5 (%)	Acc-2 (%)	F1 (%)	MAE	Corr
Random Missing Rate r = 0.8						
MISA [33]	19.92	20.99	**63.56**/62.24	63.16/61.67	1.311	**0.321**
Self-MM [29]	19.29	22.11	59.55/58.26	49.98/53.56	1.313	0.282
CENET [34]	21.14	21.67	60.93/57.53	54.68/50.80	1.314	0.274
TFR-Net [6]	23.23	27.70	58.49/55.98	44.70/41.88	1.497	0.155
ALMT [17]	23.13	23.66	60.37/58.31	**65.45**/**66.14**	1.310	0.273
GLGIS [35]	22.81	26.89	61.99/**62.66**	62.61/61.44	**1.300**	0.305
LNLN [8]	23.71	26.73	60.26/58.99	60.04/59.14	1.326	0.288
DUDF-MSA	**25.92**	**27.93**	61.89/61.81	61.73/61.98	1.327	0.306
Random Missing Rate r = 0.9						
MISA [33]	17.78	18.41	58.64/**58.21**	56.84/56.19	1.369	**0.226**
Self-MM [29]	18.32	19.78	58.59/55.25	46.16/47.46	**1.353**	0.197
CENET [34]	19.15	19.10	**58.99**/54.76	50.01/46.58	1.357	0.181
TFR-Net [6]	21.67	25.12	57.93/55.44	43.01/40.18	1.534	0.155
ALMT [17]	20.31	20.50	57.32/56.66	**64.92**/**67.82**	1.349	0.205
GLGIS [35]	20.99	22.32	54.18/55.98	54.79/56.66	1.380	0.194
LNLN [8]	21.62	23.23	54.88/53.60	54.72/53.67	1.401	0.187
DUDF-MSA	**22.40**	**24.00**	54.52/56.01	54.79/56.43	1.360	0.202

Table A3: Details of robust comparison on MOSEI with different random missing rates (The best results are shown in bold. Note: The smaller MAE indicates the better performance).

Method	Acc-7 (%)	Acc-5 (%)	Acc-2 (%)	F1 (%)	MAE	Corr
Random Missing Rate r = 0						
MISA [33]	51.79	53.85	85.28/84.10	85.10/83.75	0.552	0.759
Self-MM [29]	53.89	55.72	85.34/84.68	85.11/84.66	**0.531**	0.764
CENET [34]	**54.39**	**56.12**	85.49/82.30	85.41/82.60	**0.531**	**0.770**
TFR-Net [6]	53.71	47.91	84.96/84.65	84.71/84.34	0.550	0.745
ALMT [17]	52.18	53.89	85.62/83.99	85.69/84.53	0.542	0.752
GLGIS [35]	52.55	54.36	85.53/84.33	85.62/84.38	0.572	0.763
LNLN [8]	50.02	51.89	84.68/82.85	84.60/84.02	0.573	0.759
DUDF-MSA	52.33	53.85	**85.86**/**84.76**	**86.13**/**84.94**	0.542	0.762
Random Missing Rate r = 0.1						
MISA [33]	50.13	51.34	82.21/82.28	81.28/80.79	0.598	0.722
Self-MM [29]	51.80	53.18	83.03/83.79	82.43/83.23	0.564	0.725
CENET [34]	**52.83**	**54.23**	83.75/82.41	83.42/82.34	**0.556**	0.739
TFR-Net [6]	52.29	45.82	82.92/83.31	82.25/82.40	0.573	0.715
ALMT [17]	49.98	51.38	84.14/82.84	84.23/83.04	0.583	0.718
GLGIS [35]	52.48	53.75	**84.45**/84.14	84.57/84.26	0.560	**0.740**

(Continued)

Table A3 (continued)

