Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.085382

ARTICLE

RUAL: Uncertainty-Aware Learning for Robust Multimodal Sentiment Analysis

Weijun Gao, Ziyang Zhang* and Maotang Su

School of Computer Science and Artificial Intelligence, Lanzhou University of Technology, Lanzhou, China

*Corresponding Author: Ziyang Zhang. Email: 13935034996@163.com

Received: 10 May 2026; Accepted: 23 June 2026; Published: 13 August 2026

ABSTRACT: Multimodal sentiment analysis (MSA) has made significant progress in integrating heterogeneous information from text, speech, and vision. However, real-world multimodal data often suffer from modality noise, semantic inconsistency, and incomplete modality information, which can weaken cross-modal fusion and reduce the reliability of sentiment prediction. To address these challenges, this paper proposes RUAL, a robust uncertainty-aware learning framework for multimodal sentiment analysis. Specifically, RUAL first employs a Gathered Multi-Head Attention Pooling (GMHA) module to aggregate intra-modal features and estimate modality uncertainty based on attention entropy. Then, an Uncertainty-Aware Cross-Modal Coupled Layer (UACCL) is introduced to dynamically regulate cross-modal residual fusion according to sample confidence, thereby reducing the negative influence of unreliable modalities on fused representations. In addition, uncertainty-weighted learning and uncertainty-guided self-distillation (UWL and U-SD) are jointly integrated through an optimization strategy to further improve training stability and generalization in complex scenarios. Experimental results on CMU-MOSI, CMU-MOSEI, and MVSA-Single demonstrate that RUAL achieves strong overall performance and maintains stable prediction results under missing-modality and Gaussian-noise conditions, validating the effectiveness and robustness of the proposed framework for multimodal sentiment analysis.

KEYWORDS: Multimodal sentiment analysis; uncertainty-aware learning; robust learning; cross-modal alignment; self-distillation; weighted loss

1 Introduction

Against the backdrop of rapid advancements in digital media and intelligent perception technologies, the fusion and understanding of multimodal information have emerged as a significant research direction in artificial intelligence. This field demonstrates extensive value in applications such as video emotion recognition [1–3], cross-modal retrieval [4–6], and human-computer interaction [7,8]. As a vital branch of multimodal learning, Multimodal Sentiment Analysis (MSA) aims to jointly model heterogeneous modalities such as text, speech, and vision, thereby providing a more comprehensive and granular characterization of human emotional states. Compared to unimodal approaches, multimodal sentiment analysis integrates complementary cues such as linguistic semantics, speech prosody, and facial expressions, demonstrating superior representational capabilities in complex emotional expression scenarios.

Despite significant advancements in recent years, multimodal sentiment analysis continues to face numerous challenges in real-world applications. Existing research methods [9–12] primarily focus on designing multimodal fusion strategies, enhancing sentiment recognition performance by modeling information interactions between modalities at different levels. Based on the fusion stage, these methods are typically

categorized as early fusion [13], mid-fusion [14,15], and late fusion [16]. While achieving structural design success, these approaches often rely on carefully constructed cross-modal fusion mechanisms and implicitly assume relatively stable consistency in quality and semantic expression across modalities.

However, in real-world scenarios, multimodal data often deviates from these ideal assumptions, exhibiting substantial differences in modality quality, completeness, and semantic consistency. For example, speech may be corrupted by background noise, visual signals may be degraded by occlusion or lighting variation, and text may contain ambiguity or implicit sentiment. These factors can create reliability imbalances across modalities and cause noisy modalities to mislead multimodal fusion.

In recent years, some studies have attempted to mitigate these issues through uncertainty modeling. For example, uncertainty-weighted fusion methods employ Gaussian variance to characterize sample noise [17], while modal recovery methods alleviate representation bias caused by missing modalities through distribution consistency constraints [18]. Although these approaches enhance model robustness against noise to some extent, their uncertainty modeling often operates only at specific stages of the model. They lack holistic modeling of multimodal feature learning, cross-modal alignment, and training optimization processes, resulting in limited robustness in complex real-world scenarios.

As illustrated in Fig. 1, existing multimodal sentiment analysis models face interference from multiple uncertainty factors in real-world scenarios, including intramodal noise, intermodal semantic inconsistencies, and sample-level uncertainty. These uncertainties may accumulate during feature aggregation and sentiment prediction, leading to model decision bias and undermining the stability and reliability of prediction outcomes.

Figure 1: An example of uncertainty factors in multimodal sentiment analysis. The visual modality produces misleading negative evidence, whereas the textual modality conveys positive sentiment, illustrating the impact of intermodal semantic inconsistency on sentiment prediction.

Based on the above analysis, this paper proposes RUAL, an uncertainty-aware learning framework for robust multimodal sentiment analysis. By introducing uncertainty modeling into feature aggregation, cross-modal alignment, and training optimization, RUAL constructs an end-to-end robust learning process to improve the stability and reliability of multimodal sentiment prediction.

The main contributions of this paper are summarized as follows:

1. We propose RUAL, an uncertainty-aware robust learning framework for multimodal sentiment analysis, which integrates uncertainty modeling into feature aggregation, cross-modal interaction, and training optimization to mitigate modality noise, semantic inconsistency, and incomplete information.

2. We design GMHA and UACCL to jointly enhance intra-modal feature aggregation and cross-modal semantic interaction. GMHA adaptively aggregates salient temporal information within each modality, while UACCL performs uncertainty-regulated residual interaction between the primary textual modality and auxiliary modalities.
3. We introduce an uncertainty-guided optimization strategy that combines uncertainty-weighted learning and self-distillation-based consistency regularization, enabling the model to dynamically adjust sample contributions and improve representation stability.

2 Related Work

2.1 Fusion-Based Multimodal Sentiment Analysis

A core challenge in multimodal sentiment analysis is how to effectively integrate emotional cues from heterogeneous modalities, such as text, speech, and vision. Early studies mainly relied on feature-level or decision-level fusion, where multimodal information was combined through concatenation, weighted aggregation, or late decision integration. For example, Poria et al. [13] adopted convolutional neural networks and multi-kernel learning for multimodal sentiment fusion, while Kampman et al. [16] compared audio, visual, and textual fusion strategies in end-to-end prediction tasks. Although these methods are easy to implement, they usually operate at a shallow feature or decision level and have limited ability to model complex semantic interactions across modalities.

With the development of deep learning, related studies have gradually shifted toward latent representation learning and cross-modal interaction modeling. Liu et al. [15] proposed low-rank multimodal fusion to reduce the computational complexity of high-order multimodal interactions. MulT [19] employs cross-modal Transformers to model temporal dependencies among unaligned multimodal sequences, while MISA [20] learns modality-invariant and modality-specific representations to capture both shared and private semantic information. Recent methods such as CubeMLP [9], ConFEDE [14], and MSG-MBA [12] further improve cross-modal semantic modeling through MLP-based interaction, contrastive feature decomposition, and hybrid-modal attention. However, most fusion-based methods mainly focus on designing stronger interaction structures and usually assume that different modalities are reliable and semantically consistent, which limits their robustness under noisy, incomplete, or conflicting multimodal inputs.

