

EmoEUS: Uncertainty Supervision for Multimodal Emotion Recognition in Conversation

Zilong Huang¹, Kong Aik Lee¹, Junjie Li¹, Zhe Li², Man-Wai Mak¹

¹ Dept. of Electrical and Electronic Engineering, The Hong Kong Polytechnic University

² Speech, Language, and Cognition Laboratory, The University of Hong Kong

{zi-long.huang, kong-aik.lee, junjie98.li, man.wai.mak}@connect.polyu.hk, zheli8@hku.hk

Abstract

Multimodal emotion recognition in conversation (MERC) can leverage multimodal and contextual cues to boost recognition performance. However, existing fusion approaches in MERC often ignore modality-specific uncertainty across utterances caused by conflicting cues, varying noise, and missing modality-specific signals. We propose EmoEUS, an explicit uncertainty supervision framework for MERC. EmoEUS performs uncertainty-aware multimodal fusion by dynamically weighting modalities using learned variance estimates. We also introduce an explicitly supervised loss that aligns each utterance's predicted variance with the distance between the utterance's distributional representation and its emotion- and modality-specific cluster center. Experiments on IEMOCAP and MELD show that EmoEUS consistently outperforms state-of-the-art methods.

Index Terms: Explicit Uncertainty Supervision, Emotion Recognition in Conversation, Multimodal Fusion

1. Introduction

Multimodal emotion recognition in conversation (MERC) is important for applications such as human-computer interaction, intelligent medical care, and affective computing [1]. MERC analyzes dialogue turns from multiple modalities (e.g., text, audio, and visual) and assigns one emotion label per utterance [1, 2]. Prior work focuses on integrating multimodal information, modeling cross-modal interactions, and improving contextual feature representations [3–5].

Current MERC approaches predominantly adopt Transformer-based architectures to model multimodal conversational contexts [6, 7]. By jointly modeling multimodal input sequences, these methods effectively capture long-range contextual dependencies and cross-modal interactions, producing powerful fused representations for emotion recognition [2]. Recent studies further enhance these architectures through improved distillation mechanisms, graph-based modeling, or modality-specific encoders to better capture contextual cues and speaker dependencies in conversations [8–10]. Despite the advances in representation learning [11, 12], multimodal conversational data inherently exhibits significant uncertainty, which remains insufficiently addressed by existing fusion models [13–16]. While uncertainty modeling has been widely studied in multimodal learning to assess input reliability and improve model robustness [17–20], its exploration in MERC is still limited. Most existing MERC methods implicitly assume all modalities are uniformly reliable across every utterance, overlooking the substantial variations in their quality and informativeness [15]. In practice, modality quality and emotional contribution vary substantially due to noise, missing cues, or

conflicting signals, making uniform fusion suboptimal [16].

Although recent works have found the importance of introducing modality-specific uncertainty for MERC [21, 22], they typically model uncertainty implicitly through classification loss alone, lacking explicit supervision to dynamically leverage the emotional information of each modality for adaptive fusion. To address these limitations, we propose explicit uncertainty supervision for multimodal emotion recognition in conversation (**EmoEUS**). EmoEUS transforms point-wise features into distributional representations with variance via a context-level distribution estimator module (ContextDEM) and an explicitly supervised loss (ESL), and it performs fusion through an uncertainty-aware multimodal fusion (UAMF) mechanism. This design quantifies modality reliability and enables adaptive multimodal fusion under uncertain conversational contexts. Our main contributions are as follows:

- We propose a context-level distribution estimator module (ContextDEM) that models utterance-level uncertainty by transforming multimodal features into distributional representations with means and variances.
- We introduce an uncertainty-aware multimodal fusion (UAMF) mechanism that dynamically adjusts modality contributions according to their estimated uncertainty.
- We design an explicitly supervised loss (ESL) based on distributional cluster centers and 2-Wasserstein distance to provide explicit supervision for uncertainty estimation.

