Decoding Emotional Silences: Reliable Multimodal Sentiment Analysis with Bipolar Uncertainty

Yutao Wei*†, Hongzhu Fu*†, Yuxiang Li*, Yichen Xin*, Xovee Xu*, Fan Zhou*§and Ting Zhong*‡¶
*University of Electronic Science and Technology of China, Chengdu, China
‡Aiwen Tech (Chengdu), Chengdu, Sichuan, China
§Key Laboratory of Intelligent Digital Media Technology of Sichuan Province, Chengdu, Sichuan, China
†Equal Contribution ¶Corresponding: zhongting@uestc.edu.cn

2025 IEEE International Conference on Multimedia and Expo (ICME) | 979-8-3315-9495-4/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICME59968.2025.11209057

***Abstract*—Multimodal sentiment analysis is critical in many real-world applications like smart cities, healthcare, and human-computer interaction, where sentiment is conveyed through various modalities, including text, audio, and video. However, existing methods still face several critical challenges in mitigating the impact of random modality loss, particularly in preserving the reliability of emotional patterns. To address these limitations, we introduce UniMSA, a novel framework for multimodal sentiment analysis with missing modalities. It addresses the missing data issue by improving the bipolar emotional uncertainty learning. Our approach enhances the reliability of sentiment analysis by integrating both positive and negative emotional uncertainty estimations to recover emotional patterns in randomly missing modalities. Extensive experiments conducted on large-scale multimodal sentiment datasets demonstrate the effectiveness of UniMSA in comparison to state-of-the-art methods.**

***Index Terms*—multimedia sentiment analysis, uncertain missing modalities, uncertainty estimation, emotional patterns**

I. Introduction

In real-world scenarios, data generated by human activities often contain rich emotional information. Multimodal Sentiment Analysis (MSA), which analyzes emotions across various modalities such as video, text, and audio, plays a pivotal role in advancing domains like smart cities [1], and social analysis [2], [3]. However, challenges arise from missing modality information, necessitating effective strategies for cross-modal information transfer and integration to achieve a comprehensive understanding of emotions.

Recent advancements in MSA have facilitated the development of sophisticated models leveraging techniques such as uncertainty estimation [4], Transformers [5], and causal models [6]. While these models have improved sentiment analysis performance, the issue of missing modality information widely existing in real-world applications remains insufficiently addressed. As a result, a growing body of research is now exploring MSA under conditions of missing modalities.

For instance, MMIN [7] leverages multimodal learning to capture modality couplings, while TFR-Net [8] utilizes Transformers to extract semantic information from multiple modalities. However, they primarily depend on the remaining intact modalities to reconstruct missing information, placing significant dependence on the reliability of the intact modalities. This assumption often falls short in real-world scenarios, where missing data occurs randomly and unpredictably, leaving insufficient reliable information in the intact modalities to bridge the gaps effectively. To address this limitation, LNLN [9] incorporates randomly missing data instances into its training strategy to enhance robustness in handling incomplete data. As a result, sentiment analysis models capable of addressing randomly missing modalities have gradually become a widely adopted approach in the field.

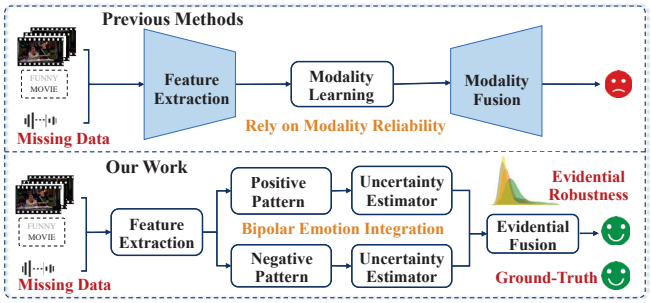


Fig. 1. Model comparison between previous methods and our proposed work.

Despite advancements in MSA methods for incomplete data, aimed at improving modality extraction and fusion techniques, several challenges remain: **(1) Low Modality Reliability:** The inherently random nature of modality dropout introduces significant noise, diminishing the reliability of extracted information and complicating its use in subsequent analysis. **(2) Severe Loss of Emotional Patterns:** Random modality dropout often results in the loss of critical emotional patterns. For example, when the key sentiment-bearing words are missing from the text modality and the original sentiments in the image and audio modalities become neutral, it becomes difficult to infer the intended emotional context accurately.

To address these challenges, we propose a novel approach, UniMSA, designed for Bipolar **Un**certa**i**nty Reliable **M**ultimodal **S**entiment **A**nalysis. As illustrated in Fig. 1, our model enhances bipolar emotional uncertainty learning, overcoming the reduced modality reliability often observed in existing approaches under conditions of missing modalities. Moreover, it explicitly explores and addresses uncertainty within emotional patterns. Specifically, **UniMSA** integrates the uncertainty of both positive and negative semantic emotional information, thereby increasing the model's sensitivity to subtle emotional patterns. This strategy significantly improves the robustness and accuracy of sentiment prediction, even in the presence of incomplete data. The key contributions of our work are summarized as follows:

(1) We introduce a novel MSA model for bipolar emotional uncertainty, effectively capturing both positive and potentially negative uncertain relationships among missing emotional patterns. Our model enhances the reliability of the recognition of fundamental emotional pattern signals.