2.2 Missing-Modality Learning and Robust Multimodal Learning

In real-world multimodal applications, models may encounter incomplete, degraded, or weakly aligned inputs caused by sensor failures, environmental changes, synchronization errors, or insufficient preprocessing. To handle incomplete multimodal inputs, existing studies have explored modality completion, representation recovery, and self-supervised learning. GCNet [21] uses graph completion to model relational structures among modalities in incomplete multimodal conversations, while Self-MM [22] learns modality-specific representations through self-supervised multi-task learning.

Beyond explicit modality missingness, multimodal data may also suffer from noise interference, modality quality imbalance, and unreliable cross-modal correspondence. Modality dominance and gradient conflict can cause optimization imbalance during multimodal training [23], and noisy correspondence may weaken cross-modal semantic alignment [24]. To improve robustness under noisy or incomplete conditions, DiCMoR [18] performs distribution-consistent modality recovery, while Wu et al. [25] proposed distribution-based feature recovery and fusion for noisy image-text sentiment analysis.

Recent prompt-based methods provide another direction for incomplete multimodal sentiment and emotion understanding. Guo et al. [26] introduced generative prompts, missing-signal prompts, and missing-type prompts to recover missing modality cues and enhance intra- and inter-modal interactions.

Wu et al. [27] further proposed a mixture-of-prompt-experts framework for few-shot multimodal semantic understanding. These studies indicate that prompt learning is a useful complementary direction for robust multimodal sentiment analysis; however, most of them do not explicitly integrate reliability estimation into fusion and optimization as RUAL does.

2.3 Uncertainty-Aware Fusion and Confidence-Guided Optimization

Uncertainty modeling provides an important perspective for improving the reliability of multimodal learning. It may arise from data-level factors, such as noise, missing values, and modality heterogeneity, or from model-level factors, such as unstable predictions caused by insufficient cross-modal modeling. Existing studies estimate predictive uncertainty through probabilistic modeling, Bayesian inference, random sampling, ensemble learning, or information-theoretic measures [28–31]. In multimodal scenarios, uncertainty is often used to characterize modality quality and sample reliability.

Recent uncertainty-aware fusion methods improve robustness by dynamically adjusting modality contributions. Gao et al. [32] modeled unimodal aleatoric uncertainty for robust multimodal fusion, Zhang et al. [17] proposed a dynamic fusion framework for low-quality multimodal data, and Gao et al. [33] introduced uncertainty-aware routing for multimodal sentiment analysis. Xie et al. [34] further proposed a trustworthy multimodal sentiment ordinal network with unimodal uncertainty estimation and Bayesian fusion. Compared with these distribution-level or fusion-stage uncertainty methods, RUAL adopts a lightweight attention-entropy-based reliability proxy and couples it with residual regulation and sample-level confidence-guided optimization.

Beyond fusion, confidence-guided optimization and self-distillation have also been used to improve training stability. Zhang and Sabuncu [35] interpreted self-distillation as instance-specific label smoothing, while Jin et al. [36] introduced uncertainty-aware distillation according to predictive uncertainty. In multimodal learning, Li et al. [37] proposed Correlation-Decoupled Knowledge Distillation for incomplete modalities, and Luo et al. [38] introduced confidence-aware self-distillation for multimodal sentiment analysis with incomplete modalities. However, most existing methods handle modality reliability or sample confidence at a single stage, whereas RUAL uniformly incorporates uncertainty-aware mechanisms into intra-modal aggregation, cross-modal fusion, and confidence-guided optimization.

3 Method

3.1 Overall Model Framework

This presents the proposed Robust Uncertainty-Aware Learning (RUAL) framework, whose overall architecture is illustrated in Fig. 2. The framework is designed to construct multimodal sentiment representations with enhanced semantic consistency and robustness through intra-modal feature aggregation, cross-modal interaction, and uncertainty-guided optimization.

Specifically, RUAL consists of three major components: the Gathered Multi-Head Attention Pooling module (GMHA), the Uncertainty-Aware Cross-Modal Coupled Layer (UACCL), and the uncertainty-driven robust optimization mechanism. First, the model extracts and projects textual, audio, and visual features into a unified latent space, yielding temporal representations with aligned dimensions. Next, the GMHA module models temporal attention dependencies within each modality to capture salient frames and core semantic information. Then, the UACCL module performs uncertainty-aware cross-modal fusion centered on the textual modality through an uncertainty-regulated residual interaction mechanism. Finally,

the optimization stage introduces uncertainty-weighted learning (UWL) and uncertainty-guided self-distillation (U-SD) to dynamically adjust the contribution of each training sample, thereby improving model robustness and generalization.

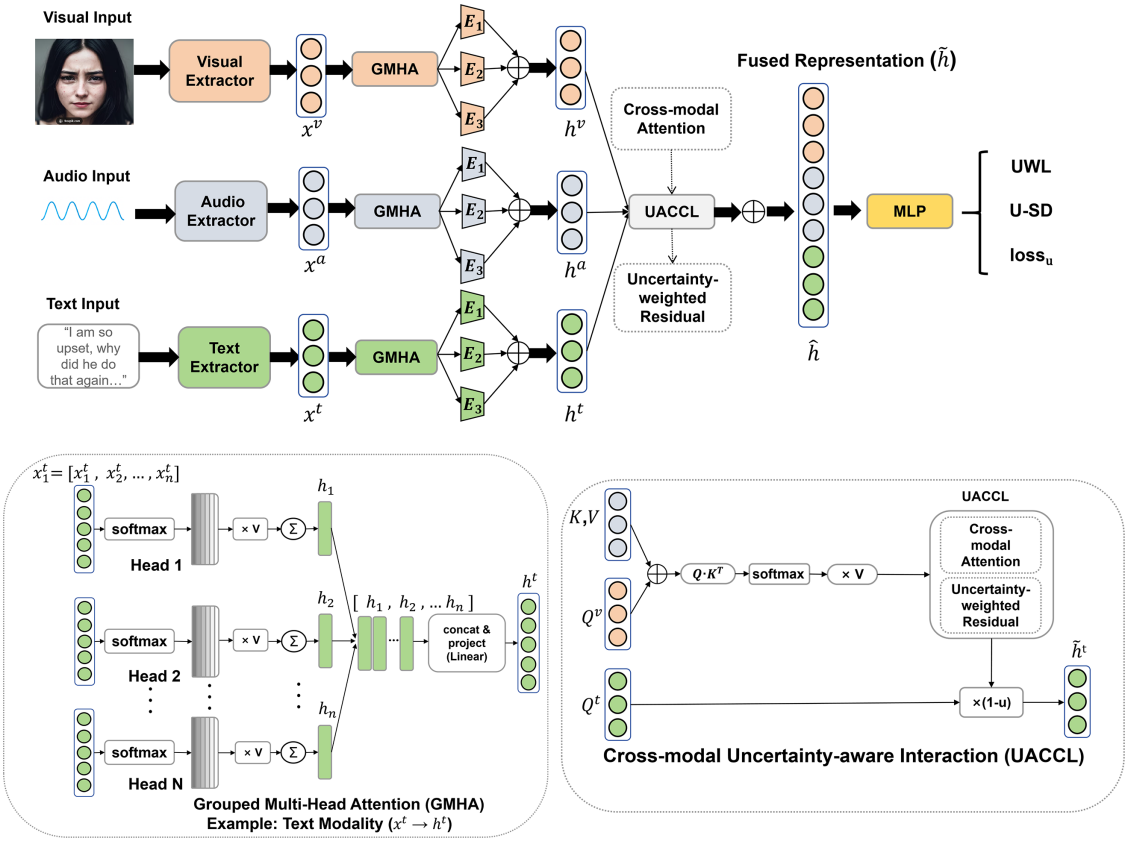

Figure 2: Overall architecture of the proposed RUAL model.