2. Methodology

2.1. Task Definition

Consider a dialogue consisting of k temporally ordered utterances, denoted as $\mathcal{U} = \{u_i\}_{i=1}^k$, where individual utterances have corresponding emotion labels $\{y_i\}_{i=1}^k$ and speaker labels $\{s_i\}_{i=1}^k$. The emotion label y_i belongs to a predefined set of emotions $\mathcal{Y}$. Each utterance u_i is accompanied by multimodal data, including a video clip, an audio segment, and a text transcript. Formally, we represent a dialogue as follows:

$$\mathcal{U} \times \mathcal{Y} \times \mathcal{S} = \left\{ \{u_i^{(\delta)}, y_i, s_i\} \mid \delta \in \{a, t, v\}, y_i \in \mathcal{Y}, s_i \in \mathcal{S} \right\}, \quad (1)$$

where $u_i^{(\delta)}$ denotes the modality-specific representation (text, audio, or video) of the i -th utterance. The ERC task involves predicting the emotion label $\hat{y}_i$ for each utterance u_i in the dialogue, leveraging information from all k utterances.

2.2. Overall Architecture

As shown in Fig. 1, EmoEUS consists of three components. First, ContextDEM maps multimodal features to Gaussian dis-

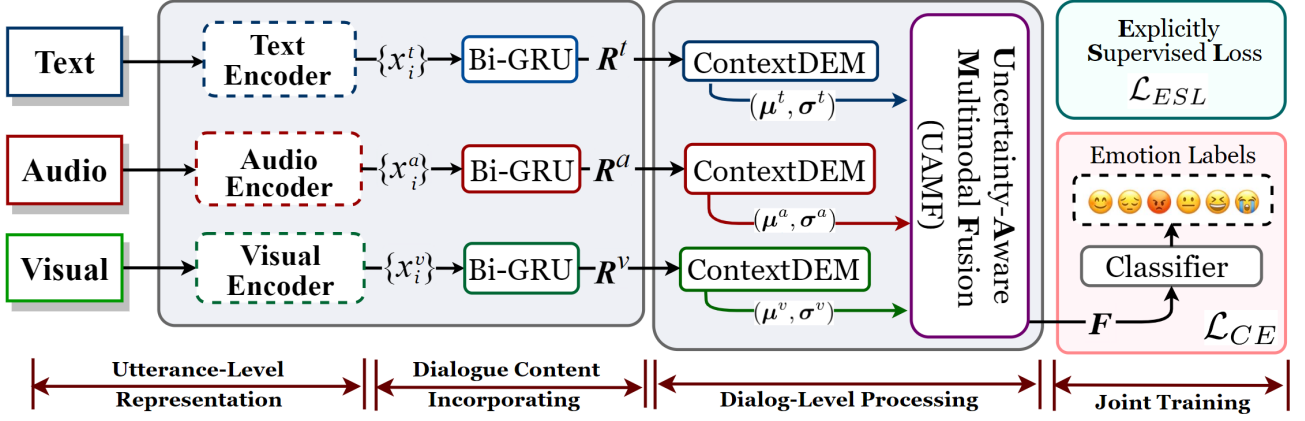


Figure 1: The overall architecture of the EmoEUS framework. EmoEUS enhances MERC performance through explicit uncertainty supervision and uncertainty-aware multimodal fusion. Solid and dashed rectangles denote trainable and frozen models, respectively.

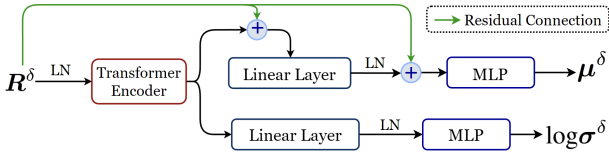


Figure 2: Framework of ContextDEM, where LN and MLP denote LayerNorm and multi-layer perceptron, respectively.

tributions (μ, σ^2) to model modality-specific uncertainty. Second, an uncertainty-aware multimodal fusion (UAMF) module integrates these distributions by weighting modalities based on their uncertainties. Finally, the explicitly supervised loss $\mathcal{L}_{ESL}$ aligns the predicted variance with the 2-Wasserstein distance between the emotion-specific cluster center and the utterance’s distribution representation, providing supervision for uncertainty estimation. The model is optimized using the classification loss $\mathcal{L}_{CE}$ together with the weighted $\mathcal{L}_{ESL}$.