Method	Acc-7 (%)	Acc-5 (%)	Acc-2 (%)	F1 (%)	MAE	Corr
LNLN [8]	50.40	51.98	83.39/82.37	83.45/82.37	0.579	0.732
DUDF-MSA	52.39	53.64	84.34/**84.35**	**84.82/84.60**	**0.559**	**0.740**
Random Missing Rate r = 0.2						
MISA [33]	47.24	47.66	77.84/79.93	75.56/76.88	0.659	0.674
Self-MM [29]	49.44	50.51	80.84/82.33	79.76/81.17	0.604	0.678
CENET [34]	50.72	51.85	81.46/81.62	80.78/81.17	0.590	0.698
TFR-Net [6]	51.04	40.13	80.47/81.61	79.29/79.99	0.604	0.672
ALMT [17]	46.61	47.82	82.71/81.65	82.82/81.83	0.607	0.669
GLGIS [35]	**51.15**	**52.61**	**83.54**/83.02	**83.70/83.18**	0.586	**0.717**
LNLN [8]	49.85	51.23	82.35/81.80	82.47/81.86	0.596	0.706
DUDF-MSA	51.00	52.09	83.43/**83.71**	84.85/84.03	**0.584**	0.707
Random Missing Rate r = 0.3						
MISA [33]	43.99	43.40	73.32/77.28	68.91/72.25	0.724	0.615
Self-MM [29]	47.23	48.07	77.63/79.99	75.69/77.74	0.653	0.610
CENET [34]	48.49	49.37	78.65/80.02	77.34/78.94	0.636	0.640
TFR-Net [6]	48.75	38.34	77.48/79.29	75.43/76.52	0.650	0.604
ALMT [17]	43.04	44.05	80.94/79.94	81.15/80.20	0.632	0.598
GLGIS [35]	49.07	**50.42**	**82.31**/81.52	82.54/81.75	0.613	**0.678**
LNLN [8]	48.42	49.50	81.10/80.80	81.27/81.00	0.622	0.671
DUDF-MSA	**49.22**	50.16	81.73/**82.52**	**82.26/82.09**	**0.611**	0.671
Random Missing Rate r = 0.4						
MISA [33]	40.87	39.53	70.46/75.04	64.02/67.93	0.780	0.561
Self-MM [29]	44.40	45.04	75.02/78.09	72.01/74.48	0.694	0.554
CENET [34]	47.12	47.74	76.03/78.57	73.87/76.75	0.678	0.587
TFR-Net [6]	46.70	35.76	74.74/77.65	71.67/73.71	0.688	0.548
ALMT [17]	40.40	41.21	79.40/79.16	79.68/79.50	0.651	0.536
GLGIS [35]	48.44	49.21	**80.79**/81.00	81.17/81.38	0.637	**0.641**
LNLN [8]	47.21	48.14	79.59/79.73	79.82/80.00	0.648	0.636
DUDF-MSA	**48.66**	**49.56**	80.68/**81.18**	**81.34/81.84**	**0.633**	**0.641**
Random Missing Rate r = 0.5						
MISA [33]	38.12	36.05	67.38/73.21	58.38/64.14	0.834	0.492
Self-MM [29]	42.70	43.14	71.97/75.81	67.40/70.38	0.733	0.477
CENET [34]	45.12	45.52	73.33/77.16	69.80/74.14	0.720	0.515
TFR-Net [6]	45.00	30.71	71.53/75.69	66.88/70.07	0.730	0.471
ALMT [17]	37.82	38.34	77.40/77.48	77.73/77.80	0.683	0.461
GLGIS [35]	46.34	47.41	78.95/79.42	79.53/79.96	0.672	0.589
LNLN [8]	45.66	46.33	77.72/78.17	78.04/78.74	0.681	0.581
DUDF-MSA	**46.83**	**47.63**	**79.22/80.30**	**80.02/81.13**	**0.669**	**0.595**

(Continued)

Table A3 (continued)