Suppose that each input sample consists of three modalities, namely text, audio, and visual features:

$$X = \{X^t, X^a, X^v\}, \quad X^m \in \mathbb{R}^{T_m \times d_m}, \quad m \in \{t, a, v\} \tag{1}$$

The model generates the final sentiment prediction value through multi-layer nonlinear mapping and attention interactions:

$$\hat{y} = f_{\text{RUAL}}(X; \Theta) \tag{2}$$

Here, Θ denotes the set of model parameters. Through collaborative modeling across three dimensions—features, alignment, and optimization—RUAL maintains stable performance in environments with multimodal inconsistencies and noisy interference, laying the foundation for the subsequent design of specific modules.

3.2 Intra-Modal Aggregation Module

To fully capture key semantic features within each modality while suppressing noise interference, RUAL introduces a Gathered Multi-Head Attention Pooling (GMHA) module in each modality channel.

GMHA performs attention-based aggregation along the temporal dimension and adaptively selects representative subspace information from the input sequence, yielding stable and semantically condensed modal representations. Given the sequence features:

$$X^m = [x_1^m, x_2^m, \ldots, x_{T_m}^m], \quad m \in \{t, a, v\} \tag{3}$$

mapping the sequence to a compact global representation $h^m \in \mathbb{R}^{1\times d_m}$.

First, perform a linear transformation on the input features to obtain the Key and Value matrices:

$$K^m = W_k X^m, \quad V^m = W_v X^m \tag{4}$$

where $W_k, W_v \in \mathbb{R}^{d_m \times d_m}$ are learnable weight matrices.

GMHA divides the feature dimension into H attention heads, each with dimension $d_h = \frac{d_m}{H}$. For the i-th attention head, a learnable query vector $q_i \in \mathbb{R}^{d_h}$ is introduced, and the attention weight is computed as:

$$\alpha_i = \text{softmax}\left(\frac{q_i K_i^{mT}}{\sqrt{d_h}}\right) \tag{5}$$

The aggregated output of each head is then obtained through weighted summation:

$$h_i^m = \alpha_i V_i^m \tag{6}$$

Finally, concatenate the outputs of all heads and apply a linear mapping to obtain the modal aggregation result:

$$h^m = W_o [h_1^m, h_2^m, \ldots, h_H^m] \tag{7}$$

where $W_o \in \mathbb{R}^{d_m \times d_m}$ is the output transformation matrix.

In GMHA, the temporal attention distribution of the i-th attention head for modality m is $\alpha_i^m \in \mathbb{R}^{T_m}$, satisfying $\sum_{t=1}^{T_m} \alpha_{i,t}^m = 1$. We employ attention entropy to characterize the concentration of this head on sequential information:

$$H_i^m = -\sum_{t=1}^{T_m} \alpha_{i,t}^m \log\left(\alpha_{i,t}^m + \epsilon\right) \tag{8}$$

Normalized using $\log T_m$ to obtain uncertainty within the range $[0, 1]$:

$$u^m = \frac{1}{H}\sum_{i=1}^{H} \frac{H_i^m}{\log T_m} \tag{9}$$

where ϵ is a numerical stability term. A larger u^m indicates that the attention distribution is more uniform and the representation of the modality is more uncertain.

To facilitate subsequent cross-modal alignment and optimization, we further take the mean uncertainty across the three modalities as the sample-level uncertainty:

$$u = \frac{1}{3}(u^t + u^a + u^v), \quad c = 1 - u \tag{10}$$

We use attention entropy as a lightweight proxy for semantic concentration rather than a direct proof of semantic unreliability. In sentiment-related temporal sequences, a highly concentrated attention

distribution usually indicates that the model can identify a small number of salient frames or tokens, whereas a nearly uniform distribution suggests that no dominant sentiment-related evidence is clearly selected. Therefore, a larger entropy value is interpreted as weaker semantic focus and is used as an uncertainty indicator in subsequent alignment and optimization. This design does not assume that uniform attention is always unreliable in all cases; instead, it provides a sample-adaptive reliability cue that is jointly used with cross-modal residual regulation and confidence-guided training. After GMHA obtains the modal-level representation h^m, a lightweight enhancement layer E is used to supplement contextual information and alleviate potential information loss caused by aggregation. In our implementation, E is a shallow Transformer encoder with 3 encoder layers, 4 attention heads, a hidden dimension of 72, and a dropout rate of 0.1. The enhanced representation is still denoted as h^m and is fed into the subsequent UACCL module for cross-modal uncertainty-aware alignment.

3.3 Cross-Modal Residual Alignment Module

After completing intra-modal aggregation, the model obtains high-level semantic representations $\mathbf{h}^t$, $\mathbf{h}^a$, and $\mathbf{h}^v$ for each modality. However, information from a single modality often suffers from semantic incompleteness or expression bias. To address this, we propose the Uncertainty-aware Cross-modal Coupled Layer (UACCL). This module maintains the semantic dominance of the primary modality (text) while incorporating supplementary information from the audio and visual modalities, thereby enabling robust cross-modal dynamic alignment.

UACCL uses the text modality as the query ($\mathbf{Q}^t$), while concatenating the encoded representations of the audio and visual modalities as the key and value (K, V). First, a cross-modal attention mechanism computes interaction weights between the primary modality and the other modalities:

$$\mathrm{Attn}(\mathbf{Q}^t, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}^t\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V} \tag{11}$$

The cross-modal attention output $\tilde{\mathbf{h}}^t$ is used to supplement the semantic information of the primary modality. Through this process, UACCL captures dynamic semantic associations across modalities and enables the model to extract complementary sentiment-related cues from both acoustic and visual modalities.