2.3. Contextual Feature Extraction

For each dialogue utterance u_i ($i = 1, \dots, k$) with textual (u_i^t), acoustic (u_i^a), and visual (u_i^v) modalities, we employ modality-specific pre-trained encoders to extract features:

$$\mathbf{x}_i^\delta = \text{ModalityEncoder}(u_i^\delta), \delta \in \{t, a, v\}, \quad (2)$$

where the features are obtained by average pooling the final layer’s outputs of individual encoders. To capture sequential dependencies, the features are processed by bidirectional GRUs:

$$\mathbf{R}^\delta = \text{Bi-GRU}(\mathbf{x}_1^\delta, \dots, \mathbf{x}_k^\delta) \in \mathbb{R}^{D_r \times k}, \quad (3)$$

where D_r is twice the hidden state dimension of the GRUs.

2.4. Context-level Distribution Estimation

To model semantic uncertainty, we propose a context-level distribution estimator module (ContextDEM), as shown in Fig. 2. Given modality-specific features $\mathbf{R}^\delta$, the ContextDEM first enhances global contextual interactions via a Transformer encoder, followed by residual connections to preserve original context. The resulting features are then fed into two parallel MLP branches to independently predict the mean μ^δ and log-variance $\log \sigma^\delta$ of a Gaussian distribution:¹

$$(\mu^\delta, \log \sigma^\delta) = \text{ContextDEM}(\mathbf{R}^\delta), \delta \in \{t, a, v\}, \quad (4)$$

¹For notational simplicity, we denote σ as the variance instead of the usual σ^2 .

where $\mu^\delta, \sigma^\delta \in \mathbb{R}^{D_g \times k}$.

2.5. Uncertainty-Aware Multimodal Fusion (UAMF)

Given the modality-wise variance vectors $\{\sigma^\delta\}$, we first quantify the modality confidence based on relative uncertainty ratios:

$$\mathbf{C}_{\text{raw}}^\delta = \mathbf{1} - \frac{\sigma^\delta + \epsilon}{\sigma^a + \sigma^v + \sigma^t + \epsilon}, \delta \in \{t, a, v\}, \quad (5)$$

where ϵ is a small constant for numerical stability, and the operations are performed element-wise along the feature dimension. This formulation assigns higher confidence to modalities with lower relative uncertainty. Softmax normalization is applied along the modality dimension to ensure fair competition:

$$\mathbf{C}^\delta = \text{softmax}(\mathbf{C}_{\text{raw}}^\delta) \in \mathbb{R}^{D_g \times k}. \quad (6)$$

Then, the confidence-weighted features are obtained by element-wise multiplication with the original mean features:

$$\mathbf{R}_u^\delta = \mathbf{C}^\delta \odot \mu^\delta, \delta \in \{t, a, v\}, \quad (7)$$

where $\odot$ denotes element-wise multiplication, enabling automatic suppression of high-uncertainty features.

To model inter-utterance contextual dependencies, we employ a multi-head attention mechanism. The query, key, and value matrices are constructed as:

$$\begin{aligned} \mathbf{Q} &= \mathbf{W}^Q \text{Concat}(\mu^t, \mu^a, \mu^v) \in \mathbb{R}^{D_C \times k}, \\ \mathbf{K} &= \text{Concat}(\mathbf{R}_u^t, \mathbf{R}_u^a, \mathbf{R}_u^v) \in \mathbb{R}^{D_C \times k}, \\ \mathbf{V} &= \mathbf{W}^V \text{Concat}(\mu^t, \mu^a, \mu^v) \in \mathbb{R}^{D_C \times k}, \end{aligned} \quad (8)$$

where $\mathbf{W}^Q$ and $\mathbf{W}^V$ are learnable projections and $D_C = 3D_g$. The uncertainty-aware attention output is computed as:

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}^\top \mathbf{K}}{\sqrt{D_C}}\right) \mathbf{V} \in \mathbb{R}^{D_C \times k}, \quad (9)$$

allowing high-confidence modalities to receive larger attention weights. The attended representations are further refined by a Transformer encoder to capture global conversational context:

$$\mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_k] = \text{TransEncoder}(\mathbf{A}), \quad (10)$$

a residual connection is applied to preserve the original multimodal information and improve training stability:

$$\mathbf{F} = [\mathbf{f}_1, \dots, \mathbf{f}_k] = \mathbf{H} + \text{Concat}(\mu^t, \mu^a, \mu^v), \quad (11)$$

where $\mathbf{f}_i \in \mathbb{R}^C$ denotes the fused feature of utterance u_i . Each fused embedding $\mathbf{f}_i$ is fed into a classifier to obtain emotion probabilities $\mathbf{p}_i$, optimized using the standard cross-entropy loss $\mathcal{L}_{CE}$.

