Multimodal Emotion Recognition Calibration in Conversations

Geng Tu*
Harbin Institute of Technology
Shenzhen, China
22b951011@stu.hit.edu.cn

Feng Xiong*
Harbin Institute of Technology
Shenzhen, China
23s151006@stu.hit.edu.cn

Bin Liang
The Chinese University of Hong Kong
Hong Kong, China
bin.liang@cuhk.edu.hk

Hui Wang
Peng Cheng Laboratory
Shenzhen, China
wangh06@pcl.ac.cn

Xi Zeng
The 30th Research Institute of China
Electronics Technology Group
Corporation
Chengdu, China
zxmm2@163.com

Ruifeng Xu†
Harbin Institute of Technology
Shenzhen, China
GD Prov. Key Lab. of Novel Security
Intelligence Technologies
Shenzhen, China
Peng Cheng Laboratory
Shenzhen, China
xuruifeng@hit.edu.cn

ABSTRACT

Multimodal Emotion Recognition in Conversations (MERC) aims to identify the emotions conveyed by each utterance in a conversational video. Current efforts focus on modeling speaker-sensitive context dependencies and multimodal fusion. Despite the progress, the reliability of MERC methods remains largely unexplored. Extensive empirical studies reveal that current methods suffer from unreliable predictive confidence. Specifically, in some cases, the confidence estimated by these models increases when a modality or specific contextual cues are corrupted, defining these as uncertain samples. This contradicts the foundational principle in informatics, namely, the elimination of uncertainty. Based on this, we propose a novel calibration framework CMERC to calibrate MERC models without altering the model structure. It integrates curriculum learning to guide the model in progressively learning more uncertain samples; hybrid supervised contrastive learning to refine utterance representations, by pulling uncertain samples and others apart; and confidence constraint to penalize the model on uncertain samples. Experimental results on two datasets demonstrate the effectiveness and generalization capabilities of our CMERC across various MERC models, surpassing state-of-the-art methods.

CCS CONCEPTS

• **Information systems** → **Multimedia information systems**; **Sentiment analysis**; **Clustering and classification**; • **Computing methodologies** → *Discourse, dialogue and pragmatics.*

KEYWORDS

Multimodal Conversational Emotion Recognition, Confidence Calibration, Curriculum Learning, Contrastive Learning

ACM Reference Format:
Geng Tu, Feng Xiong, Bin Liang, Hui Wang, Xi Zeng, and Ruifeng Xu. 2024. Multimodal Emotion Recognition Calibration in Conversations. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28–November 1, 2024, Melbourne, VIC, AustraliaProceedings of the 32nd ACM International Conference on Multimedia (MM'24), October 28-November 1, 2024, Melbourne, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3664647.3681515

1 INTRODUCTION

Emotion Recognition in Conversations (ERC) poses significant challenges due to conversations' dynamic and spontaneous nature, where individuals express a range of emotions [67]. Conventional ERC approaches primarily depend on textual data [66], but textual information alone often falls short in capturing emotional nuances [15]. Consequently, Multimodal ERC (MERC), which integrates audio and visual signals in addition to text, is attracting more research interest [63].

Current MERC research in dialogue modeling follows the ERC paradigm, requiring the modeling of both context-sensitive and speaker-sensitive dependencies [30]. This is achieved using recurrent-based networks [10, 34], transformer-based networks [30, 67], and graph-based networks [40, 66]. For multimodal fusion, significant focus is placed on aggregation-based fusion methods such as concatenation [15, 56] and attention networks [50, 63], as well as heterogeneous graph-based fusion approaches [4, 18, 20].

Despite the above progress, the reliability of current MERC methods remains largely unexplored. In classification settings, a crucial aspect of reliability involves developing a robust confidence estimator [6, 33, 37] that accurately quantifies the probability of correct predictions. Such an estimator proves particularly valuable in high-stakes situations [13]. In MERC, alongside the precise overall prediction confidence, it is crucial to consider the correlation between confidence and modalities or contextual cues, both intra-speaker (within the same speaker) and inter-speaker (between different

[*] Both authors contributed equally to this research.
[†] Corresponding author.

MM '24, October 28–November 1, 2024, Melbourne, VIC, Australia.