Method	Acc-7 (%)	Acc-5 (%)	Acc-2 (%)	F1 (%)	MAE	Corr
Random Missing Rate r = 0.6						
MISA [33]	36.16	33.30	65.55/72.30	54.64/62.12	0.875	0.415
Self-MM [29]	41.47	41.75	69.33/73.93	63.01/66.76	0.762	0.401
CENET [34]	44.45	44.64	70.50/75.39	65.27/70.86	0.749	0.446
TFR-Net [6]	43.88	28.33	68.80/74.05	62.51/67.07	0.762	0.397
ALMT [17]	35.99	36.30	74.98/76.26	75.44/76.71	0.710	0.395
GLGIS [35]	44.92	45.65	76.22/**78.56**	77.15/79.37	**0.697**	0.533
LNLN [8]	44.85	45.35	75.45/76.23	75.93/77.09	0.710	0.532
DUDF-MSA	**45.89**	**46.38**	**76.33**/78.26	**77.21**/**79.51**	0.704	**0.540**
Random Missing Rate r = 0.7						
MISA [33]	34.54	31.21	64.28/71.71	51.82/60.65	0.906	0.344
Self-MM [29]	39.93	40.12	66.79/72.55	58.05/63.45	0.786	0.329
CENET [34]	43.93	44.03	67.50/73.39	59.88/67.02	0.776	0.384
TFR-Net [6]	42.91	26.92	66.64/72.77	58.32/64.02	0.786	0.322
ALMT [17]	34.78	34.95	71.62/73.98	72.24/74.54	0.743	0.315
GLGIS [35]	44.45	44.99	73.06/76.15	74.88/77.45	0.743	0.468
LNLN [8]	43.63	43.98	73.06/74.82	73.69/76.04	**0.742**	**0.473**
DUDF-MSA	**44.99**	**45.20**	**73.50**/**76.26**	**75.54**/**78.04**	**0.742**	0.468
Random Missing Rate r = 0.8						
MISA [33]	33.29	29.51	63.43/71.30	49.95/59.69	0.927	0.267
Self-MM [29]	38.69	38.78	65.07/71.83	54.44/61.49	0.805	0.259
CENET [34]	42.71	42.74	65.88/72.16	56.80/64.67	0.798	0.316
TFR-Net [6]	42.23	27.70	65.05/71.95	54.91/61.82	0.807	0.241
ALMT [17]	34.01	34.09	68.15/71.48	69.12/72.28	0.774	0.231
GLGIS [35]	43.01	43.29	**71.08**/**74.46**	**74.30**/76.79	0.780	0.394
LNLN [8]	42.52	42.60	69.48/72.88	70.42/74.69	**0.772**	**0.396**
DUDF-MSA	**43.66**	**43.79**	70.67/74.31	73.13/**77.30**	0.790	0.386
Random Missing Rate r = 0.9						
MISA [33]	32.29	28.03	62.95/71.07	48.80/59.12	0.941	0.180
Self-MM [29]	37.46	37.50	63.85/71.24	51.32/59.72	0.821	0.188
CENET [34]	42.08	42.08	64.14/70.42	54.27/62.33	0.814	0.254
TFR-Net [6]	41.73	25.12	63.64/71.34	52.02/59.99	0.820	0.175
ALMT [17]	34.40	34.40	61.41/68.65	63.32/69.83	0.810	0.138
GLGIS [35]	41.81	41.88	**66.14**/**72.78**	**71.65**/**77.24**	0.818	**0.287**
LNLN [8]	**42.12**	**42.18**	65.39/70.73	67.23/74.00	**0.808**	0.286
DUDF-MSA	41.70	41.75	65.60/71.69	69.66/76.10	0.831	0.251

Table A4: Details of robust comparison on SIMS with different random missing rates. (The best results are shown in bold. Note: The smaller MAE indicates the better performance).

