DFMU: Distribution-based Framework for Modeling Aleatoric Uncertainty in Multimodal Sentiment Analysis

Chen Tang[1], **Tingrui Shen**[1], **Xinrong Gong**[2], **Chong Zhao**[1] and **Tong Zhang**[1,†]
[1]School of Computer Science and Engineering, South China University of Technology, Guangzhou, China
[2]School of Engineering, Huaqiao University, Quanzhou, China
Tangyueming311@gmail.com, tony@scut.edu.cn

Abstract

In Multimodal Sentiment Analysis (MSA), data noise arising from various sources can lead to uncertainty in Aleatoric Uncertainty (AU), significantly impacting model performance. Current efforts to address AU have insufficiently explored its sources. They primarily focus on modeling noise rather than implementing targeted modeling based on its origin. Consequently, these approaches struggle to effectively mitigate the influence of AU, resulting in sustained limitations in model performance. Our research identifies that the AU primarily stems from two problems: **subjective bias in the annotation process and the complex set relationships of sentiment features**. To specifically address them, we propose DFMU, a Distribution-based Framework for Modeling Aleatoric Uncertainty, which incorporates an uncertainty modeling block capable of encoding uncertainty distributions and adaptively adjusting optimization objectives. Furthermore, we introduce distribution-based contrastive learning with sentiment words replacement to better capture the complex relationships among features. Extensive experiments on three public MSA datasets, i.e., MOSI, MOSEI, and SIMS, demonstrate that the proposed model maintains robust performance even under high noise conditions and achieves state-of-the-art results on these popular datasets.

1 Introduction

MSA is a pivotal research task that seeks to comprehend individuals' sentiment states and analyze multimodal information present in online video data, to predict the intensity of expressed sentiment. This capability is central in diverse real-world applications, such as affective intelligence [Li *et al.*, 2023a] and human-computer interaction [Jiang *et al.*, 2020].

However, multimodal signals inevitably contain redundant and noisy information, which causes unavoidable Aleatoric Uncertainty(AU). According to previous studies [Kendall and Gal, 2017][Gawlikowski *et al.*, 2023], AU in deep learning reflects the inherent uncertainty issues within the data, which are related to data noise. This uncertainty significantly affects model performance and makes it difficult for the model to achieve reliability. Most current work in MSA overlooks the inherent AU problem. A few works that consider AU primarily focus on modeling noise rather than implementing targeted modeling based on its origin. There is a lack of further exploration into the sources of AU in the MSA field, with understanding limited to the presence of noise in the data. As a result, previous models that have attempted to model uncertainty in MSA struggle to genuinely mitigate the impact of AU, leading to continued limitations in model performance. According to experiments, this work finds that in the MSA field, the generation of AU primarily stems from two problems: subjective biases in the annotation of MSA datasets

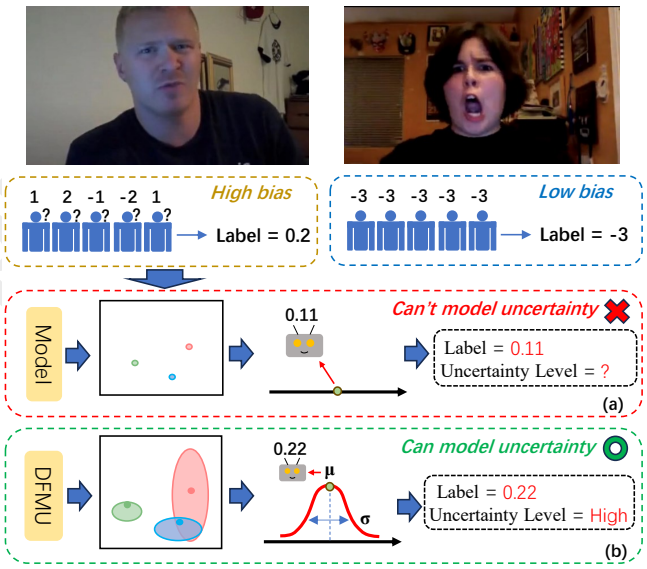


Figure 1: The annotation labels may have inconsistency as depicted in the figure. The higher the difference in annotation values, the higher the degree of uncertainty. Traditional approaches modeled the sample as a point in the representation space, as shown in (a), resulting in the loss of information about the degree of uncertainty. In contrast, DFMU employs distribution representation to predict the level of uncertainty, as shown in (b).

† Corresponding author.

and the complex set relationships of sentiment features.

As shown in Figure 1, some samples in the dataset express sentiment with ambiguity, increasing the difficulty of annotation. Additionally, the subjectivity of the annotators leads to inconsistent label values for the same sample. The degree of inconsistency reflects the bias of the samples, referred to as **subjective annotation bias**. Samples with high subjective annotation bias contain greater AU. Moreover, in Figure 1, points represent correlations solely through distances in the representation space, without expressing their inclusion relationships. For instance, in distribution representations, the sentiment features of joy and pride are almost inclusive, while they do not intersect with sadness. Therefore, there exist relationships between sentiment features that cannot be expressed through deterministic point distances, termed **complex set relationships of sentiment features**. Our model needs to specifically address this complex relationship.