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3681515

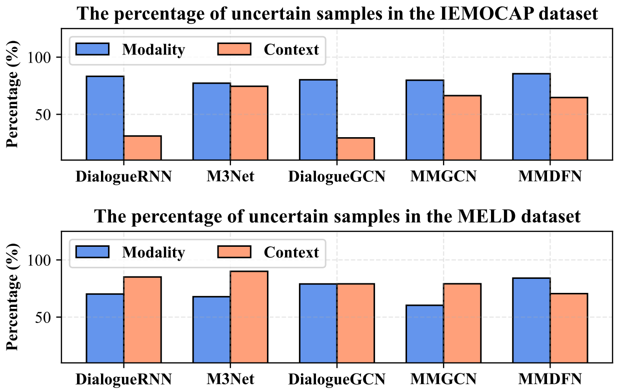

Figure 1: Percentage of uncertain samples in the testing sets of two datasets stems from removed modalities or contexts.

speakers) contexts. Intuitively, the confidence of the target category should not increase when a modality or specific contextual cues (intra- or inter-speaker contexts) are removed in a MERC model, as the observed information becomes less comprehensive.

However, empirical studies on existing methods reveal a counterintuitive trend – in some cases, the removal of certain modalities or specific contextual cues can lead to an increase in confidence as illustrated in Fig. 1, defining these as uncertain samples. Notably, the advanced models like M3Net [4] can underperform compared to simpler models like DialogueRNN [34], which suggests that current MERC research prioritizes performance metrics over the reliability of predictive confidence. This contradicts the fundamental principle in informatics that "the essence of information is to eliminate uncertainty" [2]. This further hampers the reliability of models, making them susceptible to influence when a modality or specific contextual cues are corrupted, as they lack a trustworthy confidence estimator for decision-making. To delve into the underlying reasons, M3Net [4] is examined as an example. The analysis reveals that its contradiction with the fundamental principle in informatics arises from its excessive focus on the textual modality and its struggle to balance the impact of different contexts.

To address the above issue, potential solutions such as temperature scaling [11], Bayesian learning [5, 23], etc., offer global calibration of predicted confidences. However, these methods cannot explicitly calibrate MERC models across various modalities and contexts. Therefore, we propose the CMERC, a framework for calibrating them without altering the model structure, which explores three pivotal calibrations. **(1) Calibrating the training strategy:** Intuitively, the MERC model struggles to make decisions based on reliable predictive confidence for these uncertain samples, which hampers the learning process. Our CMERC employs Curriculum Learning (CL) [1] to guide the model in progressively learning more uncertain samples that contradict the foundational principle in informatics, i.e., the removal of a modality or specific contextual cues enhances the model's confidence, thus aiding in the model's learning. **(2) Calibrating utterance representations:** If the MERC model fails to learn essential features relevant to violations of fundamental informatics principles, reliable confidence estimation becomes difficult. Understanding the factors behind uncertain samples is vital for improving training reliability. We define uncertain samples as positive and others as negative. Hybrid Supervised Contrastive Learning (SCL) is utilized to distinguish uncertain samples from others, facilitating the model in capturing the correlation and difference between uncertain samples and the rest. This process allows the model to identify factors contributing to uncertainty. **(3) Calibrating the loss function:** Introducing a Confidence Constraint aims to constrain unexpected surges in predicted confidences directly. The main contributions of this paper can be summarized as follows:

- We have conducted extensive empirical studies revealing that existing MERC methods fail to provide reliable confidence estimation for decision-making.
- We propose the CMERC framework to calibrate MERC models for the first time, focusing on calibrating the training strategy, utterance representations, and the loss function.
- Experimental results demonstrate the effectiveness and generalization capabilities of our CMERC across various MERC models, surpassing state-of-the-art methods.