Method	Acc-5 (%)	Acc-3 (%)	Acc-2 (%)	F1 (%)	MAE	Corr
Random Missing Rate r = 0						
MISA [33]	40.55	63.38	78.19	77.22	0.449	0.576
Self-MM [29]	40.77	**64.92**	78.26	78.00	0.421	0.584
CENET [34]	23.85	54.05	68.71	57.82	0.578	0.137
TFR-Net [6]	33.85	54.12	69.15	58.44	0.562	0.254
ALMT [17]	23.41	54.78	75.64	76.27	0.527	0.536
GLGIS [35]	41.06	64.33	77.82	79.86	0.434	0.567
LNLN [8]	38.73	62.91	78.69	79.73	0.453	0.538
DUDF-MSA	**41.21**	64.33	**79.73**	**80.05**	**0.423**	**0.576**
Random Missing Rate r = 0.1						
MISA [33]	38.88	63.02	77.39	75.82	0.461	0.561
Self-MM [29]	40.26	63.53	77.32	76.76	0.433	0.563
CENET [34]	22.83	53.98	68.57	57.36	0.580	0.136
TFR-Net [6]	30.12	53.25	68.85	59.38	0.596	0.203
ALMT [17]	22.10	55.14	74.40	75.19	0.530	0.537
GLGIS [35]	**41.21**	**63.90**	77.82	80.08	0.439	0.549
LNLN [8]	37.86	62.14	78.40	79.53	0.461	0.527
DUDF-MSA	41.08	63.09	**79.29**	**80.31**	**0.435**	**0.564**
Random Missing Rate r = 0.2						
MISA [33]	38.15	59.23	74.33	71.70	0.489	0.490
Self-MM [29]	38.37	61.71	74.98	73.71	0.464	0.500
CENET [34]	22.25	54.20	68.57	57.64	0.583	0.132
TFR-Net [6]	29.03	53.61	68.64	59.74	0.619	0.191
ALMT [17]	21.08	53.17	72.65	73.90	0.541	0.485
GLGIS [35]	38.61	**62.58**	75.28	77.20	0.460	0.507
LNLN [8]	36.73	59.81	**77.32**	**78.65**	0.479	0.497
DUDF-MSA	**40.91**	61.86	76.08	76.98	**0.458**	**0.515**
Random Missing Rate r = 0.3						
MISA [33]	36.40	59.30	74.11	70.40	0.505	0.464
Self-MM [29]	37.93	59.81	74.76	72.85	0.474	0.487
CENET [34]	21.44	54.05	68.42	57.41	0.578	0.175
TFR-Net [6]	27.64	52.30	68.42	59.88	0.640	0.182
ALMT [17]	20.35	50.62	72.06	73.64	0.546	0.469
GLGIS [35]	38.39	**61.48**	**75.78**	**77.69**	0.462	0.497
LNLN [8]	36.95	58.35	75.53	77.38	0.486	0.471
DUDF-MSA	**40.24**	61.41	**75.78**	76.95	**0.458**	**0.508**
Random Missing Rate r = 0.4						
MISA [33]	34.86	57.33	72.87	67.52	0.523	0.436
Self-MM [29]	34.57	58.28	73.30	70.36	0.482	0.479
CENET [34]	22.54	54.12	68.49	57.68	0.583	0.141
TFR-Net [6]	25.31	51.86	67.91	59.16	0.664	0.176
ALMT [17]	19.91	49.45	70.75	72.97	0.549	0.470
GLGIS [35]	38.51	**60.46**	**74.62**	75.80	0.481	0.464
LNLN [8]	36.97	58.93	74.43	**76.81**	0.496	0.444
DUDF-MSA	**39.02**	59.66	73.89	75.46	**0.461**	**0.506**
Random Missing Rate r = 0.5						
MISA [33]	30.56	54.78	71.26	64.16	0.552	0.367

(Continued)

Table A4 (continued)

Method	Acc-5 (%)	Acc-3 (%)	Acc-2 (%)	F1 (%)	MAE	Corr
Self-MM [29]	32.02	53.90	71.41	67.11	0.517	0.390
CENET [34]	23.12	54.05	68.71	57.92	0.588	0.107
TFR-Net [6]	24.65	52.37	67.47	58.66	0.685	0.171
ALMT [17]	18.38	47.12	68.27	71.22	0.563	0.395
GLGIS [35]	36.19	56.69	72.28	75.49	0.513	0.406
LNLN [8]	36.46	**58.16**	**73.84**	**76.74**	0.513	0.413
DUDF-MSA	**37.13**	57.99	73.30	74.93	**0.500**	**0.429**
Random Missing Rate r = 0.6						
MISA [33]	27.72	53.97	70.46	61.81	0.578	0.286
Self-MM [29]	29.10	51.86	70.02	64.21	0.548	0.313
CENET [34]	22.46	53.69	69.00	58.64	0.592	0.102
TFR-Net [6]	24.80	52.59	67.03	58.30	0.696	0.157
ALMT [17]	18.67	43.69	66.81	70.69	0.574	0.322
GLGIS [35]	33.36	55.45	70.96	73.57	0.533	0.354
LNLN [8]	32.97	**55.83**	71.22	**75.34**	0.538	0.351
DUDF-MSA	**34.21**	55.58	**71.63**	73.21	**0.532**	**0.364**
Random Missing Rate r = 0.7						
MISA [33]	24.87	52.52	**69.95**	59.54	0.601	0.167
Self-MM [29]	25.53	50.62	69.58	62.28	0.571	0.198
CENET [34]	21.81	53.32	67.69	57.87	0.599	0.070
TFR-Net [6]	23.78	52.30	67.18	58.15	0.707	0.163
ALMT [17]	18.02	38.66	65.57	70.27	0.586	0.218
GLGIS [35]	30.71	53.12	69.22	73.33	0.602	0.259
LNLN [8]	**30.93**	**53.32**	68.82	**73.98**	0.568	0.272
DUDF-MSA	30.36	51.63	69.65	72.46	**0.556**	**0.281**
Random Missing Rate r = 0.8						
MISA [33]	22.69	52.22	69.37	57.82	0.610	0.092
Self-MM [29]	22.03	50.77	**69.51**	60.68	**0.585**	0.138
CENET [34]	21.73	52.15	67.47	58.44	0.599	0.074
TFR-Net [6]	22.97	**52.74**	67.54	57.55	0.721	0.100
ALMT [17]	18.60	34.06	64.19	69.64	0.597	0.133
GLGIS [35]	25.21	50.93	69.14	73.13	0.625	0.167
LNLN [8]	**26.66**	52.55	66.88	**73.53**	0.590	0.197
DUDF-MSA	26.15	51.13	68.27	71.03	0.597	**0.217**
Random Missing Rate r = 0.9						
MISA [33]	20.64	52.95	**69.22**	57.01	0.617	0.041
Self-MM [29]	22.17	52.15	68.92	58.32	0.586	0.111
CENET [34]	20.86	48.07	65.72	58.18	0.609	−0.002
TFR-Net [6]	23.05	**53.76**	69.08	57.71	0.721	0.088
ALMT [17]	19.47	26.91	66.23	73.76	**0.596**	0.076
GLGIS [35]	22.34	48.58	67.76	74.02	0.645	0.104
LNLN [8]	26.15	51.23	66.84	**74.98**	0.604	**0.139**
DUDF-MSA	**26.48**	50.47	64.48	69.51	0.633	0.111