2.6. Explicitly Supervised Loss (ESL)

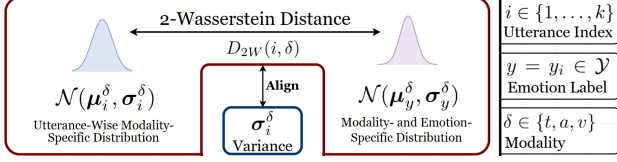


Figure 3: *Framework of explicitly supervised loss $\mathcal{L}_{ESL}^\delta$. ESL aims to align the predicted variance σ_i^δ of each utterance with the 2-Wasserstein distance $D_{2W}(i, \delta)$ between the utterance-wise distribution and the corresponding global emotion center, providing explicit supervision for uncertainty estimation.*

We hypothesize that implicit uncertainty modeling, based solely on the cross-entropy loss $\mathcal{L}_{CE}$, is insufficient to accurately capture utterance-level uncertainty in multimodal emotion recognition. To address this limitation, we propose a novel explicitly supervised loss $\mathcal{L}_{ESL}$, as shown in Fig. 3, which directly supervises the variance outputs from the ContextDEM modules, explicitly encouraging the model to learn accurate and well-calibrated uncertainty representations across modalities.

Alignment with Global Distributional Cluster Centers:

To maintain consistency with the evolving feature representations, global cluster centers are updated once per training epoch using the current feature statistics and are not involved in gradient backpropagation. Unlike conventional methods that only store cluster means, we adopt a distributional representation of cluster centers to better capture the statistical properties. For each emotion class y and modality δ , we compute both the mean and variance of the cluster center:

$$\begin{aligned}\mu_y^\delta &= \frac{1}{|\mathcal{I}_y^\delta|} \sum_{i_y \in \mathcal{I}_y^\delta} \mu_{i_y}^\delta \in \mathbb{R}^{D_g}, \delta \in \{t, a, v\}, \\ \sigma_y^\delta &= \frac{1}{|\mathcal{I}_y^\delta|} \sum_{i_y \in \mathcal{I}_y^\delta} \sigma_{i_y}^\delta \in \mathbb{R}^{D_g}, \delta \in \{t, a, v\},\end{aligned}\quad (12)$$

where $\mathcal{I}_y^\delta = \{i_y \mid y_{i_y} = y \in \mathcal{Y}\}$ denotes the set of utterance indices with emotion label y , D_g is the feature dimension, $\mu_{i_y}^\delta$ and $\sigma_{i_y}^\delta$ are the mean and variance vectors obtained from the ContextDEM modules, and $|\mathcal{I}_y^\delta|$ is the number of utterances with emotion class y . For a dataset with 3 modalities and 6 emotion labels, 18 global distributional cluster centers are maintained.

To measure the deviation between an utterance’s distributional representation and its emotion cluster center, we employ the 2-Wasserstein (2W) distance:

$$D_{2W}(i, \delta) = \|\mu_i^\delta - \mu_{y_i}^\delta\|_2^2 + \|\sigma_i^\delta - \sigma_{y_i}^\delta\|_2^2, \quad (13)$$

where y_i is the emotion label of utterance i , and $\mu_{y_i}^\delta$ and $\sigma_{y_i}^\delta$ denote the mean and variance vectors of the cluster center corresponding to emotion $y_i \in \mathcal{Y}$ and modality δ . The 2-Wasserstein distance can naturally account for the distributional uncertainty. The explicit supervision loss is then formulated as:

$$\mathcal{L}_{ESL}^\delta = \frac{1}{k} \sum_{i=1}^k \left\| \alpha \sqrt{\sigma_i^\delta} - \sqrt{D_{2W}(i, \delta)} \right\|_2^2, \quad (14)$$

where k is the number of utterances in the conversation, and α is a learnable uncertainty scaling factor that addresses potential scale mismatches between the predicted variance and the embedding deviations.