In response to the discovered source of AU generation in MSA, we designed a specialized model to achieve better dataset performance, and we obtained improved results in AU-related noise resistance experiments. To address Problem 1, we designed an uncertainty modeling block that utilizes Gaussian distributions to encode uncertain distribution representations. These representations can store AU information of samples with high subjective annotation bias in the variance dimension of the distribution. Subsequently, the proposed uncertainty regularization can adaptively adjust the optimization objectives based on the obtained uncertainty information. This regularization enables the model to learn variance during the training process, ultimately displaying AU through information from the variance dimension. To tackle Problem 2, we proposed a distribution-based contrastive learning with sentiment words replacement to model the complex set relationships of sentiment features. We employed a sentiment dictionary data augmentation method to enhance the fine-grained sentiment-capturing capability and then used distribution-based contrastive learning to endow DFMU with the ability to model complex set relationships.

The main contributions of this work can be summarized as follows:

- This paper presents DFMU, which reveals through empirical research that the source of AU generation in the MSA lies in the issues of subjective bias in MSA dataset annotations and the complex set relationships of sentiment features.
- This paper introduces an uncertainty modeling block and uncertainty regularization, modeling sentiment features as uncertain distribution representations to enable DFMU to quantify the subjective annotation bias.
- This paper proposes distribution-based contrastive learning with sentiment words replacement, utilizing distribution contrastive loss and sentiment dictionary data augmentation methods to model complex set relationships of sentiment features.
- DFMU achieves state-of-the-art performance across multiple widely adopted datasets. This work provides analysis of uncertainty issues and offers comprehensive empirical results to demonstrate the effectiveness and necessity of DFMU.

2 Related Work

In this section, we review relevant research in the areas of MSA uncertainty modeling, and contrastive learning.

2.1 Multimodal Sentiment Analysis

Previous work primarily relied on deterministic approaches to model sentiment representations, often employing contrastive learning [Yu *et al.*, 2023]or complex fusion methods [Zhang *et al.*, 2023] to address noise in modality information. Some works [Liang *et al.*, 2023] attributed the noise to redundancy and conflicting information across modalities, while [Hazarika *et al.*, 2020] suggested projecting each modality into modality-invariant and modality-specific representations to model the noise. However, these approaches overlooked the issue of uncertainty in the noise that cannot be resolved through deterministic point representations.

2.2 Uncertainty Modeling

Recently, the MSA field has gradually begun to consider the impact of AU on prediction. AU refers to the inherent noise in the provided training data. TMSON [Xie *et al.*, 2024] proposed to use Trustworthy Multimodal Fusion to estimate the reliability of each multimodality. EAU [Gao *et al.*, 2024] believes that each modality in a multimodal task has separate noise, and proposes to use a novel generic and robust multimodal fusion strategy to better model unimodal noise. TMSC [Xu *et al.*, 2024] constructs the modal private task (unique) by using the Dirichlet distribution and evidence theory to solve the uncertainty of each modal noise. However, previous methods have stopped at linking AU with noise, without exploring the source of AU in the MSA field, which prevents them from truly modeling the complete AU.

2.3 Contrastive Learning

Contrastive learning aims to learn effective representations by bringing positive samples closer together and pushing negative samples farther apart in embedding space. In multimodal tasks, HyCon [Mai *et al.*, 2023] explored relationships between samples and classes within and across modalities, while ConFEDE [Yang *et al.*, 2023] and ConKI [Yu *et al.*, 2023] process each modality into two different features to explore similarities and differences. Previous works treated sentiment features as deterministic points in the feature space, failing to capture their complex set relationships.

3 Method

The overall processing flow of DFMU for handling uncertainty in MSA tasks is illustrated in Figure 2. As shown in the figure, DFMU primarily consists of three components: Uncertainty Modeling Block, Distribution-based Contrastive Learning, and Uncertainty Regularization.