To further improve robustness, we introduce an uncertainty-weighted residual regulation mechanism. Specifically, during the intra-modal aggregation stage, the model estimates a sample-level uncertainty score $u \in [0,1]$ based on the attention distributions, which reflects the confidence of the current multimodal representation. During cross-modal fusion, this uncertainty score is used to adaptively control the strength of residual information injection. The final fused representation is formulated as

$$\hat{\mathbf{h}}^t = \mathbf{h}^t + \alpha_{\mathrm{res}}\gamma \odot \tilde{\mathbf{h}}^t, \qquad \gamma = \sigma(\alpha)\cdot(1-u), \tag{12}$$

where $\sigma(\cdot)$ denotes the Sigmoid function, which constrains the learnable gate to the range $(0,1)$, and α_{res} is the residual scaling factor. The learnable term $\sigma(\alpha)$ captures the global tendency of the model to introduce cross-modal residual information, while the sample-level confidence term $(1-u)$ adaptively modulates this residual injection for each input sample. When the sample uncertainty u is high, the confidence term becomes small and suppresses unreliable cross-modal residuals. In contrast, when u is low, more complementary acoustic and visual information is introduced to enhance the textual representation. In our implementation, α_{res} is set to 1.0.

3.4 Uncertainty-Weighted Optimization with Self-Distillation

After intra-modal aggregation and cross-modal residual alignment, the model obtains the fused multimodal representation $\hat{\mathbf{h}}$ and generates the final sentiment prediction $\hat{y}$ through a Multi-Layer Perceptron (MLP). The prediction MLP is implemented as a multi-layer feed-forward network with ReLU activation and dropout. It maps the concatenated multimodal representation to hidden layers with dimensions d, 128, 256, and 128, and finally outputs a task-specific sentiment prediction. However, in real-world multimodal scenarios, significant uncertainty variations often exist across different samples. Noise interference in acoustic signals, occlusions in visual frames, and ambiguous semantic expressions in text may all introduce prediction bias. When optimized with uniform sample weights, highly noisy samples tend to dominate gradient propagation, thereby weakening convergence stability and reducing generalization performance. To address this issue, RUAL introduces two complementary uncertainty-aware optimization mechanisms: Uncertainty-Weighted Loss (UWL) and Uncertainty-guided Self-Distillation (U-SD), enabling more robust and confidence-aware learning.

Specifically, during the modal aggregation phase, the model estimates uncertainty $u \in [0,1]$ for each sample via attention distributions and incorporates this as a dynamic weighting factor in loss computation. For each sample i, UWL applies an inverse confidence weighting to the L_1 loss:

$$L_{\text{UWL}} = \frac{1}{N}\sum_{i=1}^{N}(1-u_i)\,|\hat{y}_i - y_i| \tag{13}$$

When a sample has high confidence, i.e., a small u_i, it is assigned a larger optimization weight; conversely, high-uncertainty samples receive smaller weights. This mechanism reduces the influence of noisy samples and encourages the model to learn more from reliable samples.

However, simply reducing the weight of high-uncertainty samples may lead to insufficient learning in low-confidence regions, thereby causing representation drift. To address this, RUAL further introduces uncertainty-guided self-distillation (U-SD) to maintain prediction consistency on high-confidence samples. Instead of directly detaching the prediction from the same forward pass, U-SD adopts a teacher-student consistency mechanism. Specifically, the student prediction $\hat{y}_i^S$ is obtained from the current forward pass, while the teacher prediction $\hat{y}_i^T$ is generated by an additional no-gradient stochastic forward pass. The teacher prediction is detached from gradient propagation and serves as a soft target for high-confidence samples:

$$L_{\text{U-SD}} = \frac{\sum_{i=1}^{N}\mathbf{1}(u_i^T < \tau)\left(\hat{y}_i^S - \text{sg}(\hat{y}_i^T)\right)^2}{\sum_{i=1}^{N}\mathbf{1}(u_i^T < \tau) + \epsilon} \tag{14}$$

where τ represents the confidence threshold, u_i^T denotes the uncertainty estimated from the teacher forward pass, $\text{sg}(\cdot)$ denotes the stop-gradient operation, and ϵ is a numerical stability term. Since the teacher prediction is produced by an independent no-gradient stochastic forward pass rather than by simply detaching the student output from the same computation graph, the self-distillation objective avoids collapsing into a trivial identity constraint. By imposing consistency constraints only on high-confidence samples, U-SD stabilizes predictions while reducing the propagation of unreliable pseudo-targets.

Ultimately, the overall optimization objective of RUAL is defined as:

$$L_{\text{total}} = L_{\text{aux}} + \lambda_{\text{sup}} L_{\text{UWL}} + \lambda_{\text{sd}} L_{\text{U-SD}} \tag{15}$$

where L_{aux} denotes the auxiliary regularization loss returned by the fusion module, and λ_{sup} and λ_{sd} are the weights for the uncertainty-weighted supervised loss and the uncertainty-guided self-distillation loss,

respectively. In our implementation, $\lambda_{\text{sup}} = 10.0$, $\lambda_{\text{sd}} = 0.2$, and the confidence threshold in U-SD is set to $\tau = 0.4$. Through joint optimization, RUAL emphasizes reliable samples and suppresses uncertain ones, improving robustness and generalization under noisy and incomplete multimodal conditions.

4 Experiments

4.1 Experimental Setup

To validate the proposed method's robustness and generalization capabilities under multimodal incompleteness and real-world noise conditions, we conducted experiments on three widely used multimodal sentiment analysis datasets: CMU-MOSI [39], CMU-MOSEI [40], and MVSA-Single [41]. CMU-MOSI comprises 2199 short video comments, partitioned into training (1284), validation (229), and test (686) sets. CMU-MOSEI comprises 22,856 YouTube video comments, with 16,326 for training, 1871 for validation, and 4659 for testing. Both datasets provide text, visual, and acoustic multimodal information, employing continuous sentiment labels within the range [−3, 3]. The MVSA-Single dataset comprises image-text samples from authentic social media, featuring inherent noise and modal quality inconsistencies characteristic of real-world scenarios. It is used to evaluate model performance under complex noise conditions.

Experimental tasks focus on multimodal sentiment analysis. For CMU-MOSI and CMU-MOSEI, we evaluate continuous sentiment prediction using Binary Classification Accuracy (Acc2), Seven-Class Classification Accuracy (Acc7), and Weighted F1 Score. For MVSA-Single, predictions are performed directly in its original image-text setting without additional noise injection.

For reproducibility, we specify the feature settings used in our experiments. For CMU-MOSI and CMU-MOSEI, we use the aligned multimodal features commonly adopted in MSA studies. Textual features are extracted with BERT-base-uncased, with an original dimension of 768. Acoustic and visual features are represented by COVAREP and FACET/OpenFace descriptors, with original dimensions of 74 and 35, respectively. All modality sequences are padded or truncated to a length of 50 and projected into a shared hidden dimension of 72 before being fed into GMHA. For MVSA-Single, the text branch follows the same BERT-based encoding strategy, and the image branch is projected into the same hidden dimension.