2 RELATED WORK

Emotion Recognition in Conversations: The rapid growth of social media has led to the availability of extensive interaction data, including datasets like IEMOCAP [3] and MELD [43]. Consequently, ERC has attracted substantial research interest. Unlike typical emotion classification tasks focused on individual sentences, in conversations, the speaker's emotions are not only influenced by their internal state but also by their interactions with other speakers [10]. Therefore, the ERC model requires modeling both context-sensitive and speaker-sensitive dependencies [54]. This includes approaches such as recurrent-based networks [19, 27, 34], transformer-based networks [22, 30, 47], and graph-based networks [10, 48, 55].
Multimodal Emotion Recognition in Conversations: Traditional ERC focuses on text, while MERC integrates modalities such as audio and visual cues along with text to better capture conveyed emotions [61]. This integration makes modeling the dependencies among modalities a significant challenge. Recent research in MERC has focused on using speaker embeddings to model speaker-sensitive dependencies [4, 18, 20]. For context-sensitive dependencies, approaches can be categorized into two types: those based on previous text-based modeling paradigms [29, 49] and those exploring intra- and cross-modal interactions to capture contextual clues [4, 20, 50]. For multimodal fusion, MERC aims to combine information from different modalities, including aggregation-based methods like concatenation [15, 56] and attention networks [45, 58]. However, aggregation-based fusion methods overlook the complex interactions between modalities, resulting in insufficient utilization of contextual information [18]. Recently, researchers have explored graph-based fusion methods to capture intra- and inter-modal interactive information [4, 18, 20, 40, 66]. Despite the progress, the reliability of these methods remains largely unexplored.
Uncertainty estimation: Uncertainty estimation is crucial for reliable predictions [33]. Various models have been proposed to address uncertainty, including Bayesian neural networks [7, 24], Dropout [36], and Deep ensembles [14, 26]. In classification tasks,

prediction confidence is essential, but models often exhibit overconfidence due to the rapid growth of softmax probabilities [17]. To mitigate this, methods have been developed to calibrate confidence scores to reflect predictive uncertainty [35]. Some approaches aim to train well-calibrated models directly [16, 28, 31, 60, 62, 64, 68], employing techniques like mixup [53], label smoothing [39], and focal loss [37, 38, 69]. Others rely on post-processing methods for calibration [12, 25, 42, 44], with temperature scaling [11] being a prominent example, adjusting probabilities using a single scalar parameter. However, these methods fail to consider the relationship between various modalities or contexts, solely adjusting overall confidence without specific calibration for individual modalities or contexts. This limits their effectiveness in the MERC task.

3 METHODOLOGY

In this section, we offer a comprehensive introduction to each component of the CMERC framework as illustrated in Fig. 2.

3.1 Task Definition

Let $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_\mathrm{N}]$ be a conversation uttered by $\mathrm{M} \geq 2$ speakers, consisting of N utterances. Each utterance $\mathbf{u}_k$ is represented by a triplet $\mathbf{x}_k = \{\mathbf{x}_k^\mathrm{A}, \mathbf{x}_k^\mathrm{V}, \mathbf{x}_k^\mathrm{T}\}$. $\mathbf{x}_k^\mathrm{A} \in \mathbb{R}^{d_a}$, $\mathbf{x}_k^\mathrm{V} \in \mathbb{R}^{d_v}$, and $\mathbf{x}_k^\mathrm{T} \in \mathbb{R}^{d_t}$ denote the acoustic, visual, and textual features of $\mathbf{u}_k$, respectively. MERC aims to predict the emotion label of each utterance $\mathbf{u}_k$ according to its context $\mathbf{c}_k = \{c_k^\mathrm{X}, c_k^\mathrm{O}\}$. c_k^X and c_k^O are the intra-speaker and inter-speaker contexts of the utterance $\mathbf{u}_k$.

3.2 Feature Representation

Following Ghosal et al. [9], we employ layer normalization and average operation on the last four hidden layers of the Roberta model [32] to obtain textual features. For acoustic and visual feature extraction, following Hu et al. [18], Wen et al. [59], we utilize OpenSmile [8] and a pre-trained DenseNet model [21], respectively.

3.3 Overview

Considering the unreliable emotional inferences in MERC models, we propose a CMERC framework to calibrate them, as shown in Fig. 2, which integrates three key calibrations: First, it employs CL to progressively train the model on uncertain samples, aiding in the model's learning. Secondly, Hybrid SCL pulls uncertain samples and others further apart, reinforcing the model's focus on factors contributing to uncertainty. Finally, introducing a Confidence Constraint to penalize the model on uncertain samples, ensures a trustworthy confidence estimator for decision-making.

3.4 Curriculum Learning for MERC

To design the curriculum for MERC models, we gauge the difficulty of each conversation across different modalities and contexts by measuring confidence levels after the removal operation. Intuitively, the MERC model faces challenges in making decisions based on reliable predictive confidence for uncertain samples, thereby hindering the learning process. The greater the number of uncertain samples within a conversation, the more difficult it becomes, as the model's confidence becomes increasingly unreliable, making it more challenging to grasp the emotion of the utterances.