(2) The proposed model shifts the focus from modality reliability to the reliability of underlying emotional patterns. By completing missing emotional patterns through bipolar information fusion, it significantly improves MSA performance.

(3) Extensive experiments conducted on three multimodal datasets demonstrate that UniMSA outperforms state-of-the-art baselines in sentiment analysis accuracy under conditions of missing modality, highlighting its robustness in uncertain environments. The source codes and datasets are available at https://github.com/SuperPower97/UniMSA.

II. Related Work

Most existing MSA studies operate under the assumption of complete modality data availability [5], [6]. However, in real-world scenarios, data completeness is often compromised due to sensor limitations, necessitating robust mechanisms for handling missing modality issue and ensuring reliable decision-making. Therefore, many modality missing methods primarily aim to learn the correlations between complete and missing modalities to construct more robust representations. For instance, MMIN [7] proposes a cross-modal learning to capture coupling relationships. TFR-Net [10] introduces Transformers to extract intra-modal and inter-modal relationships, constructing a feature reconstruction network to recover missing modality semantics. However, these modality missing methods heavily rely on the quality of the complete modalities, often overlooking the randomness inherent in modality loss. Consequently, recent research has increasingly focused on the randomly modality missing methods that explore arbitrary inter-modality correlations. For instance, LNLN [9] uses randomly missing data instances as a strategic approach to enhance MSA's robustness when dealing with incomplete data.

Additionally, existing robustness studies in MSA primarily address either data quality uncertainty or modality-specific uncertainty. TMSON [11] proposes Gaussian distributions and Bayesian fusion to manage robust multimodal uncertainty, while DiCMoR [12] introduces normalizing flows to recover missing modality features aligned with real data distributions.

However, they often overlook the substantial reduction in modality reliability caused by random modality dropout, which undermines the effectiveness of sentiment modality extraction. In contrast, our proposed UniMSA enhances bipolar emotional uncertainty learning to capture clearer emotional messages.

III. Methodology

The overall framework of our proposed UniMSA – bipolar emotional uncertainty learning network in multimodal sentiment analysis with incomplete data – is depicted in Fig. 2.

A. Multimodal Feature Extraction

Data Missing Preprocessing. Following the methodology outlined in [8], we simulate the missing information by randomly removing varying proportions of data (ranging from 0% to 100%) for each modality, where the missing modalities inherently exhibit high uncertainties. Specifically, for the visual and audio modalities, the missing information is replaced with zeros; for the language modality, the erased data is substituted with the [UNK] token in BERT, which is a pretrained language model.

Multimodal Data Preprocessing. Following previous work on multimodal processing [13], we first obtain input representations by processing language (l) with BERT, visual information (v) with OpenFace , and audio (a) with Librosa. These representations are denoted as $\mathbf{X}_m^0 \in \mathbb{R}^{T_m \times d_m}$, where $m \in \{l, v, a\}$, T_m is the sequence length, and d_m is the vector dimension for each modality. Subsequently, we perform missing data preprocessing on $\mathbf{X}_m^0$, resulting in a multimodal input $\mathbf{X}_m^1 \in \mathbb{R}^{T_m \times d_m}$ with random noise.

B. Bipolar Emotion Integration

Drawing on the previous work [13], due to the significant interference caused by the loss of semantic emotional information in the dominant language modality after random data missing, we design *Bipolar Emotion Integration*, a novel module that leverages bipolar emotional information directly from other modalities to perform dual-semantic emotional correction for the language-dominant module.

Initial Modality Embedding. We first propose a two-layer Transformer encoder to extract and integrate features from the multimodal input $\mathbf{X}_m^1$. Each modality begins with a randomly initialized low-dimensional vector $\mathbf{H}_m^0 \in \mathbb{R}^{T \times d_m}$, which is processed by the encoder to obtain the initial feature representation $\mathbf{H}_m^1 \in \mathbb{R}^{T \times d}$:

$$\mathbf{H}_m^1 = \mathrm{Encoder}\left(\mathrm{concat}(\mathbf{H}_m^0, \mathbf{X}_m^1), \theta_m\right), \tag{1}$$

where Encoder$(\cdot)$ extracts each modality's features from the concatenation of $\mathbf{H}_m^0$ and $\mathbf{X}_m^1$, under the condition of different parameters θ_m.

Adaptive Hyper-modality Fusion (AHF). Since we consider the language feature representation $\mathbf{H}_l^1$ as the primary determinant for predicting sentiment, we utilize two layers of Transformer-based encoders to learn the dominant representations at different scales [13]. The process of learning middle- and high-scale representations ($\mathbf{H}_l^2$ and $\mathbf{H}_l^3$) is defined as:

$$\mathbf{H}_l^i = \mathrm{E}_l^i\left(\mathbf{H}_l^{i-1}, \theta_m\right) \in \mathbb{R}^{T \times d}, \tag{2}$$

where $i \in \{2, 3\}$ represents the i-th layer of the multi-scale Transformer encoder $\mathrm{E}_l^i(\cdot)$. Unlike *Initial Modality Embedding*, which primarily aims to perform the embedding process, the purpose of this encoder is to effectively transfer essential information into the initial embedding.