2.7. Jointly Training

The final training objective combines the classification loss and the explicitly supervised uncertainty loss for all three modalities:

$$\mathcal{L} = \mathcal{L}_{CE} + \kappa(\mathcal{L}_{ESL}^t + \mathcal{L}_{ESL}^a + \mathcal{L}_{ESL}^v), \quad (15)$$

where κ denotes the weighting coefficient for the explicitly supervised uncertainty loss. To stabilize training, the explicitly supervised loss $\mathcal{L}_{ESL}$ is introduced after an initial training phase. Specifically, the weight κ remains 0 before a predefined start epoch ep_{start} , and is then set to a fixed constant λ :

$$\kappa(\text{epoch}) = \begin{cases} 0, & \text{if epoch} < ep_{start}, \\ \lambda, & \text{otherwise.} \end{cases} \quad (16)$$

3. Experimental Settings

3.1. Datasets and Evaluation Metric

IEMOCAP [23] is a widely used dataset for emotion recognition in conversation. For partitioning the data, we utilized the commonly used “Leave-One-Session-Out” (LOSO) strategy. **MELD** [24] is a multi-modal, multi-speaker conversational dataset derived from the TV series “Friends”. To ensure a fair comparison, we followed the predefined train/val/test splits provided by the dataset, ensuring data allocation aligned with [25]. Similar to other recent studies, we adopt the accuracy and weighted F1-score (w-F1) for evaluation.

3.2. Implementation Details

For the feature extraction, we used RoBERTa [26] for text, Wav2vec2.0 [27] for audio, and CLIP [28] for visual modality. All experiments were conducted on a single NVIDIA RTX 4090 GPU, trained for 40 and 50 epochs on IEMOCAP and MELD with a batch size of 15 and 100, respectively. The model was optimized using the Adam optimizer with a learning rate of 2×10^{-4} , loss weight of $\lambda = 2 \times 10^{-5}$ with the starting epoch $ep_{start} = 10$, and a dropout rate of 0.2. Finally, we averaged the weights from the last 10 checkpoints to obtain the final model for evaluation.

4. Experimental Results and Analysis

4.1. Effect of Different Fusion Methods and Overall Results

Table 1 shows the detailed results on both datasets. The proposed EmoEUS performs the best among all models, demonstrating its effectiveness. By explicitly modeling uncertainty through $\mathcal{L}_{ESL}$, EmoEUS effectively identifies and emphasizes reliable modalities while suppressing uncertain ones, leading to more robust multimodal representations.

While EmoEUS shows slightly lower scores on the *Sad* and *Angry* classes compared to some baselines, it exhibits significantly more balanced and robust performance across all emotion categories (achieving w-F1 scores above 70% for every emotion class on IEMOCAP), avoiding the severe bias towards specific classes observed in other methods and thereby leading to superior overall accuracy and w-F1 scores.

Table 1: Accuracy and weighted-average F1 score (w-F1) compared with other baselines. Best and second-best results are displayed in **bold** and underline, respectively.

Baseline Model	Proposed Year	IEMOCAP								MELD	
		Happy	Sad	Neutral	Angry	Excited	Frustrated	Acc	w-F1	Acc	w-F1
DialogRNN [29]	2019	32.20	80.26	57.89	62.82	73.87	59.76	63.52	62.89	60.31	57.66
DialogGCN [30]	2019	42.75	84.54	63.54	64.19	63.08	66.99	65.25	64.18	-	58.10
MMGCN [31]	2021	45.14	77.16	64.36	68.82	74.71	61.40	66.36	66.26	60.42	58.31
M2FNet [6]	2022	-	-	-	-	-	-	69.69	69.86	<u>67.85</u>	66.71
CFN-ESA [7]	2023	53.67	80.60	<u>71.65</u>	70.32	74.82	68.06	71.04	70.78	<u>67.85</u>	66.70
AdaIGN [32]	2024	53.04	81.47	71.26	65.87	76.34	67.79	-	70.74	-	<u>66.79</u>
MDAG [33]	2024	45.26	<u>81.40</u>	69.53	70.33	71.61	66.94	69.11	69.08	64.41	64.00
DER-GCN [34]	2025	58.80	79.80	61.50	72.10	73.30	67.80	69.70	69.40	66.80	66.10
FEMI [35]	2025	60.60	85.55	70.92	70.98	75.13	<u>69.00</u>	71.97	<u>73.53</u>	64.88	66.41
EmoEUS	Ours	77.42	72.87	72.04	<u>71.20</u>	78.66	73.58	74.33	74.36	68.32	67.53