Algorithm 1: Training process of CMERC using audio and intra-speaker context.

```
Input: Dataset D; the number of buckets $\mathrm{N}_d$; the difficulty
       measurer $\mathrm{DF}(.)$.
Output: $\mathcal{L}_m$, $\mathcal{L}_c$, $\mathcal{L}_s$.
▷ Curriculum Learning in the CMERC.
$\hat{\mathrm{D}} = \{\hat{\mathcal{D}}_i\}_{i=1}^{\mathrm{N}_d} \leftarrow sort(\mathrm{D}, \mathrm{DF})$;   $\mathcal{D}_{train} = \varnothing$
for $i = 1$ to $\mathrm{N}_d$ do
    $\mathcal{D}_{train} = \mathcal{D}_{train} \cup \hat{\mathcal{D}}_i$
    ▷ Traversing $\mathcal{D}_{train}$ with $\mathrm{N}_b$ mini-batches.
    for $\mathbf{d} = \{\mathbf{d}_1, ..., \mathbf{d}_{\mathrm{N}_b}\}$ do
        $\{\mathbf{x}_k^{\{\mathrm{A,V,T}\}}, \mathbf{c}_k^{\{\mathrm{X,O}\}}\}_{k=1}^{\mathcal{B}} \leftarrow \mathbf{d}$;   $\Gamma_s \leftarrow 0$
        ▷ Hybrid SCL in the CMERC.
        $\mathbf{z}^\mathrm{A}, \mathbf{z}^\mathrm{X} \leftarrow \{0, 0\}_{k=1}^{\mathcal{B}}$
        for $k = 1$ to $\mathcal{B}$ do
            $\mathbf{o}_k, \mathbf{H}_k \leftarrow \mathcal{M}(\mathbf{x}_k^{\{\mathrm{A,V,T}\}}, \mathbf{c}_k^{\{\mathrm{X,O}\}})$
            $\mathbf{o}_k^\mathrm{A} \leftarrow \mathcal{M}(\mathbf{x}_k^{\{[mask],\mathrm{V,T}\}}, \mathbf{c}_k^{\{\mathrm{X,O}\}})$
            $\mathbf{o}_k^\mathrm{X} \leftarrow \mathcal{M}(\mathbf{x}_k^{\{\mathrm{A,V,T}\}}, \mathbf{c}_k^{\{[mask],\mathrm{O}\}})$
            ▷ Pseudo labeling for the utterance $\mathbf{u}_k$.
            for $\xi = \{\mathrm{A}, \mathrm{X}\}$ do
                if $\mathbf{o}_k[\varphi] < \mathbf{o}_k^\xi[\varphi]$ then
                    $\mathbf{z}_k^\xi \leftarrow 1$;   $\Gamma_s$ += $\mathbf{o}_k^\xi[\varphi] - \mathbf{o}_k[\varphi]$
        $\Gamma_\mathrm{A}, \Gamma_\mathrm{X} \leftarrow 0, 0$;   $\ell_{(+)}^\mathrm{A}, \ell_{(-)}^\mathrm{A}, \ell_{(+)}^\mathrm{X}, \ell_{(-)}^\mathrm{X} \leftarrow [\,], [\,], [\,], [\,]$
        for $n = 1$ to $\mathcal{B}$ and $n \neq m$ do
            for $\xi = \{\mathrm{A}, \mathrm{X}\}$ do
                if $\mathbf{z}_n^\xi == \mathbf{z}_m^\xi$ then
                    $\ell_{(+)}^\xi$ += $\ell(\mathrm{H}_n, \mathrm{H}_m)$
                $\ell_{(-)}^\xi$ += $\ell(\mathrm{H}_n, \mathrm{H}_m)$;   $\Gamma_\xi$ += $\ell_{(+)}^\xi$ / $\ell_{(-)}^\xi$
        ▷ Computing the confidence constraint.
        $\mathcal{L}_s \leftarrow \Gamma_s/\mathcal{B}$
        ▷ Computing the contrastive losses.
        $\mathcal{L}_m, \mathcal{L}_c \leftarrow -\Gamma_\mathrm{A}/\mathcal{B}, -\Gamma_\mathrm{X}/\mathcal{B}$
```

$$\mathbf{DF}(\mathbf{d}_i) = \frac{\mathbf{dF}_{\mathbb{C}}(\mathbf{d}_i) + \mathbf{dF}_{\mathbb{M}}(\mathbf{d}_i) + \mathrm{N}_s(\mathbf{d}_i)}{\mathrm{N}_u(\mathbf{d}_i) + \mathrm{N}_s(\mathbf{d}_i)} \tag{1}$$