After obtaining the language features of different scales, a hyper-modality feature $\mathbf{H}_{\text{hyper}} \in \mathbb{R}^{T \times d}$ is calculated by a multi-

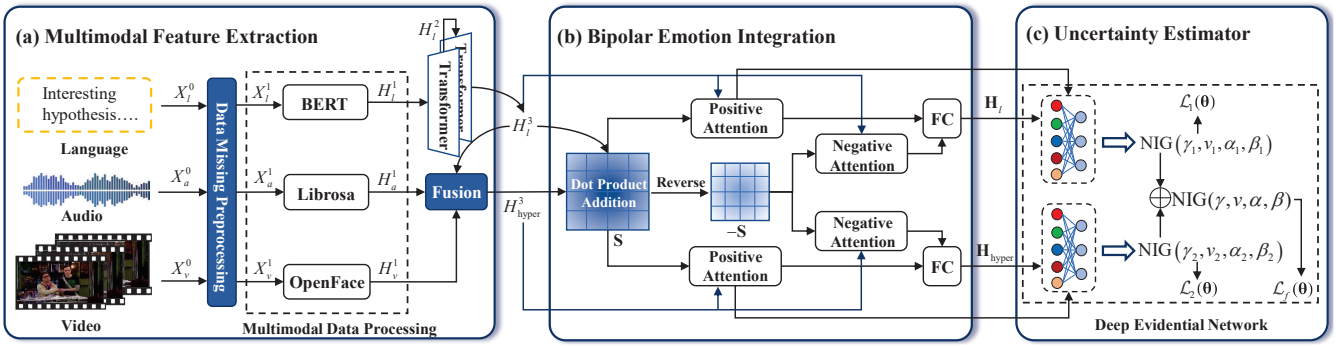


Fig. 2. The framework of UniMSA: **(a)** We introduce data imputation and preprocess multimodal data from raw inputs. **(b)** We capture positive and negative fundamental emotional signals by multimodal bipolar emotional attention mechanism. **(c)** We propose an evidential uncertainty estimator to precisely model the real uncertainty inherent in incomplete modality data.

head attention relationship between the language features and the two remaining modalities:

$$\mathbf{H}_{\text{hyper}}^{j} = \mathbf{H}_{\text{hyper}}^{j-1} + \text{MHA}\left(\mathbf{H}_{l}^{j}, \mathbf{H}_{a}^{1}\right) + \text{MHA}\left(\mathbf{H}_{l}^{j}, \mathbf{H}_{v}^{1}\right), \quad (3)$$

where $j \in \{1, 2, 3\}$ represents the number of iterations in the above formula, and $\text{MHA}(\cdot)$ denotes the multi-head attention.

Bipolar Emotional Attention Learning. Now we compute the fine-grained similarity matrix $\mathbf{S} \in \mathbb{R}^{T \times T}$, representing the similarity between each feature of $\mathbf{H}_{\text{hyper}}^{3}$ and $\mathbf{H}_{3}^{i}$. The similarity scores are computed using the addition attention mechanism [14] and the scaled dot-product attention mechanism, as follows:

$$\mathbf{S} = \tanh\left(\mathbf{H}_{\text{hyper}}^{3} + \mathbf{H}_{l}^{3}\right)\mathbf{W}_{s}, \ \mathbf{S} = \mathbf{H}_{\text{hyper}}^{3}(\mathbf{H}_{l}^{3})^{\mathrm{T}}/\sqrt{d_e}, \quad (4)$$

where $\mathbf{W}_{s} \in \mathbb{R}^{d_e \times 1}$ represents a learnable weight vector. Then we introduce the positive and negative learnings. Positive learning aims to identify the most similar emotional features, while negative learning seeks to uncover the contrasting emotional information.

Positive Learning. After calculating the most similar emotional features between the modalities, we obtain the features $\mathbf{H}_{\text{hyper}}^{p} \in \mathbb{R}^{T \times d}$ and $\mathbf{H}_{l}^{p} \in \mathbb{R}^{T \times d}$ guided by positive emotional information as follows:

$$\mathbf{H}_{\text{hyper}}^{p} = \text{softmax}(\mathbf{S})\mathbf{H}_{\text{hyper}}^{3}, \quad \mathbf{H}_{l}^{p} = \text{softmax}\left(\mathbf{S}^{\mathrm{T}}\right)\mathbf{H}_{l}^{3}. \quad (5)$$

Negative Learning. In addition to positively correlated emotional information, conflicting emotional information also plays a crucial role in sentiment analysis. To obtain the negatively correlated attention vectors $\mathbf{H}_{\text{hyper}}^{n} \in \mathbb{R}^{T \times d}$ and $\mathbf{H}_{l}^{n} \in \mathbb{R}^{T \times d}$, the similarity matrix S is multiplied by the high-scale representations with a negative constant:

$$\mathbf{H}_{\text{hyper}}^{n} = \text{softmax}(-\mathbf{S})\mathbf{H}_{\text{hyper}}^{3}, \ \mathbf{H}_{l}^{n} = \text{softmax}(-\mathbf{S}^{\mathrm{T}})\mathbf{H}_{l}^{3}. \quad (6)$$