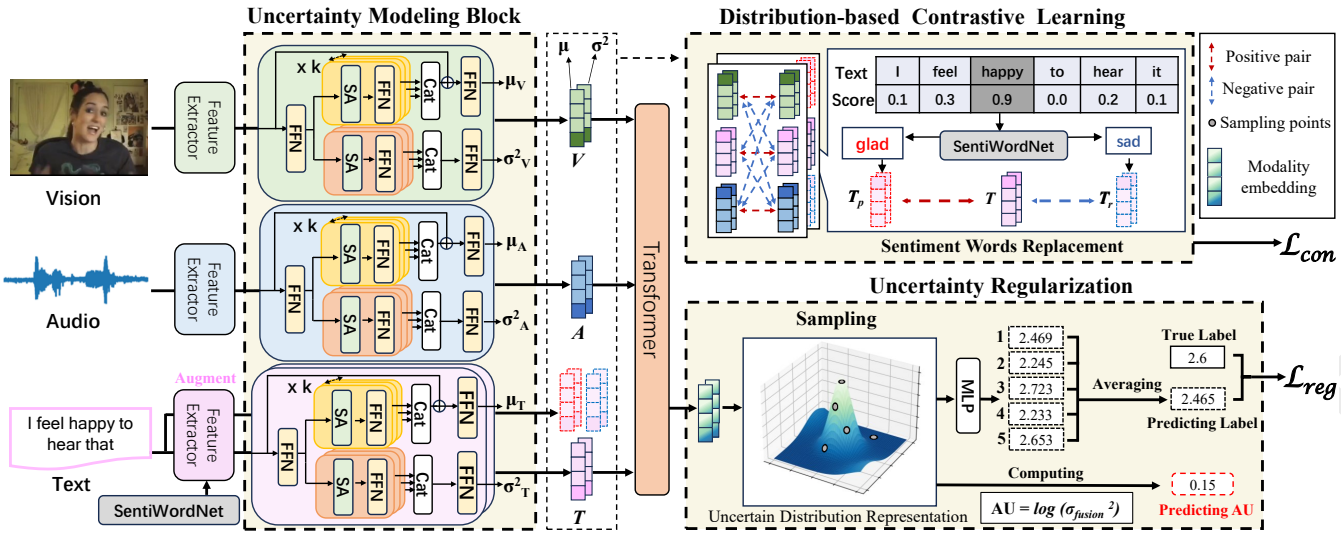


Figure 2: DFMU begins by extracting unified modality features from the input data. Then DFMU models the sentiment features as uncertain high-dimensional distributions by Uncertainty Modeling Block. Within the high-dimensional space of sentiment distribution representation, a Distribution-based Contrastive Learning approach is utilized to model the complex set relationships between different sentiments. The novel Sentiment Words Replacement method is applied to improve the fine-grained modeling capability of contrastive learning. With the fused sentiment distribution features, uncertainty regularization is employed to fit the label, thus establishing the Distribution-based Framework.

3.1 Uncertainty Modeling Block

To encode uncertain distribution representations, DFMU utilizes Gaussian distributions $\mathcal{N}(\mu, \sigma^2)$ to represent the sentiment features of different modalities in the sentiment distribution space. We assume Gaussian distributions to model our uncertain representations and learn mappings for mean and covariance vectors. As shown in Figure 2, the Uncertainty Modeling Block (UMB) is proposed to better model the single-modality data relationships by using multi-head self-attention. Let $E \in \{V, A, T\}$ be a matrix in $\mathbb{R}^{L\times D}$, where V, A, T represent the initial visual, audio, and textual embeddings, respectively. Here L is the sequence length and D is the embedding size. E is sent to an Feed Forward layers(FFD) and two pathways, μ and σ^2. Then it is split into k heads. Within each head, the input hidden state of each path is $E^{(i)} \in \mathbb{R}^{L\times D/2k}$ in i-th head. After $E^{(i)}$ processing in corresponding Self-Attention heads, the k outputs are concatenated and then passed through an additional self-attention(SA) layer and an FFD to model the relationships across the sequences. The σ^2 path is similar to the μ path. Since the input point representations are related to the mean of the distributional representations, a residual network is employed to learn the mean vector. Formally, the operation in the μ path is:

$$Head_{\mu}^{(i)} = \text{Self-Attention}\left(E^{(i)}\right), \quad (1)$$

$$\mu_E = \text{FFD}\left(\text{Concat}_{i\in k}\left[Head_{\mu}^{(i)}\right] + E^{(i)}\right). \quad (2)$$

3.2 Distribution-Based Contrastive Learning

The Distribution-Based Contrastive Learning module is proposed to model the complex set relationships of features. To measure the distance between multivariate Gaussian distributions, the 2-Wasserstein distance [Mallasto and Feragen, 2017] is employed, different modality embeddings $E \in \{T, A, V\}$, as an example:

$$D_{2w}(E_i, E_j) = \left\|\mu_{E_i} - \mu_{E_j}\right\|_2^2 + \left\|\sigma_{E_i}^2 - \sigma_{E_i}^2\right\|_2^2, \quad (3)$$

where E_i, E_j represent the samples of corresponding modality embeddings. μ_{E_i} and $\sigma_{E_i}^2$ are the mean and variance values obtained from E_i after UMB processing.

Sentiment Words Replacement

To further enhance the model's learning of textual sentiment distribution, we design a distribution-based text augmentation for contrastive learning: Sentiment Words Replacement (SWR) method, which is shown in Figure 3. Unlike random words processing, we first explore the Sentiment scores of each word in the sentence by using SentiWordNet [Baccianella *et al.*, 2010], where higher scores represent stronger sentiment. The words with high sentiment scores are chosen for a series of augmentations. Specifically, two types of text augmentations are constructed: positive sentences and negative sentences. Each type of augmentation operates on $k_t\%$ of words in the sentence. Positive sentences are constructed by replacing words with synonyms and negative sentences are constructed by replacing words with antonyms.

In each training iteration, DFMU feeds the batches of original sentences, positive sentences, and negative sentences into the Feature Extractor, obtaining corresponding sets of text embeddings, which denoted as $\{T^1, ..., T^b\}$, $\{T_p^1, ..., T_p^b\}$, and $\{T_n^1, ..., T_n^b\}$, respectively. Here, b is batchsize, T represents the text embedding of an original sentence, T_p and T_n represent the text embedding of a positive sentence and negative sentence respectively.

The SWR contrastive learning loss is defined as $\mathcal{L}_c^{SWR}$, using InfoNCE [van den Oord *et al.*, 2018] and r represents either positive sentences or negative sentences:

$$\mathcal{L}_c^{SWR} = -\sum_{i=1}^{b} \log \frac{\exp(D_{2w}(T^i, T_p^i))}{\sum_{j=1}^{b} \exp\left(D_{2w}\left(T^i, T_r^j\right)\right)}, \quad (4)$$

where $r \in \{p,n\}$.