References

1. Liu Y, Yuan Z, Mao H, Liang Z, Yang W, Qiu Y, et al. 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 Nov 7–11; Bengaluru, India. p. 247–58.

2. Sun Z, Sarma P, Sethares W, Liang Y. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2020 Feb 7–12; New York, NY, USA. p. 8992–9.
3. Hu G, Lin TE, Zhao Y, Lu G, Wu Y, Li Y. UniMSE: towards unified multimodal sentiment analysis and emotion recognition. arXiv:2211.11256. 2022.
4. Aguilar G, Rozgić V, Wang W, Wang C. Multimodal and multi-view models for emotion recognition. arXiv:1906.10198. 2019.
5. Pham H, Liang PP, Manzini T, Morency LP, Póczos B. Found in translation: learning robust joint representations by cyclic translations between modalities. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2019 Jan 27–Feb 1; Honolulu, HI, USA. p. 6892–9.
6. Yuan Z, Li W, Xu H, Yu W. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In: Proceedings of the 29th ACM International Conference on Multimedia; 2021 Oct 20–24; Chengdu, China. p. 4400–7.
7. Guo Z, Jin T, Zhao Z. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (volume 1: Long Papers); 2024 Aug 11–16; Bangkok, Thailand. p. 1726–36.
8. Zhang H, Wang W, Yu T. Towards robust multimodal sentiment analysis with incomplete data. Adv Neural Inf Process Syst. 2024;37:55943–74. doi:10.52202/079017-1779.
9. Li X, Cheng X, Miao D, Zhang X, Li Z. TF-Mamba: text-enhanced fusion mamba with missing modalities for robust multimodal sentiment analysis. In: Christodoulopoulos C, Chakraborty T, Rose C, Peng V, editors. Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2025; 2025 Nov 4–9; Suzhou, China; 2025. p. 11252–67.
10. Golagana V, Row SV, Rao PS. Adaptive multimodal sentiment analysis: improving fusion accuracy with dynamic attention for missing modality. J Electr Syst. 2024;20:134.
11. Zeng J, Zhou J, Liu T. Mitigating inconsistencies in multimodal sentiment analysis under uncertain missing modalities. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing; 2022 Dec 7–11; Abu Dhabi, United Arab Emirates. p. 2924–34.
12. Wu Y, Lin Z, Zhao Y, Qin B, Zhu LN. A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis. In: Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021; 2021 Aug 1–6; Virtual. p. 4730–8. doi:10.18653/v1/2021.findings-acl.417.
13. Yang J, Yu Y, Niu D, Guo W, Xu Y. Confede: contrastive feature decomposition for multimodal sentiment analysis. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2023 Jul 9–14; Toronto, ON, Canada. p. 7617–30.
14. Li X, Zhang H, Dong Z, Cheng X, Liu Y, Zhang X. Learning fine-grained representation with token-level alignment for multimodal sentiment analysis. Expert Syst Appl. 2025;269:126274.
15. Sun K, Tian M. Sequential fusion of text-close and text-far representations for multimodal sentiment analysis. In: Rambow O, Wanner L, Apidianaki M, Al-Khalifa H, Eugenio BD, Schockaert S, editors. Proceedings of the 31st International Conference on Computational Linguistics; 2025 Jan 19–24; Abu Dhabi, United Arab Emirates, p. 40–9.
16. Han W, Chen H, Poria S. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. arXiv:2109.00412. 2021.
17. Zhang H, Wang Y, Yin G, Liu K, Liu Y, Yu T. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. arXiv:2310.05804. 2023.
18. Li M, Yang D, Lei Y, Wang S, Wang S, Su L, et al. A unified self-distillation framework for multimodal sentiment analysis with uncertain missing modalities. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2024 Feb 20–27; Vancouver, BC, Canada. p. 10074–82.
19. Li L, Xu G, Shen X, Xu Z, Zhang Y, Zhang Z, et al. Senti-iFusion: an integrity-centered hierarchical fusion framework for multimodal sentiment analysis under uncertain modality missingness. arXiv:2511.16990. 2025.