$$\mathbf{dF}_{\mathbb{M}}(\mathbf{d}_i) = \sum_{k=1}^{\mathcal{B}} \sum_{\xi}^{\{\mathrm{A,V,T}\}} \mathbf{CF}_k^\xi \tag{2}$$

$$\mathbf{dF}_{\mathbb{C}}(\mathbf{d}_i) = \sum_{k=1}^{\mathcal{B}} \sum_{\xi}^{\{\mathrm{X,O}\}} \mathbf{CF}_k^\xi \tag{3}$$

$$\mathbf{CF}_k^\xi = \mathbf{max}(0, \mathbf{o}_k^\xi[\varphi] - \mathbf{o}_k[\varphi]) \tag{4}$$

where $\mathbf{CF}_k^\xi$ represents the confidence boost when modalities or contexts are removed. $\mathbb{C} = \{\mathbf{X}, \mathbf{O}\}$ denotes intra- and inter-speaker contexts. $\mathbb{M} = \{\mathbf{A}, \mathbf{V}, \mathbf{T}\}$ denotes the acoustic, visual, and textual modalities. $\mathbf{dF}_{\mathbb{M}/\mathbb{C}}(.)$ denotes the modality- or context-specific difficulty measurer. φ is the index of target categories. $\mathbf{o}_k^\xi$ is the predictive distribution of the MERC model $\mathcal{M}$ with a modality or context removed. $\mathcal{B}$ is the size of mini-batch $\mathbf{d}_i$. $\mathrm{N}_u(\mathbf{d}_i)$ represents the total number of utterances in a mini-batch $\mathbf{d}_i$. $\mathrm{N}_s(\mathbf{d}_i)$ is the number of speakers take part in $\mathbf{d}_i$ and it acts as a smoothing factor. We utilize

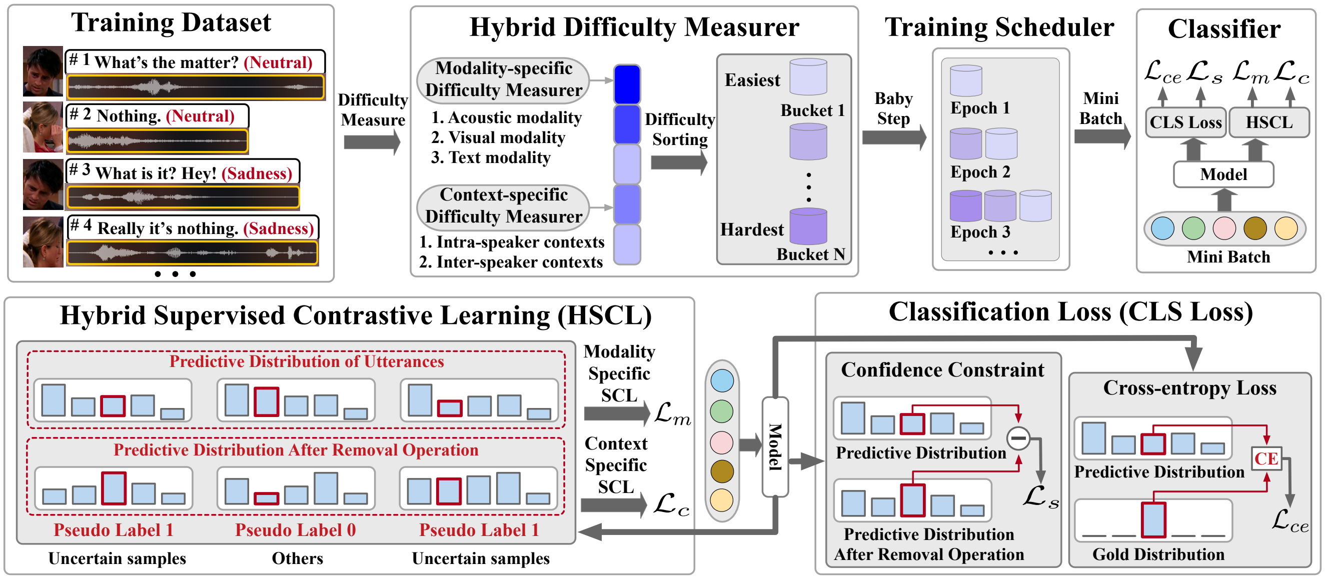

Figure 2: The proposed CMERC framework. Mathematical symbols are consistent with the formulas in the paper.

baby step training scheduler [52] to arrange conversations and organize the training process, described as *Lines* 1 - 4 in Algorithm 1.