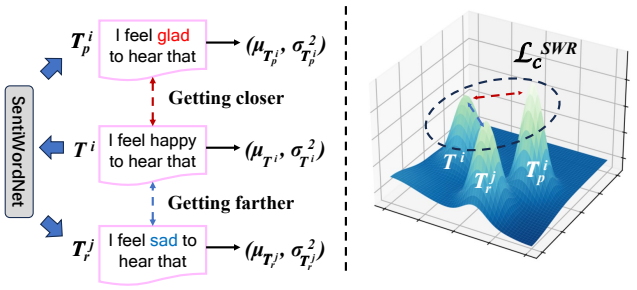


Figure 3: In Sentiment Words Replacement, the text modality of input samples is enhanced using SentiWordNet to create positive and negative counterpart samples. These samples are then processed through the UMB to convert them into corresponding distributional representations, as depicted in the figure on the right. Subsequently, distribution-based contrastive learning is applied.

Intra-Modal and Cross-Modal Contrastive Learning

If the sentiment scores of two samples are close, we consider their intended sentiment intents to be similar, and uncertain distribution representations should be close. Specifically, positive pairs are defined as samples whose absolute difference in sentiment scores is less than a threshold value k_e and negative pairs are on the contrary:

$$P = \{(E_i, E'_j) | E, E' \in \{T, A, V\} \quad \& \quad i \neq j \quad \& \quad |Z_i - Z_j| < k_e\}, \quad (5)$$

$$N = \{(E_i, E'_j) | E, E' \in \{T, A, V\} \quad \& \quad i \neq j \quad \& \quad |Z_i - Z_j| > k_e\}, \quad (6)$$

where, P and N represent positive pairs and negative pairs respectively. E_i, E'_j represent the samples of corresponding modality embeddings, Z represents the sentiment score of corresponding sample, which can be acquired from datasets. The contrastive learning loss for both Intra-modal and Cross-moda Contrastive Learning is defined as $\mathcal{L}$, where $(E_p, E'_q), (E_s, E'_t)$ represent the corresponding combination in P or N:

$$\mathcal{L} = -\log \frac{\sum_{(E_p, E'_q) \in P} \exp(D_{2w}(E_p, E'_q))}{\sum_{(E_s, E'_t) \in P \cup N} \exp(D_{2w}(E_s, E'_t))}. \quad (7)$$

For **Intra-modal Learning**, we compare samples in the same modality of data to model intra-modal information. Based on Equation 5, 6, it requires $E = E'$, that is, the two samples' modalities in combinations in P and N are the same. Then uses Equation 7 to compute Intra-modal loss $\mathcal{L}_c^{intra}$.

For **Cross-modal Contrastive Learning**, we compare samples in the different modalities of data to learn the relationships between different modalities and contribute to model cross-modal information. Based on Equation 5, 6, it requires $E \neq E'$, that is, the two samples' modalities in combinations in P and N are different. Then uses Equation 7 to compute Cross-modal loss $\mathcal{L}_c^{cross}$.

3.3 Uncertainty Regularization

To learn the distribution of the representation's mean and variance, the original Mean Square Error(MSE) regression loss is modified to prevent the variance from collapsing to zero or exploding to extremely large values. In our experiments, we discovered that replacing the MSE loss entirely with the regularization loss made the model difficult to converge. Therefore, a hyperparameter α is proposed to balance the MSE loss and the regularization loss. The $\mathcal{L}_{\text{mse}}$ is Mean squared error loss, and the $\mathcal{L}_{\text{reg}}$ is regularization loss:

$$\mathcal{L}_{reg} = \alpha\left(\frac{\mathcal{L}_{mse}}{2\sigma^2} + \frac{1}{2}\log\sigma^2\right) + (1-\alpha)\mathcal{L}_{mse}. \quad (8)$$

To simulate the discrepancies among n annotators during data annotating, we designed the method to perform multi-point sampling within the distribution, as shown in Figure 1 (b) and Figure 2. Each sampled point is individually predicted and takes the average of the multiple predictions. Subsequent ablation experiments confirmed that this approach helps model the subjective biases that may arise during data annotation. It is worth noting that the sampling operation of the distribution poses challenges for backpropagation. By applying the reparameterization trick [Kingma and Welling, 2014], the model firstly samples random noise from the standard normal distribution instead of directly sampling from the distribution :

$$z = \mu + \sigma\epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \quad (9)$$

Following the equation, the output z follows the predictive distribution of the partial differential equation. Hence, the computation of mean and standard deviation from the sampling operation can be separated, and they are trainable.

The final loss for all the tasks is:

$$\mathcal{L}_{all} = \alpha_{SWR}\mathcal{L}_c^{SWR} + \alpha_{cross}\mathcal{L}_c^{cross} + \alpha_{intra}\mathcal{L}_c^{intra} + \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_{reg}^i, \quad (10)$$

where $\alpha_{SWR}, \alpha_{cross}, \alpha_{intra}$ are hyperparameters of losses, n is the number of samples, and $\mathcal{L}_{reg}^i$ is the regularization loss of the i-th sample.

4 Experiment

4.1 Experimental Settings

The experiments were conducted on three publicly available benchmark datasets in MSA: **CMU-MOSI** [Zadeh *et al.*, 2016], **CMU-MOSEI** [Zadeh *et al.*, 2018] and **CH-SIMS** [Yu *et al.*, 2020].