20. Bao J, Qian J, Yan M, Yang J. Progressive representation learning for multimodal sentiment analysis with incomplete modalities. arXiv:2603.09111. 2026.
21. Liu W, Wang X, Owens J, Li Y. Energy-based out-of-distribution detection. Adv Neural Inf Process Syst. 2020;33:21464–75.
22. Ming Y, Fan Y, Li Y. Poem: out-of-distribution detection with posterior sampling. In: Proceedings of the International Conference on Machine Learning, PMLR; 2022 Jul 17–23; Baltimore, MD, USA. p. 15650–65.
23. Tellamekala MK, Amiriparian S, Schuller BW, André E, Giesbrecht T, Valstar M. COLD fusion: calibrated and ordinal latent distribution fusion for uncertainty-aware multimodal emotion recognition. IEEE Trans Pattern Anal Mach Intell. 2023;46(2):805–22.
24. Chen YT, Shi J, Ye Z, Mertz C, Ramanan D, Kong S. Multimodal object detection via probabilistic ensembling. In: European Conference on Computer Vision. Berlin/Heidelberg, Germany: Springer; 2022. p. 139–58.
25. Ma H, Han Z, Zhang C, Fu H, Zhou JT, Hu Q. Trustworthy multimodal regression with mixture of normal-inverse gamma distributions. Adv Neural Inf Process Syst. 2021;34:6881–93.
26. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T, editors. 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); 2019 Jun 2–7; Minneapolis, MN, USA, 4171–86.
27. McFee B, Raffel C, Liang D, Ellis DP, McVicar M, Battenberg E, et al. librosa: audio and music signal analysis in python. SciPy. 2015;2015:18–24.
28. Baltrusaitis T, Zadeh A, Lim YC, Morency LP. OpenFace 2.0: facial behavior analysis toolkit. In: Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018); 2018 May 15–19; Xi'an, China. p. 59–66.
29. Yu W, Xu H, Yuan Z, Wu J. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2021 Feb 2–9; Virtual. p. 10790–7.
30. Zadeh A, Zellers R, Pincus E, Morency LP. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv:1606.06259. 2016.
31. Zadeh AB, Liang PP, Poria S, Cambria E, Morency LP. 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 Jul 15–20; Melbourne, Australia. p. 2236–46.
32. Yu W, Xu H, Meng F, Zhu Y, Ma Y, Wu J, et al. 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 Jul 5–10; Virtual. p. 3718–27.
33. Hazarika D, Zimmermann R, Poria S. Misa: modality-invariant and-specific representations for multimodal sentiment analysis. In: Proceedings of the 28th ACM International Conference on Multimedia; 2020 Oct 12–16; Seattle, WA, USA. p. 1122–31.
34. Tao H, Xie C, Wang J, Xin Z. CENet: a channel-enhanced spatiotemporal network with sufficient supervision information for recognizing industrial smoke emissions. IEEE Internet Things J. 2022;9(19):18749–59.
35. Yang Y, Dong X, Qiang Y. CLGSI: a multimodal sentiment analysis framework based on contrastive learning guided by sentiment intensity. In: Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2024; 2024 May 22–27; Mexico City, Mexico. p. 2099–110.