Table 2: Performance of using different model settings.

Fusion Method	IEMOCAP		MELD	
	Acc	w-F1	Acc	w-F1
Concat	71.90	72.20	66.82	65.57
Attention	72.39	72.33	66.64	65.75
Transformer	72.53	72.57	67.64	66.58
EmoEUS (Full)	74.33	74.36	68.32	67.53
w/o $\mathcal{L}_{ESL}$	73.32	73.33	67.78	66.73
w/o UAMF	72.63	72.69	66.53	66.01

Table 2 presents the performance of different fusion methods. ‘‘Concat’’ directly concatenates features ($\mathbf{R}^\delta$) from different modalities and feeds them to the classifier. ‘‘Attention’’ and ‘‘Transformer’’ employ a multi-head attention module and a simple Transformer encoder to fuse the concatenated multimodal features, respectively. Compared with these baselines, EmoEUS consistently achieves better performance on both datasets, demonstrating the effectiveness of the uncertainty-aware fusion mechanism. Performance drops in the variants without UAMF and $\mathcal{L}_{ESL}$ further verify the importance of these components. Removing UAMF degrades performance, showing that uncertainty-aware fusion is crucial for multimodal integration. The variant without $\mathcal{L}_{ESL}$ also declines, suggesting that explicit uncertainty supervision helps learn more accurate and well-calibrated uncertainty estimates, which further facilitate more effective uncertainty-aware fusion.

4.2. Performance on Ambiguous Emotion Pairs

The results in Table 3 demonstrate that the proposed EmoEUS effectively reduces misclassification rates between ambiguous emotion pairs (e.g., *Hap-Exc* denotes misclassifying ‘‘Happy’’ as ‘‘Excited’’), thereby improving overall emotion prediction performance. Notably, the full model with explicit uncertainty supervision $\mathcal{L}_{ESL}$ achieves further reductions in misclassification rates for most ambiguous pairs, which enables the model to learn more separable and discriminative representations for semantically similar emotions. By directly supervising the variance outputs and aligning them with emotion-specific distributional cluster centers, the model better captures fine-grained uncertainty patterns, thereby leading to more robust overall prediction performance.

4.3. Ablation Studies on ContextDEM

Table 4 demonstrates the advantages of distributional representation over point estimation. Modeling both mean (μ^δ) and variance (σ^δ) with the 2-Wasserstein distance consistently out-

Table 3: Misclassification Rates Reduction on the Ambiguous Emotion Pairs in IEMOCAP (**Lower is Better**).

Methods	Hap-Neu	Hap-Exc	Neu-Exi	Sad-Fru	Exc-Hap
Transformer	7.68	16.74	5.41	24.62	7.54
w/o $\mathcal{L}_{ESL}$ *	7.57	19.05	5.23	16.92	6.56
EmoEUS (Full)	6.51	14.88	4.96	16.92	6.89

*Train without the explicit uncertainty supervision loss $\mathcal{L}_{ESL}$

performs single-point representation with the standard MSE distance across all settings, validating that explicitly capturing uncertainty through variance leads to more robust feature learning. Furthermore, the residual connection proves effective in preserving the original contextual features while integrating the global interactions from the Transformer encoder.

Table 4: Effect of Distributional Representation Modeling and Residual Connection on ContextDEM. For comparison, we employed an MSE distance for point representation using (μ^δ).