For fairness, we employ the same feature extractors adopted in ALMT [Zhang *et al.*, 2023] for vision, audio, text

Models	CMU-MOSI					CMU-MOSEI				
	Corr (↑)	Acc-7 (↑)	Acc-2 (↑)	F1 (↑)	MAE (↓)	Corr (↑)	Acc-7 (↑)	Acc-2 (↑)	F1 (↑)	MAE (↓)
MISA	0.795	-	83.53 / 85.52	83.46 / 85.51	0.714	0.761	-	83.34 / 85.88	83.91 / 85.76	0.542
MMIM	0.800	46.65	84.14 / 86.06	84.00 / 85.98	0.700	0.772	54.24	82.24 / 85.97	82.66 / 85.94	0.526
Self-MM	0.795	45.79	82.54 / 84.77	83.68 / 84.91	0.712	0.767	53.46	82.68 / 84.96	82.95 / 84.93	0.529
ALMT	0.801	48.27	84.21 / 86.38	84.25 / 86.29	0.699	0.762	53.73	84.52 / 86.34	85.06 / 86.44	0.529
TMSON	0.803	46.9	85.36 / 87.01	85.32 / 86.99	0.690	0.758	54.42	85.13 / 86.28	85.02 / 86.24	0.531
EAU	0.801	48.47	84.73 / 86.12	84.66 / 86.15	0.722	0.778	54.31	85.29 / 86.56	85.25 / 86.33	0.533
TMSC	0.793	48.66	83.78 / 85.01	83.64 / 85.02	0.689	0.768	53.41	84.02 / 85.5	83.94 / 85.53	0.527
ConKI	0.809	48.22	84.52 / 86.17	84.53 / 86.18	0.688	0.771	54.05	82.87 / 86.45	83.34 / 86.47	0.526
ConFEDE	0.784	43.12	84.21 / 85.61	84.23 / 85.62	0.737	0.774	54.25	82.34 / 86.07	82.36 / 86.09	0.524
HyCon	0.787	45.84	83.82 / 85.27	83.91 / 85.16	0.711	0.769	52.21	83.93 / 85.44	83.17 / 85.59	0.597
DFMU	**0.830**	**49.45**	**85.78 / 87.66**	**86.01 / 87.68**	**0.663**	**0.780**	**54.50**	**85.65 / 87.59**	**85.61 / 87.62**	**0.513**

Table 1: Performance of DFMU compared to SOTA approaches on MOSI and MOSEI datasets. Higher metric values indicate better performance except "MAE". The "MAE" metric is better with lower values.

Models	Corr (↑)	Acc-5(↑)	Acc-3 (↑)	Acc-2(↑)	F1(↑)	MAE (↓)
ALMT	0.602	45.53	67.28	80.87	81.02	0.426
UDMF	0.617	45.14	66.28	80.89	80.94	0.427
EAU	0.605	42.20	64.31	78.98	79.35	0.501
TMSC	0.608	44,37	65.85	80.41	80.33	0.414
ConKI	0.568	43.37	65.21	78.12	78.22	0.463
DFMU	**0.627**	**46.13**	**68.96**	**81.32**	**81.91**	**0.399**

Table 2: Performance of DFMU compared to SOTA approaches on CH-SIMS dataset. Higher metric values indicate better performance except "MAE". The "MAE" metric is better with lower values.

and also retrain all baselines under the same conditions. All the experiment records are aggregated over three independent runs.

Experimental results are reported in two forms: regression and classification. For regression, MAE (mean absolute error) and Corr (Pearson correlation) are reported. For classification, Acc-2 (binary classification accuracy), Acc-7 (seven-class classification accuracy) and F1 score are reported. Higher values indicate better performance for all metrics except for MAE.

4.2 Baselines

We compare DFMU with the following state-of-the-art baseline models in MSA: based on contrastive learning methods: ConKI [Yu *et al.*, 2023], ConFEDE [Yang *et al.*, 2023], Hycon [Mai *et al.*, 2023]; uncertainty modeling models: TMSON [Xie *et al.*, 2024], EAU [Gao *et al.*, 2024], TMSC [Xu *et al.*, 2024] ; deterministic modeling models: MMIM [Han *et al.*, 2021], MISA [Hazarika *et al.*, 2020], ALMT [Zhang *et al.*, 2023], Self-MM [Yu *et al.*, 2021].

4.3 Performance Comparison

Table 1 and Table 2 list the comparison results of our proposed and state-of-the-art methods on the MOSI, MOSEI, and CH-SIMS, respectively. It can be observed from these tables that DFMU yields better results to a range of baseline models on all datasets.

It was noteworthy that DFMU was the only model to exceed an F1 score of 87 on both the MOSI and MOSEI datasets. On MOSEI, it surpassed the second-best model, ConKI, by a significant margin of 1.15 points, which showed

Method	MAE(↓)	Acc-7(↑)	Corr(↑)
w/o Text modality	1.364	24.31	0.224
w/o Vision modality	0.704	47.53	0.807
w/o Audio modality	0.701	48.18	0.812
w/o $\mathcal{L}_c^{SWR}$	0.712	46.07	0.796
w/o $\mathcal{L}_c^{cross}$	0.694	47.60	0.818
w/o $\mathcal{L}_c^{intra}$	0.684	47.89	0.813
w/o Uncertainty Modeling Block	0.693	44.83	0.793
w/o Distribution-based CL	0.728	44.03	0.789
w/o Uncertainty Regularizaion	0.685	46.36	0.807
DFMU	**0.663**	**49.45**	**0.830**

Table 3: Ablation results when (1) without different modalities (2) without different uncertainty contrastive learning methods (3) without different uncertainty modules.

that DFMU can accurately model uncertainty and demonstrate superior performance in fitting high subjective annotation bias samples. Furthermore, the Acc-7 metric of the DFMU on the MOSI exceeded the performance of the best baseline model by 0.79, though the task was particularly a challenging seven-class classification problem, within the context of a regression task. The superiority of DFMU in Acc-7 metric demonstrated that it was excellent performance in predicting labels across the entire range.