Next, two fully-connected (FC) layers are employed to combine positive and negative attention vectors for obtaining $\mathbf{H}_{\text{hyper}}^{*}$ and $\mathbf{H}_{l}^{*}$, while allowing two types of guided information to merge and produce final vectors $\mathbf{H}_{\text{hyper}}$ and $\mathbf{H}_{l}$:

$$\mathbf{H}_{\text{hyper}}^{*} = \text{FC}(\mathbf{H}_{\text{hyper}}^{p} \oplus \mathbf{H}_{\text{hyper}}^{n}), \mathbf{H}_{l}^{*} = \text{FC}(\mathbf{H}_{l}^{p} \oplus \mathbf{H}_{l}^{n}), \quad (7)$$

$$\mathbf{H}_{\text{hyper}} = \text{FC}(\mathbf{H}_{\text{hyper}}^{3} \oplus \mathbf{H}_{l}^{*}), \mathbf{H}_{l} = \text{FC}(\mathbf{H}_{l}^{3} \oplus \mathbf{H}_{\text{hyper}}^{*}). \quad (8)$$

C. Emotional Evidence Uncertainty Estimator

Now we have updated the representations of multimodal features $\mathbf{H}_{\text{hyper}}$ and dominant linguistic features $\mathbf{H}_{l}$. Given the high uncertainty associated with randomly missing sentiment modalities, we evaluate the emotional uncertainty for sentiment prediction, $\hat{y}$, by modeling it using a Gaussian distribution, $\mathcal{N}(\mu, \sigma^2)$, which satisfies the relationships:

$$\hat{y} \sim \mathcal{N}(\mu, \sigma^2), \quad \mu \sim \mathcal{N}(\gamma, \sigma_v^2) \quad \sigma^2 \sim \Gamma^{-1}(\alpha, \beta), \quad (9)$$

where $\Gamma(\cdot)$ denotes the Gamma function, $\gamma \in \mathbb{R}^2$, $v > 0$, $\alpha > 1$, and $\beta > 0$. Our objective is to estimate the posterior distribution $q(\mu, \sigma^2|\hat{y})$, which has the approximate form of a Normal Inverse-Gamma (NIG) distribution:

$$p\left(\{\mu, \sigma^2\} \mid \Omega\right) = \frac{\beta^{\alpha}\sqrt{v}}{\Gamma(\alpha)\sqrt{2\pi\sigma^2}}\left(\frac{1}{\sigma^2}\right)^{\alpha+1}\exp\left\{-\frac{2\beta + v(\gamma-\mu)^2}{2\sigma^2}\right\}, \quad (10)$$

where the NIG distribution is denoted as $\Omega = \{\gamma, v, \alpha, \beta\}$. Next, we constrain the training process to generate the distribution hyperparameters: first increasing the emotional evidence to support observations, then reducing evidence (increasing uncertainty) when predictions are incorrect.

Maximizing the emotional evidence. In Bayesian theory, emotional evidence represents the likelihood of the observation $\hat{y}$ given the evidential distribution parameters, Ω. When applying the NIG prior, the analytical solution for the evidence is:

$$p(\hat{y}|\Omega) = St(\hat{y}; \gamma, \frac{\beta}{v\alpha + v}, 2\alpha), \quad (11)$$

where $St(\cdot)$ is the student's t distribution. To train the model, we use the negative log of model evidence as the loss function:

$$\mathcal{L}^{\text{NLL}}(\boldsymbol{\theta}) = \frac{1}{2}\log\left(\frac{\pi}{v}\right) - \alpha\log(\omega) + \log\left(\frac{\Gamma(\alpha)}{\Gamma\left(\alpha + \frac{1}{2}\right)}\right) \quad (12)$$

$$+ \left(\alpha + \frac{1}{2}\right)\log\left((\hat{y} - \gamma)^2 v + \omega\right), \quad (13)$$

where $\omega = 2\beta(1 + v)$. By maximizing the model emotional evidence, the neural network can output the appropriate NIG parameters to fit the observed data.

Minimizing the emotional evidence. We regularize the training by applying an incorrect evidence penalty to reduce the

evidence for incorrect predictions. Specifically, we use the following evidence regularizer:

$$\mathcal{L}^{\mathrm{R}}(\boldsymbol{\theta}) = |\hat{y} - \mathbb{E}[\mu]| \cdot \phi = |\hat{y} - \gamma| \cdot (2v + \alpha), \tag{14}$$

herein, $E[\mu]$ is actually the prediction γ. Now the total loss scaled by the trade-off coefficient λ is defined as follows:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}^{\mathrm{NLL}}(\boldsymbol{\theta}) + \lambda \mathcal{L}^{\mathrm{R}}(\boldsymbol{\theta}). \tag{15}$$