For the proposed RUAL model, fixed hyperparameter configurations were employed to ensure reproducible training and evaluation. Specific parameter settings are detailed in Table 1, covering training-related parameters (e.g., batch size and learning rate) and model architecture parameters (e.g., multi-head attention heads and Transformer encoder layers). The proposed RUAL model is implemented within the PyTorch framework and trained using the Adam optimizer. A learning rate scheduling strategy and an early stopping mechanism are incorporated during training to enhance stability. Experiments were conducted on an Ubuntu environment using PyTorch 2.1 and CUDA 11.8.

4.2 Baseline Models

To comprehensively evaluate the performance of RUAL, we compare it with representative multimodal sentiment analysis baselines, including MulT [19], MISA [20], CubeMLP [9], MHMF-BERT [42], GCNet [21], MSG-MBA [12], ConFEDE [14], DiCMoR [18], Self-MM [22], MMIM [43], and MAG-BERT [44]. These methods cover major technical directions in multimodal sentiment analysis, including cross-modal attention, representation alignment, graph-based completion, mutual information maximization, and pretrained transformer-based fusion. For the MVSA-Single image-text sentiment dataset, we further compare RUAL with Late Fusion [16], MMTM [45], TMC [46], MVCN [47], and QMF [17].

Table 1: Best-performing hyperparameter settings.

Hyperparameter	Value
Batch size	16
Number of runs	3
Learning rate (BERT)	3×10^{-5}
Learning rate (other modules)	1×10^{-4}
Dropout rate	0.1
Hidden dimension	72
Number of GMHA heads	4
Number of transformer encoder layers	3
Feature alignment length	50
UWL supervised weight λ_{sup}	10.0
U-SD weight λ_{sd}	0.2
Confidence threshold τ	0.4
UACCL residual scale α_{res}	1.0
Weight decay	1×10^{-4}
Early stopping patience	10
Learning rate scheduler factor	0.5
Learning rate scheduler patience	5

For transparency, the baseline results in Tables 2 and 3 are collected from the corresponding original papers or recent publicly reported benchmark results under commonly used dataset splits and evaluation metrics. Since these methods may differ in feature inputs, pretrained backbones, and training details, the comparisons should be regarded as benchmark-level comparisons rather than fully unified re-implementations. RUAL is trained and evaluated under the setting described in Section 4.1, and the reported results are averaged over three runs.

Table 2: Comparison results on the CMU-MOSI and CMU-MOSEI datasets.

Method	CMU-MOSI			CMU-MOSEI		
	Acc7	Acc2	F1	Acc7	Acc2	F1
MulT	–	81.5	80.6	–	–	–
MISA	42.5	80.5	80.5	52.2	82.5	82.5
CubeMLP	45.5	85.6	85.5	54.9	85.1	84.5
MHMF-BERT	–	85.3	85.3	–	85.6	85.6
GCNet	44.9	85.1	85.1	51.5	85.2	85.2
MSG-MBA	45.3	85.7	85.6	52.8	85.4	85.4
ConFEDE	42.3	85.5	85.5	54.9	85.8	85.8
DiCMoR	45.3	85.6	85.6	53.4	85.1	85.1
Self-MM	45.7	82.3	82.7	53.5	82.5	82.5

(Continued)

Table 2 (continued)

Method	CMU-MOSI			CMU-MOSEI		
	Acc7	Acc2	F1	Acc7	Acc2	F1
MMIM	46.2	82.9	83.0	54.0	82.3	82.4
MAG-BERT	–	82.5	82.6	–	83.8	83.7
RUAL (ours)	**46.2**	**86.1**	**86.3**	**54.9**	**85.8**	**85.9**

Note: The results of RUAL are highlighted in bold. Baseline results are collected from the corresponding original papers or recent publicly reported benchmark results under commonly used CMU-MOSI and CMU-MOSEI splits and metrics. RUAL is trained and evaluated under the setting in Section 4.1, and the reported results are averaged over three runs. "–" denotes unavailable results.

Table 3: Comparison results on the MVSA-Single dataset.

Method	Acc	F1
Late Fusion	65.6	65.4
MMTM	75.2	75.0
TMC	76.1	74.6
QMF	78.1	77.2
MVCN	76.1	74.6
RUAL (ours)	**78.6**	**78.2**

Note: The results of RUAL are highlighted in bold. Baseline results are collected from the corresponding original papers or recent publicly reported benchmark results under the MVSA-Single setting. RUAL is trained and evaluated under the setting in Section 4.1, and the reported results are averaged over three runs.

To further clarify the methodological differences between RUAL and representative multimodal sentiment analysis methods, Table 4 summarizes their fusion strategies, reliability or uncertainty modeling, missing-modality handling, and robustness evaluation settings.

Table 4: Methodological comparison between RUAL and representative multimodal sentiment analysis methods.

Method	Fusion Strategy	Reliability Modeling	Missing-Modality Handling	Robustness Evaluation
MulT	Cross-modal attention	Not explicit	Not considered	Not emphasized
MISA	Shared-specific representation learning	Not explicit	Not considered	Not emphasized
Self-MM	Self-supervised multimodal learning	Implicit	Partially considered	Limited

(Continued)

Table 4 (continued)

Method	Fusion Strategy	Reliability Modeling	Missing-Modality Handling	Robustness Evaluation
GCNet	Graph-based completion	Not explicit	Explicit completion	Missing-modality setting
DiCMoR	Distribution-consistent recovery	Distribution consistency	Explicit recovery	Incomplete-modality setting
QMF	Quality-aware dynamic fusion	Quality estimation	Not mainly considered	Low-quality multimodal setting
Guo et al.	Prompt-based recovery	Not explicit	Missing-modality prompts	Incomplete-modality setting
Xie et al.	Bayesian uncertainty-aware fusion	Unimodal uncertainty	Partially considered	Robustness-oriented MSA setting
RUAL	GMHA aggregation + UACCL fusion	Attention-entropy uncertainty + sample confidence	Zero-tensor simulation + adaptive residual regulation	Missing-modality + multi-level Gaussian-noise settings

Note: This table provides a methodological comparison rather than a direct performance comparison. Some recent studies are included to clarify the technical positioning of RUAL in relation to prompt-based missing-modality learning and uncertainty-aware fusion.

4.3 Comparative Experiments

Tables 2 and 3 present performance comparisons between RUAL and multiple baseline models across three datasets.

The results in Tables 2 and 3 show that RUAL achieves competitive or superior performance across all three datasets. On CMU-MOSI, RUAL obtains 86.1% Acc2 and 86.3% F1, outperforming most baseline methods and demonstrating the effectiveness of uncertainty-guided cross-modal fusion. On CMU-MOSEI, RUAL also maintains stable and competitive performance, with 85.8% Acc2 and 85.9% F1. Moreover, on the MVSA-Single image-text dataset, RUAL achieves the highest Acc and F1 scores, indicating that the proposed uncertainty-aware mechanism can generalize from video-based multimodal sentiment analysis to image-text sentiment scenarios. These results validate the effectiveness of RUAL in robust multimodal representation learning and cross-modal fusion.