4.4 Ablation Study

An ablation study about modalities was conducted first, as shown in Table 3. This finding validated that the text modality plays a dominant role in the MSA. In addition, a series of ablation experiments on subtasks of our proposed distribution-based contrastive learning were conducted. When the three contrastive learning subtasks were removed, the Acc-7 metric of the model decreased by 3.38, 1.85, and 1.56, which showed that among the three sub-tasks, the SWR task showed the highest improvement in performance.

Finally, a series of ablation experiments were conducted on our proposed Uncertainty Regularizaion, Uncertainty Modeling Block and Distribution-based Contrastive Learning. We found that the absence of Uncertainty Regularization and the Uncertainty Modeling Block resulted in a decrease of 3.09

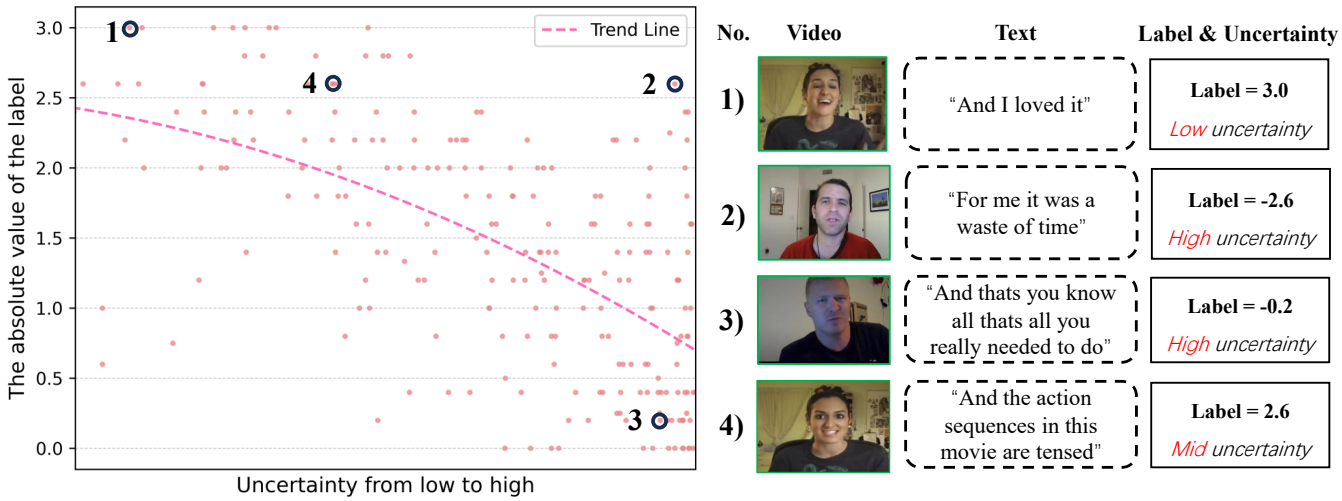


Figure 4: Visualization evidence of uncertainty issues in the dataset, the x-axis represents the degree of AU, and the y-axis represents the sentiment value.

Hyperparameter	Value	Acc-7(↑)	F1(↑)
α_{SWR}	0.05	47.28	84.76
	0.1	43.68	82.54
	0.3	29.54	79.18
α_{cross}	0.05	45.52	86.11
	0.1	45.95	85.19
	0.3	45.95	85.66
α_{intra}	0.05	45.83	86.10
	0.1	45.39	85.95
	0.3	45.25	84.56

Table 4: Ablation results when DFMU removed the other two uncertainty contrastive learning methods. The Values meant the values of the hyperparameter of the remaining method.

and 4.62, respectively, in the model's Acc-7 scores. Moreover, without the Distribution-based Contrastive Learning, the performance suffered the most, which led to a substantial 5.42 reduction in the Acc-7 metric. The Distribution-based Contrastive Learning helped our model learn better representations of sentiment distributions.

4.5 Hyperparameters of Contrastive Learning

To study the impact of hyperparameters of different uncertainty contrastive learning methods in more detail, we conducted the ablation experiments of three hyperparameters in MOSI. According to Table 4, we found that the cross-modal contrastive learning and Intra-modal Learning methods had little impact on the results, and were insensitive to hyperparameters. On the contrary, sentiment words replacement contrastive learning method was very sensitive to hyperparameters. The value of Acc-7 metric dropped 37.52% when α_{SWR} changed from 0.05 to 0.3. We speculated that this may be because the loss value of this method is very large, which caused the situation.

Models	CMU-MOSI	CMU-MOSEI
UniSA	84.11	84.93
UniMSE	**85.85** / 86.9	**85.86** / 87.50
GPT-4V	80.43	-
DFMU	85.78 / **87.66**	85.65 / **87.59**
DFMU w/o Audio modality	84.67 / 86.45	84.88 / 86.95

Table 5: Performance of DFMU compared to the pre-trained models and the multimodal large language model. The metric in the table is "Acc-2" and higher values indicate better performance. The results of UniSA [Li *et al.*, 2023b] and UniMSE [Hu *et al.*, 2022] are from their original papers. The result of GPT-4V is from [Lian *et al.*, 2024].