Training. Given two NIG distributions $\mathrm{NIG}\left(\gamma_1, v_1, \alpha_1, \beta_1\right)$ and $\mathrm{NIG}\left(\gamma_2, v_2, \alpha_2, \beta_2\right)$ from different perspectives, $\mathbf{H}_{\text{hyper}}$ and $\mathbf{H}_l$, the summation of these two NIG distributions is:

$$\begin{gathered}\mathrm{NIG}(\gamma, v, \alpha, \beta) := \mathrm{NIG}\left(\gamma_1, v_1, \alpha_1, \beta_1\right) \oplus \mathrm{NIG}\left(\gamma_2, v_2, \alpha_2, \beta_2\right), \\ \text{where} \quad \gamma = \left(v_1 + v_2\right)^{-1}\left(v_1\gamma_1 + v_2\gamma_2\right), \quad v = v_1 + v_2 \\ \alpha = \alpha_1 + \alpha_2 + \frac{1}{2}; \beta = \beta_1 + \beta_2 + \frac{1}{2}v_1\left(\gamma_1 - \gamma\right)^2 + \frac{1}{2}v_2\left(\gamma_2 - \gamma\right)^2.\end{gathered} \tag{16}$$

We define the final loss function $\mathcal{L}_{all}(\boldsymbol{\theta})$ as the sum of the losses from two views and the fused distribution:

$$\mathcal{L}_{all}(\boldsymbol{\theta}) = \mathcal{L}_1(\boldsymbol{\theta}) + \mathcal{L}_2(\boldsymbol{\theta}) + \mathcal{L}_f(\boldsymbol{\theta}). \tag{17}$$

At last, the final loss combined with emotional uncertainty constraint $\mathcal{L}_{all}(\boldsymbol{\theta})$ and the MSE loss, can be written as:

$$\mathcal{L} = \mathcal{L}_{all}(\boldsymbol{\theta}) + \frac{1}{N}\sum_{n=0}^{N}\|y^n - \hat{y}^n\|_2^2, \tag{18}$$

where N is the number of samples in the training set.

IV. Experiment

A. Experimental Configuration

Datasets. MOSI [15] is a benchmarking dataset for multimodal sentiment analysis, containing 2,199 video clips with text, audio, and visual modalities. Each clip is labeled with sentiment intensity scores. **MOSEI** [16] is a larger version of MOSI, with over 22,856 video clips covering diverse topics from YouTube. It includes sentiment intensity labels across text, audio, and visual modalities, facilitating more comprehensive emotion recognition tasks. Both datasets assign sentiment scores to each sample, ranging from -3 (strongly negative) to 3 (strongly positive). **SIMS** [17] is a large-scale Chinese sentiment dataset consists of 2,281 video clips extracted from various movies and TV series. Each sample is manually annotated with a sentiment score on a scale from -1 (negative) to 1 (positive).

Benchmarks. In this study, we compare our method with eight baseline models as below: MISA [18], MMIN [7], Self-MM [19], TFR-Net [10], CENET [20], TETFN [21], ALMT [13] and LNLN [9].

Evaluation Criteria. To evaluate performance, we report several metrics including seven-class accuracy (Acc-7), five-class accuracy (Acc-5), binary classification accuracy (Acc-2), binary classification F1 score, mean absolute error (MAE), and the correlation coefficient between predictions and labels (Corr) on the MOSI and MOSEI datasets. For Acc-2, we compute both accuracy and F1 scores in two configurations: negative/positive (left) and negative/non-negative (right). For the SIMS dataset, we report the Acc-5, three-class accuracy (Acc-3), Acc-2, F1 score, MAE, and Corr.

Implementation. We conduct ten experiments with missing rates from 0 to 0.9 (increment of 0.1) and average the results to evaluate performance. We employ the Adam optimizer with a learning rate of $1e^{-4}$ for optimizing the model parameters. Our model is trained for 100 epochs to achieve best results. The model is implemented in PyTorch and run on a GeForce RTX 4090 GPU.

B. Overall Performance

We evaluated the performance of our method UniMSA and baselines across multiple evaluation metrics on the MOSI, MOSEI, and SIMS datasets. The results are shown in Table I and Table II. **(1) Outstanding Overall Performance**: UniMSA outperforms other models, including recent advanced approaches such as LNLN and ALMT, in critical evaluation metrics including Acc-7, Acc-5, Acc-3, Acc-2, F1, and Corr. This confirms that UniMSA excels in capturing nuanced sentiment patterns information across different modalities, establishing its superiority in multimodal sentiment analysis tasks. **(2) Enhanced Reliability of Missing Emotional Patterns**: A notable strength of UniMSA is its ability to effectively recover missing emotional patterns. Even when modalities are randomly missing, our model consistently maintains high accuracy, demonstrating its robustness in scenarios with incomplete or noisy input data. This is a significant advantage in real-world applications, where modality or data missing issues are common. **(3) Impact of Bipolar Emotional Uncertainty Learning**: The innovation of bipolar emotional uncertainty learning primarily lies in reorienting attention from modality reliability towards the stability of fundamental emotional pattern signals. This shift significantly enhances the performance of binary sentiment analysis. The method not only improves the Acc-2 metric but also leads to substantial gains in more fine-grained metrics such as Acc-5 and Acc-7.