4.4 Ablation Study

To verify the contribution of each module, we conducted ablation experiments on the CMU-MOSI and CMU-MOSEI datasets. Since MVSA-Single is mainly used to evaluate generalization under real-world noisy conditions, module-level ablation was not conducted on this dataset. The experiments progressively remove GMHA, UACCL, UWL, and U-SD, and the results are reported in Tables 5 and 6.

Table 5: Ablation study on the CMU-MOSEI dataset.

Description	Acc7 ↑	Acc2 ↑	F1 ↑
RUAL	**54.92 ± 0.22**	**85.82 ± 0.13**	**85.93 ± 0.14**
w/o GMHA	49.60 ± 0.56	83.98 ± 0.38	82.96 ± 0.40
w/o UACCL	51.35 ± 0.43	84.32 ± 0.31	83.25 ± 0.34
w/o UWL	53.03 ± 0.34	85.50 ± 0.18	84.25 ± 0.30
w/o U-SD	52.85 ± 0.29	84.66 ± 0.24	84.46 ± 0.28

Note: "w/o" denotes the variant model without the corresponding component. Values are reported as mean ± standard deviation over three independent runs. Bold values indicate the best mean performance.

Table 6: Ablation study on the CMU-MOSI dataset.

Description	Acc7 ↑	Acc2 ↑	F1 ↑
RUAL	**46.15 ± 0.28**	**86.13 ± 0.16**	**86.29 ± 0.18**
w/o GMHA	43.94 ± 0.52	83.51 ± 0.35	83.19 ± 0.37
w/o UACCL	44.96 ± 0.45	85.27 ± 0.28	84.32 ± 0.33
w/o UWL	45.84 ± 0.31	84.89 ± 0.26	84.86 ± 0.27
w/o U-SD	45.72 ± 0.36	85.65 ± 0.22	85.58 ± 0.24

Note: "w/o" denotes the variant model without the corresponding component. Values are reported as mean ± standard deviation over three independent runs. Bold values indicate the best mean performance.

The ablation results show that each component contributes to the final performance, with GMHA producing the most obvious impact. Removing GMHA causes the largest performance drop on both CMU-MOSEI and CMU-MOSI, indicating that intra-modal temporal aggregation is essential for selecting sentiment-related salient information and suppressing noisy sequence features before cross-modal interaction.

Removing UACCL also leads to clear degradation, which confirms the importance of uncertainty-regulated cross-modal residual interaction. Without UACCL, the model cannot dynamically control the injection of auxiliary acoustic and visual information, making the textual representation more vulnerable to noisy or semantically inconsistent cues. UWL and U-SD show complementary effects: UWL adjusts the supervised loss contribution according to sample confidence, whereas U-SD improves prediction consistency through high-confidence teacher-student regularization. The performance drops caused by removing either module indicate that both sample-level weighting and confidence-guided consistency learning are beneficial for robust multimodal sentiment analysis. Overall, these results demonstrate that RUAL benefits from the cooperation of feature-level aggregation, fusion-level uncertainty regulation, and optimization-level confidence guidance. All ablation results are reported as mean ± standard deviation over three independent runs.

4.5 Robustness Analysis under Missing and Noisy Modalities

To further verify the robustness of RUAL under incomplete inputs and noise interference, we conducted two types of inference-stage robustness experiments on the CMU-MOSI and CMU-MOSEI datasets: auxiliary modality missing experiments and Gaussian noise perturbation experiments. During testing, the trained model parameters were kept unchanged, and only the input features of the auxiliary modalities were modified.

For the missing-modality setting, the unavailable modality was simulated by replacing its feature tensor with a zero tensor of the same shape, while the remaining modalities were kept unchanged. Specifically, we considered three cases: missing audio, missing vision, and missing both audio and vision. The textual modality was retained in all missing-modality settings because it serves as the primary semantic modality in the proposed framework.

For the Gaussian-noise setting, noise was injected into the auxiliary modality features during inference. Given an auxiliary modality feature X^m, where $m \in \{a, v\}$, the perturbed feature was generated as

$$\tilde{X}^m = X^m + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{16}$$

where σ denotes the noise intensity. Table 7 reports the representative noise setting of $\sigma = 0.5$. To further examine the performance trend under progressively stronger perturbations, we additionally evaluated RUAL under multiple noise intensities, i.e., $\sigma \in \{0.0, 0.1, 0.3, 0.5, 0.7, 1.0\}$. For each noise intensity, three perturbation settings were tested, including audio noise, vision noise, and audio+vision noise. The robustness curves were averaged over three independent noise perturbation runs.

Table 7: Robustness evaluation under Gaussian noise perturbation on the CMU-MOSI and CMU-MOSEI datasets ($\sigma = 0.5$).

Dataset	Setting	Acc7 ↑	Acc2 ↑	F1 ↑
CMU-MOSI	Full	46.15	86.13	86.29
	Audio Noise	46.48	85.52	85.69
	Vision Noise	46.15	85.67	85.83
	Audio + Vision Noise	46.62	85.21	85.32
CMU-MOSEI	Full	54.92	85.82	85.93
	Audio Noise	54.59	85.77	85.91
	Vision Noise	54.49	85.54	85.61
	Audio + Vision Noise	54.40	84.43	85.54

As shown in Table 8, RUAL maintains stable performance under different missing-modality settings. On both CMU-MOSI and CMU-MOSEI, the Acc2 and F1 scores remain close to the full-modality results when audio, vision, or both auxiliary modalities are missing, indicating that RUAL can reduce its dependence on unavailable auxiliary modalities.

Table 8: Robustness evaluation under missing auxiliary modalities on the CMU-MOSI and CMU-MOSEI datasets.

Dataset	Setting	Acc7 ↑	Acc2 ↑	F1 ↑
CMU-MOSI	Full	46.15	86.13	86.29
	Missing Audio	46.67	86.13	86.24
	Missing Vision	46.23	85.84	85.04
	Missing Audio + Vision	46.36	86.53	85.74
CMU-MOSEI	Full	54.92	85.82	85.93
	Missing Audio	55.10	85.77	85.84
	Missing Vision	55.46	85.66	85.74
	Missing Audio + Vision	55.36	84.71	85.79

Table 7 further shows that RUAL is robust to Gaussian noise perturbation at $\sigma = 0.5$. Although performance slightly decreases when auxiliary modalities are disturbed, the overall degradation remains limited, especially under single-modality noise. Fig. 3 reports the F1 trends under multiple noise intensities. As σ increases from 0.0 to 1.0, RUAL shows only moderate fluctuations rather than severe performance collapse, suggesting that the proposed uncertainty-aware fusion and optimization mechanisms can suppress the negative influence of noisy auxiliary modalities.