4.6 Comparison With Large Models

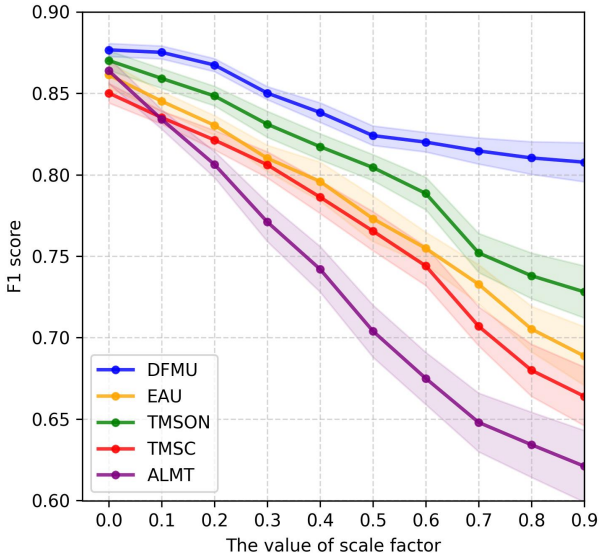
To prove the effectiveness of DFMU, we compared it to the pre-trained models and the multimodal large language model, as shown in Table 5. Our model remained ahead in performance although the number of parameters and training data was much smaller than that of large models. It was notable that DFMU exceeded GPT-4V by 7.23 in CMU-MOSI. The weak performance of GPT-4V can be partly attributed to the fact that GPT-4V does not support audio input and the loss of information. The performance of DFMU without audio modality still exceeded GPT-4V in CMU-MOSI by 6.02. The result proved that the necessity of the ability to model the uncertainty and the damage of AU to performance.

4.7 MSA Uncertainty Problems Analysis

Subjective Annotation Bias

To verify one of the proposed sources of AU: subjective annotation bias, we used the model to output the AU values corresponding to each sample in MOSI and visualized them. As in Figure 4, we selected representative samples for analysis:

- In sample 1, a person had a consistently highly positive facial expression and the corresponding text. This is the easiest scenario for annotation.



Model	0.1	0.2	0.3	0.4	-	0.9
DFMU	0.17%	1.05%	3.02%	4.38%	-	7.85%
EAU	1.86%	3.58%	5.93%	7.59%	-	20.03%
TMSON	1.26%	2.5%	4.49%	6.07%	-	16.31%
TMSC	1.76%	3.38%	5.17%	7.51%	-	21.89%
ALMT	3.44%	6.65%	10.74%	14.1%	-	28.09%

Figure 5: The figure compares our model's performance with baselines on data with different values of scale factor. The higher value means higher uncertainty. The table below represents the percentage decrease in performance for each model relative to the original scores at the corresponding scale factor. The translucent area represents the confidence interval.

- In sample 2, the person's facial expression remained neutral, but the text was obviously negative. Annotators will disagree on how to annotate the sentiment score when text and facial expressions are inconsistent, causing high sample uncertainty.
- In sample 3, the person's expression switched between frowning and expressionless, and the text is neutral, which caused complex sentiments that are difficult to annotate, causing high uncertainty.
- In sample 4, the person had a highly positive facial expression but with neutral text. Compared to Sample 1, the difficulty of annotation increases, and uncertainty also rises.

Meanwhile, a trend line was plotted in the scatter plot, the polynomial fitting function is used here to calculate the trend line. The line showed a clear negative correlation between the sentiment score and the AU. This is because samples with lower absolute label values are more likely to express complex or ambiguous sentiment, increasing the difficulty of labeling and resulting in subjective annotation bias, which in turn generates AU.

Comparison in High Uncertainty

We added different degrees of noise to the MOSI test data to simulate uncertainty. Specifically, we extract noise vectors from a Gaussian distribution and then multiply these noise vectors by a scaling factor to contaminate the test data, simulating the AUs present in the data. We compared the performance of DFMU and the baselines in the test data and visualized the results as line graphs in Figure 5. By comparing the results, we observed that the performance of all baselines decreased significantly as uncertainty increased. When the scaling factor was set to 9, our model's score only decreased by 7.85%, while the score of ALMT, which cannot model AU, dropped by 28.09%.

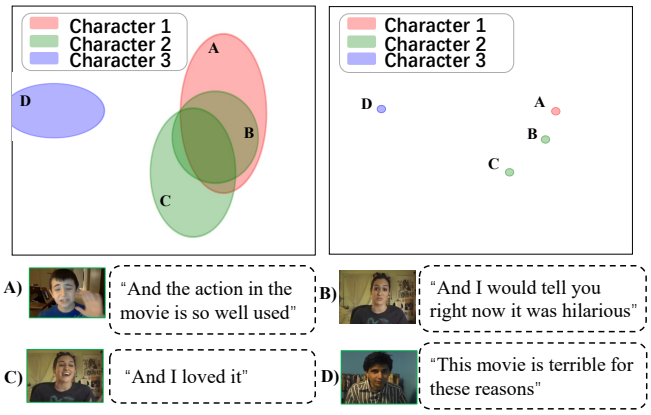


Figure 6: A 2-D toy experiment was performed by using T-SNE [Rauber *et al.*, 2016]. In the 2-D Uncertain Distribution Representation, sample B is almost an intersection of samples A and C. Sample B represents only positive text without gestures or positive expressions, while samples A and C represent positive text with gestures and positive text with positive expressions, respectively.