C. Ablation Study

Table III and Table IV show the ablation results, where we evaluate the contributions of three key modules in our model. The variants tested include: (1) **w/o AHF**: Removing the adaptive hyper-modality fusion module which facilitates learning across language, audio, and visual modalities results in a noticeable performance drop, underscoring its importance in multimodal interaction. (2) **w/o Bipolar**: Excluding the bipolar emotion integration which integrates positive and negative information leads to reduced performance, highlighting its critical role in sentiment analysis. (3) **w/o Uncertainty**: Excluding the emotional evidence uncertainty estimator – which captures multimodal distribution uncertainty – causes a performance degradation. This emphasizes its significance in handling noisy data. The ablation study confirms that each module is essential for the improved model performance, and any removal leads to a decrease in accuracy.

TABLE I
SENTIMENT ANALYSIS PERFORMANCE COMPARISON OF THE OVERALL PERFORMANCE ON MOSI AND MOSEI DATASETS.

Method	MOSI						MOSEI					
	Acc-7	Acc-5	Acc-2	F1	MAE	Corr	Acc-7	Acc-5	Acc-2	F1	MAE	Corr
MISA [18]	29.79	33.12	71.50 / 70.36	71.29 / 70.01	1.087	0.525	40.69	39.28	71.27 / 75.82	64.75 / 68.29	0.784	0.512
MMIM [7]	31.19	33.57	69.21 / 67.16	66.64 / 64.12	1.073	0.510	39.68	40.72	72.74 / 78.25	69.38 / 71.68	0.742	0.489
Self-MM [19]	29.49	34.76	70.42 / 69.31	66.51 / 67.60	1.068	0.514	44.70	45.38	74.54 / 77.55	68.83 / 72.45	0.695	0.498
TFR-Net [10]	29.55	34.78	68.32 / 66.46	61.78 / 60.16	1.212	0.461	45.32	34.67	73.62 / 77.23	68.72 / 70.32	0.739	0.488
CENET [20]	30.66	35.21	71.45 / 67.82	68.52 / 64.90	1.083	0.508	44.47	43.68	74.67 / 77.34	70.68 / 74.08	0.685	0.535
TETFN [21]	30.28	34.31	69.69 / 67.67	65.79 / 63.49	1.086	0.507	32.40	45.42	69.76 / 67.68	65.69 / 63.29	1.087	0.508
ALMT [13]	30.34	33.24	70.88 / 68.58	72.36 / 71.67	1.079	0.496	40.84	41.52	76.66 / 77.56	77.22 / 78.32	0.692	0.481
LNLN [9]	32.73	35.94	72.45 / 70.99	72.73 / 71.52	1.094	0.535	45.01	45.85	76.58 / 78.17	77.67 / 79.86	**0.684**	0.540
UniMSA	**33.80**	**36.36**	**72.50 / 71.06**	**72.75 / 71.89**	**1.045**	**0.536**	**45.35**	**46.43**	**77.17 / 79.22**	**77.68 / 80.65**	0.686	**0.547**

TABLE II
SENTIMENT ANALYSIS PERFORMANCE COMPARISON OF THE OVERALL PERFORMANCE ON SIMS DATASET.

Method	Acc-5	Acc-3	Acc-2	F1	MAE	Corr
MISA [18]	32.21	53.97	71.24	67.40	0.538	0.352
MMIN [7]	32.28	52.76	69.78	67.41	0.534	0.342
Self-MM [19]	31.24	54.29	70.94	69.46	0.528	0.367
TFR-Net [10]	25.96	52.98	67.84	57.62	0.657	0.189
CENET [20]	24.28	53.34	69.62	58.89	0.586	0.124
TETFN [21]	32.64	53.87	72.57	68.67	0.525	0.381
ALMT [13]	22.12	47.58	69.66	72.58	0.557	0.364
LNLN [9]	32.16	54.61	69.69	71.31	**0.519**	0.368
UniMSA	**33.64**	**55.63**	**72.81**	**74.99**	0.524	**0.372**

D. Hyperparameter Analysis

As illustrated in Fig. 3, we conducted a sensitivity analysis on the trade-off coefficient parameter λ, which plays a crucial role in balancing the uncertainty measurement objectives. We explored various λ values including $1e^{\{-4,-3,-2,-1\}}$. As λ increases, the trends for Acc-2 and Acc-5 show that the performance of most datasets steadily improves, with a gradual increase in model performance. However, there are exceptions, such as the trend for Acc-5 on the MOSI dataset, which initially decreases before rising again. Nonetheless, in all cases, the peak performance is observed at $\lambda = 1e^{-1}$. Beyond this point, higher values of λ lead to diminishing returns, with excessive penalization of imbalance potentially causing overfitting, thereby hindering the model's ability to generalize. Taking into account the overall performance of λ across Acc-2, Acc-5, and all other parameters, the sentiment analysis model achieves the optimal configuration when $\lambda = 1e^{-1}$.