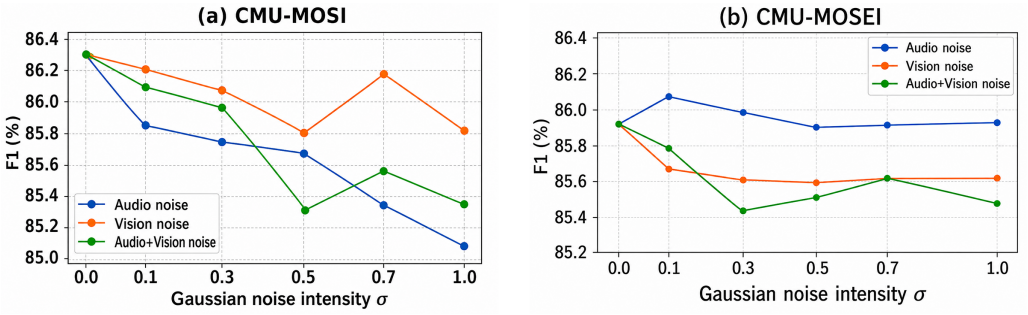

Figure 3: Robustness curves of RUAL under different Gaussian noise intensities. (**a**) F1 trends on CMU-MOSI. (**b**) F1 trends on CMU-MOSEI.

4.6 Modality Weight Variation Analysis

To further examine the adaptive behavior of UACCL, we analyze the average modality weights of text, audio, and vision under Clean, Noisy, and Missing scenarios on CMU-MOSEI, as shown in Fig. 4. The weights are obtained by averaging the cross-modal gating responses over test samples. Text consistently receives the highest weight and increases from 0.38 to 0.50 as input quality degrades, indicating that RUAL relies more on stable linguistic information under noisy or missing conditions. In contrast, the audio weight decreases from 0.34 to 0.00 when the audio modality becomes invalid, while the visual weight changes from 0.28 to 0.22 and then to 0.30, suggesting that the model can reduce unreliable modality contributions and perform cross-modal compensation. These results further support the adaptive robustness of the proposed uncertainty-aware fusion mechanism.

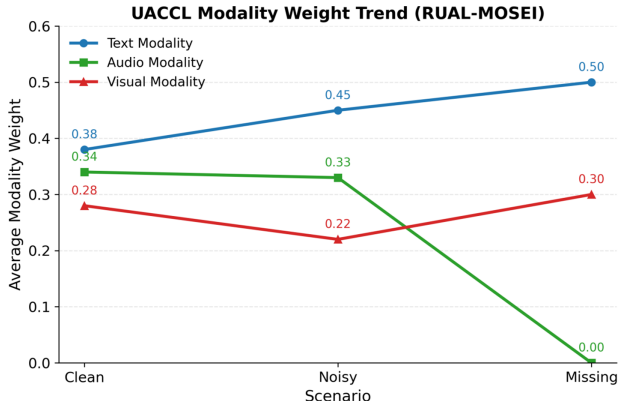

Figure 4: Average modality weight trends of text, audio, and visual modalities under Clean, Noisy, and Missing scenarios on CMU-MOSEI.

5 Conclusions and Future Work

This paper proposes RUAL, a robust uncertainty-aware learning framework for multimodal sentiment analysis under noisy and incomplete conditions. By integrating uncertainty modeling into intra-modal aggregation, cross-modal residual fusion, and training optimization, RUAL adaptively characterizes modality reliability and sample confidence, thereby reducing the negative influence of unreliable modalities and low-confidence samples. Experimental results on CMU-MOSI, CMU-MOSEI, and MVSA-Single demonstrate that RUAL achieves competitive performance and maintains stable prediction results under missing-modality and Gaussian-noise settings.

Although RUAL shows promising robustness, its performance may still be limited under highly ambiguous sentiment semantics or extreme sample distributions. Future work will explore more fine-grained uncertainty modeling, cross-domain generalization, and interpretable robust multimodal sentiment learning.

Acknowledgement: Not applicable.

Funding Statement: This work was supported by the National Natural Science Foundation of China under Grant No. 61762059. The funder's website is available at https://www.nsfc.gov.cn/.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Weijun Gao and Ziyang Zhang; methodology and experiments: Weijun Gao; analysis and interpretation of results: Weijun Gao, Ziyang Zhang and Maotang Su; draft manuscript preparation: Weijun Gao; manuscript review and editing: Ziyang Zhang and Maotang Su. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets used in this study are publicly available from their original public sources. CMU-MOSI and CMU-MOSEI are publicly available through the CMU Multimodal Data SDK, and MVSA-Single is available from its original public dataset source.

Ethics Approval: Not applicable.

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

References

1. Jiang X, Xu X, Chen Z, Zhang J, Song J, Shen F, et al. DHHN: dual hierarchical hybrid network for weakly-supervised audio-visual video parsing. In: Proceedings of the 30th ACM International Conference on Multimedia; 2022 Oct 10–14; Lisboa, Portugal. p. 719–27.
2. Chen M, Gao J, Xu C. Uncertainty-aware dual-evidential learning for weakly-supervised temporal action localization. IEEE Trans Pattern Anal Mach Intell. 2023;45(12):15896–911.
3. Jiang X, Xu X, Zhang J, Shen F, Cao Z, Shen HT. Semi-supervised video paragraph grounding with contrastive encoder. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 19–24; New Orleans, LA, USA. p. 2466–75.
4. Xu X, Lin K, Yang Y, Hanjalic A, Shen HT. Joint feature synthesis and embedding: adversarial cross-modal retrieval revisited. IEEE Trans Pattern Anal Mach Intell. 2020;44(6):3030–47.
5. Jiang X, Xu X, Zhou Z, Yang Y, Shen F, Shen HT. Zero-shot video moment retrieval with angular reconstructive text embeddings. IEEE Trans Multim. 2024;26(2):9657–70. doi:10.1109/tmm.2024.3396272.
6. Mustafa B, Riquelme C, Puigcerver J, Jenatton R, Houlsby N. Multimodal contrastive learning with LIMoE: the language-image mixture of experts. Adv Neural Inf Process Syst. 2022;35:9564–76.
7. Mathur L, Mataric M, Morency LP. Expanding the role of affective phenomena in multimodal interaction research. In: Proceedings of the 25th International Conference on Multimodal Interaction; 2023 Oct 9–13; Paris, France. p. 253–60.