	Residual Connection	IEMOCAP		MELD	
		Acc	w-F1	Acc	w-F1
Point-Rep (μ^δ) with MSE Distance D_{MSE}					
(μ^δ)	×	72.86	72.42	67.53	67.45
(μ^δ)	✓	73.28	72.47	66.98	67.26
Distribution-Rep ($\mu^\delta, \sigma^\delta$) with 2-Wasserstein Distance D_{2W}					
$(\mu^\delta, \sigma^\delta)$	×	73.98	73.89	68.08	67.31
$(\mu^\delta, \sigma^\delta)$	✓	74.33	74.36	68.32	67.53

5. Conclusions and Future Work

We propose the EmoEUS, a novel framework that incorporates explicit uncertainty supervision for MERC. Our approach introduces a context-level distribution estimator module (ContextDEM) to model utterance-level distribution, dynamically weights modality reliability through an uncertainty-aware multimodal fusion (UAMF) mechanism, and supervises the uncertainty estimation using an explicitly supervised loss that aligns predicted variances with distributional deviations. Extensive experiments on IEMOCAP and MELD demonstrate state-of-the-art performance, with significant reductions in misclassification rates between ambiguous emotion pairs, validating the effectiveness of explicit uncertainty supervision. In future work, we will enhance the uncertainty estimation module to better capture cross-speaker emotional dynamics in real-time settings.

6. Acknowledgment

The work presented in this article is supported by the Research Platform for Advanced Audio and Speech Signal Processing (P0049192) funded by Innovation Technology Co. Ltd.

7. Use of Generative AI Disclosure

Generative AI tools were used only for language polishing and formatting assistance. All scientific content, experiments and analyses were produced and verified by the authors.