Complex Set Relationships of Sentiment

To verify one of the sources of AU: complex set relationships of sentiment, we conducted a visual analysis of the distribution representations of sentiment features in the model. As shown in Figure 6, In the left diagram, samples A B C D preserved the fine-grained set relationships among the features, where samples A B C intersected with one another but do not overlap with D. However, in the right diagram, all relationships degenerated into mere distance relationships, indicating that point representations lost more information compared to distribution representations, particularly for samples with similar label values. This indicated that the loss of sentiment information in point representations also contributed to AU.

5 Conclusion

In this work, we explore two causes for the emergence of Aleatoric Uncertainty in MSA: subjective bias in the annotation process and the complex relationships among sentiment features. To address these issues, we propose DFMU, a Distribution-based Framework for Modeling Aleatoric Uncertainty, which includes an uncertainty modeling module and distribution-based contrastive learning with sentiment word replacement. Our proposed model demonstrates robust performance even under high noise conditions, achieving state-of-the-art results on various popular datasets.

Acknowledgments

This work was supported by the National Natural Science Foundation of China grant under number 62222603, National Training Program of Innovation and Entrepreneurship for Undergraduates (202510561094).

Contribution Statement

The authors of this paper, Chen Tang and Tingrui Shen, have made equal contributions to the paper.

References

[Baccianella *et al.*, 2010] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*. European Language Resources Association, 2010.

[Gao *et al.*, 2024] Zixian Gao, Xun Jiang, Xing Xu, Fumin Shen, Yujie Li, and Heng Tao Shen. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26876–26885, 2024.

[Gawlikowski *et al.*, 2023] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589, 2023.

[Han *et al.*, 2021] Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 9180–9192. Association for Computational Linguistics, 2021.

[Hazarika *et al.*, 2020] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. MISA: modality-invariant and -specific representations for multimodal sentiment analysis. In Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann, editors, *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 1122–1131. ACM, 2020.

[Hu *et al.*, 2022] Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7837–7851, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[Jiang *et al.*, 2020] Yingying Jiang, Wei Li, M. Shamim Hossain, Min Chen, Abdulhameed Alelaiwi, and Muneer H. Al-Hammadi. A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition. *Inf. Fusion*, 53:209–221, 2020.

[Kendall and Gal, 2017] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

[Kingma and Welling, 2014] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[Li *et al.*, 2023a] Shuzhen Li, Tong Zhang, Bianna Chen, and C. L. Philip Chen. Mia-net: Multi-modal interactive attention network for multi-modal affective analysis. *IEEE Trans. Affect. Comput.*, 14(4):2796–2809, 2023.

[Li *et al.*, 2023b] Zaijing Li, Ting-En Lin, Yuchuan Wu, Meng Liu, Fengxiao Tang, Ming Zhao, and Yongbin Li. Unisa: Unified generative framework for sentiment analysis. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6132–6142, 2023.

[Lian *et al.*, 2024] Zheng Lian, Licai Sun, Haiyang Sun, Kang Chen, Zhuofan Wen, Hao Gu, Bin Liu, and Jianhua Tao. Gpt-4v with emotion: A zero-shot benchmark for generalized emotion recognition, 2024.

[Liang *et al.*, 2023] Paul Pu Liang, Zihao Deng, Martin Q. Ma, James Y. Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[Mai *et al.*, 2023] Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Trans. Affect. Comput.*, 14(3):2276–2289, 2023.

[Mallasto and Feragen, 2017] Anton Mallasto and Aasa Feragen. Learning from uncertain curves: The 2-wasserstein metric for gaussian processes. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5660–5670, 2017.

[Rauber *et al.*, 2016] Paulo E. Rauber, Alexandre X. Falcão, and Alexandru C. Telea. Visualizing time-dependent data

using dynamic t-sne. In Enrico Bertini, Niklas Elmqvist, and Thomas Wischgoll, editors, *18th Eurographics Conference on Visualization, EuroVis 2016 - Short Papers, Groningen, The Netherlands, June 6-10, 2016*, pages 73–77. Eurographics Association, 2016.

[van den Oord *et al.*, 2018] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.

[Xie *et al.*, 2024] Zhuyang Xie, Yan Yang, Jie Wang, Xiaorong Liu, and Xiaofan Li. Trustworthy multimodal fusion for sentiment analysis in ordinal sentiment space. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[Xu *et al.*, 2024] Guoxia Xu, Lizhen Deng, Yansheng Li, Yantao Wei, Xiaokang Zhou, and Hu Zhu. Trusted multimodal socio-cyber sentiment analysis based on disentangled hierarchical representation learning. *IEEE Transactions on Computational Social Systems*, 11(4):5287–5297, 2024.

[Yang *et al.*, 2023] Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. Confede: Contrastive feature decomposition for multimodal sentiment analysis. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 7617–7630. Association for Computational Linguistics, 2023.

[Yu *et al.*, 2020] Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. CH-SIMS: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3718–3727. Association for Computational Linguistics, 2020.

[Yu *et al.*, 2021] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 10790–10797. AAAI Press, 2021.

[Yu *et al.*, 2023] Yakun Yu, Mingjun Zhao, Shiang Qi, Feiran Sun, Baoxun Wang, Weidong Guo, Xiaoli Wang, Lei Yang, and Di Niu. Conki: Contrastive knowledge injection for multimodal sentiment analysis. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13610–13624. Association for Computational Linguistics, 2023.

[Zadeh *et al.*, 2016] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *CoRR*, abs/1606.06259, 2016.

[Zadeh *et al.*, 2018] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2236–2246. Association for Computational Linguistics, 2018.

[Zhang *et al.*, 2023] Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 756–767. Association for Computational Linguistics, 2023.