8. Kovacevic N, Holz C, Gross M, Wampfler R. On multimodal emotion recognition for human-chatbot interaction in the wild. In: Proceedings of the 26th International Conference on Multimodal Interaction; 2024 Nov 4–8; San José, CA, USA. p. 12–21.
9. Sun H, Wang H, Liu J, Chen YW, Lin L. CubeMLP: an MLP-based model for multimodal sentiment analysis and depression estimation. In: Proceedings of the 30th ACM International Conference on Multimedia; 2022 Oct 10–14; Lisbon, Portugal. p. 3722–9.
10. Li Y, Wang Y, Cui Z. Decoupled multimodal distilling for emotion recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023 Jun 17–23; Vancouver, BC, Canada. p. 6631–40.
11. Mai S, Zeng Y, Hu H. Multimodal information bottleneck: learning minimal sufficient unimodal and multimodal representations. IEEE Trans Multimedia. 2022;25:4121–34.
12. Lin R, Hu H. Dynamically shifting multimodal representations via hybrid-modal attention for multimodal sentiment analysis. IEEE Trans Multimedia. 2023;26:2740–55.
13. Poria S, Chaturvedi I, Cambria E, Hussain A. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In: Proceedings of the 2016 IEEE 16th International Conference on Data Mining; 2016 Dec 12–15; Barcelona, Spain. p. 439–48.
14. 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; 2023 Jul 9–14; Toronto, ON, Canada. p. 7617–30.
15. Liu Z, Shen Y, Lakshminarasimhan VB, Liang PP, Zadeh AB, Morency LP. Efficient low-rank multimodal fusion with modality-specific factors. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics; 2018 Jul 15–20; Melbourne, VIC, Australia. p. 2247–56.
16. Kampman O, Barezi EJ, Bertero D, Fung P. Investigating audio, video, and text fusion methods for end-to-end automatic personality prediction. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics; 2018 Jul 15–20; Melbourne, VIC, Australia. p. 606–11.
17. Zhang Q, Wu H, Zhang C, Hu Q, Fu H, Zhou JT, et al. Provable dynamic fusion for low-quality multimodal data. In: Proceedings of the 40th International Conference on Machine Learning; 2023 Jul 23–29; Honolulu, HI, USA. p. 41753–69.
18. Wang Y, Cui Z, Li Y. Distribution-consistent modal recovering for incomplete multimodal learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023 Oct 2–6; Paris, France. p. 22025–34.
19. Tsai YHH, Bai S, Liang PP, Kolter JZ, Morency LP, Salakhutdinov R. Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019 Jul 28–Aug 2; Florence, Italy. p. 6558–69.
20. 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; Virtual. p. 1122–31.
21. Lian Z, Chen L, Sun L, Liu B, Tao J. GCNet: graph completion network for incomplete multimodal learning in conversation. IEEE Trans Pattern Anal Mach Intell. 2023;45(7):8419–32.
22. Yu W, Xu H, Yuan Z, Wu J. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. Proc AAAI Conf Artif Intell. 2021;35(12):10790–7.
23. Wang W, Tran D, Feiszli M. What makes training multi-modal classification networks hard? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020 Jun 13–19; Seattle, WA, USA. p. 12695–705.
24. Huang Z, Niu G, Liu X, Ding W, Xiao X, Wu H, et al. Learning with noisy correspondence for cross-modal matching. Adv Neural Inf Process Syst. 2021;34:29406–19.
25. Wu D, Yang D, Zhou Y, Ma C. Robust multimodal sentiment analysis of image-text pairs by distribution-based feature recovery and fusion. In: Proceedings of the 32nd ACM International Conference on Multimedia; 2024 Oct 28–Nov 1; Melbourne, VIC, Australia. p. 5780–9.

26. 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; 2024 Aug 11–16; Bangkok, Thailand. p. 1726–36. doi:10.18653/v1/2024.acl-long.94.
27. Wu Z, Huang HY, Qu F, Wu Y. Mixture-of-prompt-experts for multi-modal semantic understanding. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation; 2024 May 20–25; Torino, Italy. p. 11381–93.
28. Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learning for computer vision? In: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017); 2017 Dec 4–9; Long Beach, CA, USA.
29. Gal Y, Ghahramani Z. Dropout as a Bayesian approximation. arXiv:1506.02157. 2015.
30. Wu M, Goodman N. Multimodal generative models for scalable weakly-supervised learning. In: Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018); 2018 Dec 3–8; Montreal, QC, Canada.
31. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. arXiv:1612.01474. 2017.
32. Gao Z, Jiang X, Xu X, Shen F, Li Y, Shen HT. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024 Jun 16–24; Seattle, WA, USA. p. 26876–85.
33. Gao Z, Hu D, Jiang X, Lu H, Shen HT, Xu X. Enhanced experts with uncertainty-aware routing for multimodal sentiment analysis. In: Proceedings of the 32nd ACM International Conference on Multimedia; 2024 Oct 28–Nov 1; Melbourne, VIC, Australia. p. 9650–9.
34. Xie Z, Yang Y, Wang J, Liu X, Li X. Trustworthy multimodal fusion for sentiment analysis in ordinal sentiment space. IEEE Trans Circuits Syst Video Technol. 2024;34(8):7657–70. doi:10.1109/TCSVT.2024.3376564.
35. Zhang Z, Sabuncu M. Self-distillation as instance-specific label smoothing. Adv Neural Inf Process Syst. 2020;33:2184–95.
36. Jin X, Lan C, Zeng W, Chen Z. Uncertainty-aware multi-shot knowledge distillation for image-based object re-identification. Proc AAAI Conf Artif Intell. 2020;34(7):11165–72. doi:10.1609/aaai.v34i07.6774.
37. Li M, Yang D, Zhao X, Wang S, Wang Y, Yang K, et al. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024 Jun 16–22; Seattle, WA, USA. p. 12458–68.
38. Luo Y, Wang S, Xu Z, Li Y, Tang F, Su J. Confidence-aware self-distillation for multimodal sentiment analysis with incomplete modalities. arXiv:2506.01490. 2025.
39. Zadeh A, Zellers R, Pincus E, Morency LP. Multimodal sentiment intensity analysis in videos: facial gestures and verbal messages. IEEE Intell Syst. 2016;31(6):82–8.
40. 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; 2018 Jul 15–20; Melbourne, VIC, Australia. p. 2236–46.
41. Niu T, Zhu S, Pang L, El Saddik A. Sentiment analysis on multi-view social data. In: Proceedings of the International Conference on Multimedia Modeling; 2016 Jan 4–6; Miami, FL, USA. p. 15–27.
42. Zou W, Ding J, Wang C. Utilizing BERT intermediate layers for multimodal sentiment analysis. In: Proceedings of the 2022 IEEE International Conference on Multimedia and Expo; 2022 Jul 18–22; Taipei, Taiwan. p. 1–6.
43. Han W, Chen H, Poria S. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing; 2021 Nov 7–11; Punta Cana, Dominican Republic. p. 9180–92.
44. Rahman W, Hasan MK, Lee S, Zadeh AB, Mao C, Morency LP, et al. Integrating multimodal information in large pretrained transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020 Jul 5–10; Online. p. 2359–69.

45. Joze HRV, Shaban A, Iuzzolino ML, Koishida K. MMTM: multimodal transfer module for CNN fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020 Jun 13–19; Seattle, WA, USA. p. 13289–99.
46. Han Z, Zhang C, Fu H, Zhou JT. Trusted multi-view classification with dynamic evidential fusion. IEEE Trans Pattern Anal Mach Intell. 2022;45(2):2551–66. doi:10.1109/TPAMI.2022.3171983.
47. Wei Y, Yuan S, Yang R, Shen L, Li Z, Wang L, et al. Tackling modality heterogeneity with multi-view calibration network for multimodal sentiment detection. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics; 2023 Jul 9–14; Toronto, ON, Canada. p. 5240–52.