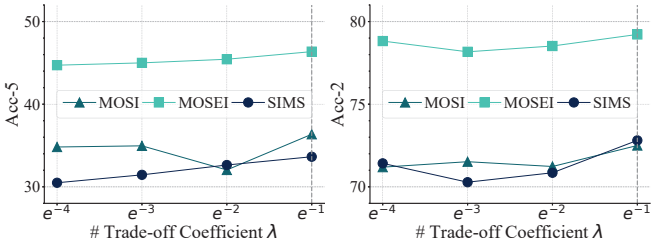


Fig. 3. Parameter Analysis of UniMSA.

E. Visualization in Bipolar Emotion Integration

To further elaborate on the transmission mechanisms of positive and negative information in positive learning and negative learning, we apply a softmax operation to the fine-grained similarity matrix $\mathbf{S}$ and its adverse matrix $-\mathbf{S}$, followed by heatmap visualization of the resulting matrices. As shown in Fig. 4, particularly in the regions enclosed by the red box, areas with larger weights in positive learning are scarcely learned in negative learning. Conversely, information that is not emphasized in positive learning occupies a central role in negative learning, which precisely corresponds to the information often overlooked by other related methods.

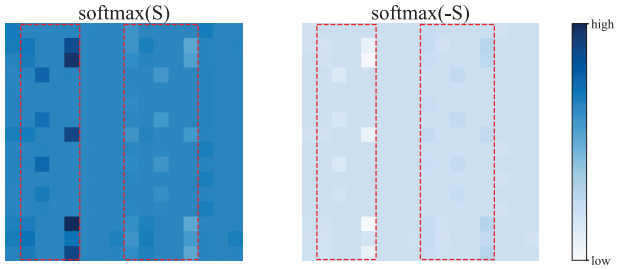
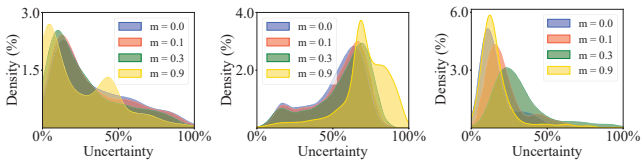


Fig. 4. Heatmap visualization of softmax($\mathbf{S}$) and softmax($-\mathbf{S}$)

F. Evidential Robustness

To validate the robustness of our UniMSA against random modality missingness, we visualize the uncertainty distribution under simulated measurement noise in Fig. 5. With the parameter m governing the missing rate, we observe that the uncertainty scores exhibit a smooth distribution across the entire range, even in the absence of supplementary noise. This indicates that, despite inherent measurement noise in real-world datasets, the evidence bipolar sentiment uncertainty learning can capture these patterns. Additionally, as noise levels increase, uncertainty also rises, confirming that the proposed method effectively detects the intensity of modality missing noise and provides timely feedback.



(a) MOSI (b) MOSEI (c) SIMS

Fig. 5. Evidential robustness under different missing ratio.

TABLE III
ABLATION STUDIES OF UNIMSA'S MODULES ON MOSI AND MOSEI DATASETS.

Method	MOSI						MOSEI					
	Acc-7	Acc-5	Acc-2	F1	MAE	Corr	Acc-7	Acc-5	Acc-2	F1	MAE	Corr
UniMSA	**33.80**	**36.36**	**72.50 / 71.06**	**72.75 / 71.89**	**1.045**	**0.536**	**45.35**	**46.43**	**77.17 / 79.22**	**77.68 / 80.65**	**0.686**	**0.547**
w/o AHF	31.07	34.57	69.77 / 69.27	69.60 / 69.20	1.126	0.469	44.76	45.87	76.09 / 76.60	76.92 / 76.86	0.711	0.536
w/o Bipolar	32.37	35.48	68.92 / 71.29	69.44 / 71.23	1.133	0.533	44.95	45.96	74.73 / 67.72	74.33 / 69.13	0.702	0.533
w/o Uncertainty	30.39	33.51	72.40 / 70.88	72.45 / 71.04	1.074	0.520	45.20	46.21	77.08 / 78.43	77.52 / 80.46	0.689	0.534

TABLE IV
ABLATION STUDIES OF UNIMSA'S MODULES ON SIMS.

Method	Acc-5	Acc-3	Acc-2	F1	MAE	Corr
UniMSA	**33.64**	**55.63**	**72.81**	**74.42**	**0.524**	**0.372**
w/o AHF	32.56	54.00	72.04	73.37	0.526	0.353
w/o Bipolar	31.43	54.02	71.20	73.59	0.525	0.368
w/o Uncertainty	31.15	52.51	70.35	71.18	0.538	0.332

G. Case Study

As illustrated in Fig. 6, we present two successful prediction cases selected from the MOSI dataset. Compared to the LNLN method, UniMSA exhibits a superior ability to recognize sentiment in samples with randomly missing data. This is achieved through the reinforcement of both positive and negative sentiment cues, enabling the model to effectively capture reliable sentiment patterns. This highlights UniMSA's capacity to approximate uncertainty in noisy environments, ultimately enabling it to extract accurate sentiment signals even when faced with incomplete or ambiguous data.