8. References

- [1] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE Access*, vol. 7, pp. 100 943–100 953, 2019.
- [2] Z. Huang, M.-W. Mak, and K. A. Lee, "MM-NodeFormer: Node transformer multimodal fusion for emotion recognition in conversation," in *Proc. Interspeech*, 2024, Conference Proceedings, pp. 4069–4073.
- [3] M. Glodek, S. Tschechne, G. Layher, M. Schels, T. Brosch, S. Scherer, M. Kächele, M. Schmidt, H. Neumann, G. Palm *et al.*, "Multiple classifier systems for the classification of audio-visual emotional states," in *Affective Computing and Intelligent Interaction: Fourth International Conference, ACHI 2011, Memphis, TN, USA, October 9–12, 2011, Proceedings, Part II*. Springer, 2011, pp. 359–368.
- [4] T. Shi and S.-L. Huang, "MultiEMO: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 14 752–14 766. [Online]. Available: <https://aclanthology.org/2023.acl-long.824/>
- [5] W. Ren, K. Li, R. Chao, J. Li, Z. Huang, S. Ahmed, Y.-J. Li, K.-H. Hung, S.-S. Wang, H.-M. Wang *et al.*, "BAV-MossFormer2: Enhanced MossFormer2 for binaural audio-visual speech enhancement," in *Proc. AVSEC 2025*, 2025, pp. 79–80.
- [6] V. Chudasama, P. Kar, A. Gudmalwar, N. Shah, P. Wasnik, and N. Onoe, "M2FNET: Multi-modal fusion network for emotion recognition in conversation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4652–4661.
- [7] J. Li, Y. Liu, X. Wang, and Z. Zeng, "CFN-ESA: A cross-modal fusion network with emotion-shift awareness for dialogue emotion recognition," *arXiv preprint arXiv:2307.15432*, 2023.
- [8] C.-X. Gan, Z. Li, Z. Jin, Z. Huang, M.-W. Mak, and K. A. Lee, "IDIR: Identifying and Distilling Informative Relations for Speaker Verification," in *Interspeech 2025*, 2025, pp. 5758–5762.
- [9] J. Li, R. Yu, H. Huang, and H. Yan, "Unimodal-driven distillation in multimodal emotion recognition with dynamic fusion," in *2025 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2025, pp. 1–6.
- [10] Z. Jin, S. Liu, Z. Li, C.-X. Gan, Z. Huang, M.-W. Mak, and K. A. Lee, "Distilling attention knowledge for speaker verification," in *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026, pp. 16 447–16 451.
- [11] C.-X. Gan, P. Bell, M.-W. Mak, Z. Li, Z. Jin, Z. Huang, and K. A. Lee, "UNET-based fusion and exponential moving average adaptation for noise-robust speaker recognition," *arXiv preprint arXiv:2604.25624*, 2026.
- [12] Z. Jin, Y. Tu, Z. Li, Z. Huang, C.-X. Gan, and M.-W. Mak, "Denosing student features with diffusion models for knowledge distillation in speaker verification," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.
- [13] C.-H. Wu, J.-C. Lin, and W.-L. Wei, "Survey on audiovisual emotion recognition: Databases, features, and data fusion strategies," *APSIPA Transactions on Signal and Information Processing*, vol. 3, p. e12, 2014.
- [14] H. Ma, J. Wang, H. Lin, B. Zhang, Y. Zhang, and B. Xu, "A transformer-based model with self-distillation for multimodal emotion recognition in conversations," *IEEE Transactions on Multimedia*, vol. 26, pp. 776–788, 2023.
- [15] X. Wang, S. Zhao, H. Sun, H. Wang, J. Zhou, and Y. Qin, "Enhancing multimodal emotion recognition through multi-granularity cross-modal alignment," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.
- [16] H. Sun, S. Zhao, S. Li, X. Kong, X. Wang, J. Zhou, A. Kong, Y. Chen, W. Zeng, and Y. Qin, "Enhancing emotion recognition in incomplete data: A novel cross-modal alignment, reconstruction, and refinement framework," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.
- [17] Z. Li, M.-W. Mak, J.-T. Chien, M. Pilanci, Z. Jin, and H. Meng, "Disentangling speech representations learning with latent diffusion for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 3896–3907, 2025.
- [18] Y. Shi and A. K. Jain, "Probabilistic face embeddings," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6902–6911.
- [19] J. Chang, Z. Lan, C. Cheng, and Y. Wei, "Data uncertainty learning in face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5710–5719.
- [20] Y. Ji, J. Wang, Y. Gong, L. Zhang, Y. Zhu, H. Wang, J. Zhang, T. Sakai, and Y. Yang, "Map: Multimodal uncertainty-aware vision-language pre-training model," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 23 262–23 271.
- [21] T. Meng, F. Zhang, Y. Shou, H. Shao, W. Ai, and K. Li, "Masked graph learning with recurrent alignment for multimodal emotion recognition in conversation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.
- [22] M. K. Tellamekala, S. Amiriparian, B. W. Schuller, E. André, T. Giesbrecht, and M. Valstar, "COLD Fusion: Calibrated and ordinal latent distribution fusion for uncertainty-aware multimodal emotion recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 2, pp. 805–822, 2024.
- [23] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.
- [24] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," *arXiv preprint arXiv:1810.02508*, 2018.
- [25] D. Hu, X. Hou, L. Wei, L. Jiang, and Y. Mo, "MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 7037–7041.
- [26] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
- [27] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.
- [28] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. International Conference on Machine Learning*, 2021, pp. 8748–8763.

- [29] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "DialogueRNN: An attentive rnn for emotion detection in conversations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 6818–6825.
- [30] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "DialogueGCN: A graph convolutional neural network for emotion recognition in conversation," *arXiv preprint arXiv:1908.11540*, 2019.
- [31] J. Hu, Y. Liu, J. Zhao, and Q. Jin, "MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation," *arXiv preprint arXiv:2107.06779*, 2021.
- [32] G. Tu, T. Xie, B. Liang, H. Wang, and R. Xu, "Adaptive graph learning for multimodal conversational emotion detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, 2024. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/29876>
- [33] C.-B. Nguyen, D.-T. Le, Q. T. Ha *et al.*, "Curriculum learning meets directed acyclic graph for multimodal emotion recognition," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 4259–4265.
- [34] W. Ai, Y. Shou, T. Meng, and K. Li, "DER-GCN: Dialog and event relation-aware graph convolutional neural network for multimodal dialog emotion recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 3, pp. 4908–4921, 2025.
- [35] Y. Fu, X. Yan, W. Chen, and J. Zhang, "Feature-enhanced multimodal interaction model for emotion recognition in conversation," *Knowledge-Based Systems*, vol. 309, p. 112876, 2025.