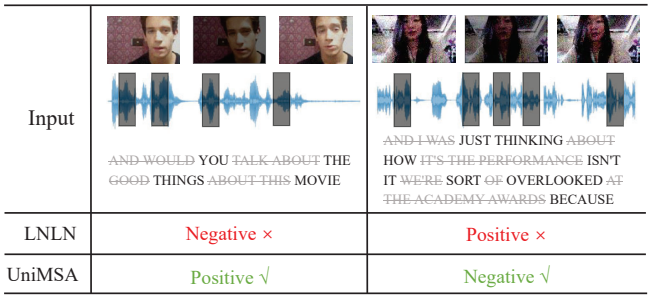


Fig. 6. Visualization of correct predictions. Note: Random data missing is applied to the input sequence for illustration.

V. CONCLUSION

We introduced UniMSA, a novel bipolar uncertainty learning framework for reliable multimodal sentiment analysis with missing modalities. By enhancing the bipolar emotional uncertainty learning, our method integrates the positive and negative evidence uncertainty to recover the emotional signals, improving the reliability of missing data for accurate sentiment analysis. Our future work would explore more advanced uncertainty estimation techniques and focus on improving reliability for more multimodal applications such as malicious content detection.

VI. ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation of China (Grant No.62176043, No.62072077, and No.U22A2097).

REFERENCES

[1] X. Liu, X. Li, Y. Cao, F. Zhang, X. Jin, and J. Chen, "Mandari: Multimodal temporal knowledge graph-aware sub-graph embedding for next-poi recommendation," in *ICME*, 2023, pp. 1529–1534.
[2] Z. Cheng, J. Zhang, X. Xu, G. Trajcevski, T. Zhong, and F. Zhou, "Retrieval-augmented hypergraph for multimodal social media popularity prediction," in *KDD*, 2024, pp. 445–455.
[3] X. Xu, Y. Zhang, F. Zhou, and J. Song, "Improving multimodal social media popularity prediction via selective retrieval knowledge augmentation," in *AAAI*, 2025.
[4] Z. Gao, X. Jiang, H. Chen, Y. Li, Y. Yang, and X. Xu, "Uncertainty-debiased multimodal fusion: Learning deterministic joint representation for multimodal sentiment analysis," in *ICME*, 2024, pp. 1–6.
[5] Y. Li, Y. Wang, and Z. Cui, "Decoupled multimodal distilling for emotion recognition," in *CVPR*, 2023, pp. 6631–6640.
[6] D. Yang, Z. Chen, Y. Wang *et al.*, "Context de-confounded emotion recognition," in *CVPR*, 2023, pp. 19 005–19 015.
[7] J. Zhao, R. Li, and Q. Jin, "Missing modality imagination network for emotion recognition with uncertain missing modalities," in *ACL*, 2021, pp. 2608–2618.
[8] Z. Yuan, W. Li *et al.*, "Transformer-based feature reconstruction network for robust multimodal sentiment analysis," in *ACM MM*, 2021, pp. 4400–4407.
[9] H. Zhang, W. Wang, and T. Yu, "Towards robust multimodal sentiment analysis with incomplete data," in *ACL*, 2024.
[10] Z. Yuan, W. Li, H. Xu, and W. Yu, "Transformer-based feature reconstruction network for robust multimodal sentiment analysis," in *ACM MM*, 2021, pp. 4400–4407.
[11] Z. Xie, Y. Yang, J. Wang, X. Liu, and X. Li, "Trustworthy multimodal fusion for sentiment analysis in ordinal sentiment space," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.
[12] Y. Wang, Z. Cui, and Y. Li, "Distribution-consistent modal recovering for incomplete multimodal learning," in *ICCV*, 2023, pp. 22 025–22 034.
[13] H. Zhang, Y. Wang, G. Yin, K. Liu, Y. Liu, and T. Yu, "Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis," in *EMNLP*, 2023.
[14] D. Bahdanau, "Neural machine translation by jointly learning to align and translate," *arXiv:1409.0473*, 2014.
[15] A. Zadeh, R. Zellers *et al.*, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.
[16] A. B. Zadeh, P. P. Liang *et al.*, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *ACL*, 2018, pp. 2236–2246.
[17] W. Yu, H. Xu, F. Meng *et al.*, "Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in *ACL*, 2020, pp. 3718–3727.
[18] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *ACM MM*, 2020, pp. 1122–1131.
[19] Z. Yuan, W. Li, H. Xu, and W. Yu, "Transformer-based feature reconstruction network for robust multimodal sentiment analysis," in *ACM MM*, 2021, pp. 4400–4407.
[20] D. Wang, S. Liu, Q. Wang *et al.*, "Cross-modal enhancement network for multimodal sentiment analysis," *IEEE Transactions on Multimedia*, vol. 25, pp. 4909–4921, 2022.
[21] D. Wang, X. Guo, Y. Tian, J. Liu, L. He, and X. Luo, "Tetfn: A text enhanced transformer fusion network for multimodal sentiment analysis," *Pattern Recognition*, vol. 136, p. 109259, 2023.