3.5 Hybrid Supervised Contrastive Learning

Understanding the factors contributing to uncertain samples is pivotal for bolstering the reliability of MERC models during training. Consequently, we advocate for the adoption of a Hybrid SCL framework, which seamlessly integrates modality- and context-specific SCL components. This approach serves to discern uncertain samples caused by the removal of modalities or contextual cues from their counterparts, thereby capturing their nuanced correlations and distinctions. As a result, the MERC model can adeptly identify the underlying factors driving uncertainty during the training process, thereby enhancing its overall effectiveness.

$$\mathcal{L}_m = -\frac{1}{\mathcal{B}} \sum_{\xi}^{\{\mathrm{A,V,T}\}} \log\left(\Gamma(\mathbf{z}^{\xi})\right) \tag{5}$$

$$\mathcal{L}_c = -\frac{1}{\mathcal{B}} \sum_{\xi}^{\{\mathrm{X,O}\}} \log\left(\Gamma(\mathbf{z}^{\xi})\right) \tag{6}$$

$$\Gamma(\triangle) = \frac{\sum_{j=1}^{\mathcal{B}} \mathbb{1}_{[i\neq j]} \mathbb{1}_{[\triangle_i=\triangle_j]}\, \ell(\mathbf{H}_i,\mathbf{H}_j)}{\sum_{k=1}^{\mathcal{B}} \mathbb{1}_{[i\neq k]}\, \ell(\mathbf{H}_i,\mathbf{H}_k)} \tag{7}$$

where $\mathbf{H}$ denotes the hidden representation of the model $\mathcal{M}$. $\mathbf{z}$ indicates the set of pseudo labels, generated according to the process described in *Lines* 14 - 17 of Algorithm 1. $\ell(\star,\star) = e^{simi(\star,\star)/\tau}$, where τ is the temperature parameter. $simi(\star,\star)$ denotes the cosine similarity function. The calculation process of modality- and context-specific SCL, taking audio and intra-speaker context as an example, is detailed in *Lines* 18 - 27 of Algorithm 1.

3.6 Confidence Constraint

To enhance the reliability of predicted confidences in MERC models, we utilize the difference in confidence increase after removal operations as the regularization constraint for mini-batch $\mathbf{d}_i$.

$$\mathcal{L}_s = \sum_{k=1}^{\mathcal{B}} \sum_{\xi}^{\{\mathbb{M},\mathbb{C}\}} \mathbf{CF}_k^{\xi} \tag{8}$$

3.7 Model Training

We jointly train our proposed framework by minimizing the sum of the following four losses.

$$\mathcal{L} = \mathcal{L}_{ce} + \gamma_m\mathcal{L}_m + \gamma_c\mathcal{L}_c + \gamma_s\mathcal{L}_s + \lambda \left\|\Theta\right\|^2 \tag{9}$$

where γ_m, γ_c, and γ_s are tuned hyperparameters. $\mathcal{L}_{ce}$ denotes the loss function for the MERC task, typically implemented as a cross-entropy loss. Θ is the set of trainable parameters within the CMERC. λ is the coefficient of $\mathbf{L}_2$-regularization.

4 EXPERIMENTS

4.1 Datasets

We evaluate the CMERC on two datasets: **IEMOCAP** [3] has dyadic conversation videos with ten speakers, featuring 7,433 utterances and 151 dialogues. Each utterance has one of six emotions. **MELD** [43] contains multiparty conversations collected from the 'Friends' TV series, having 1,433 conversations, 13,708 utterances, and 304 speakers. Each utterance holds one of seven emotions. Following Ghosal et al. [9], the data splitting for datasets is detailed in Table 1. As the IEMOCAP dataset lacks a predefined train/validation split, we allocate 10% of the training dialogues for validation.

4.2 Experimental Settings

All re-implementation methods have released their source codes, ensuring identical settings as the original papers. For the CMERC, γ_m, γ_c, γ_s, and τ are manually tuned for each dataset using hold-out validation. We adopt M3Net [4] as the **Baseline** in this paper. The reported results are the average score of 5 random runs on the test set. Our experiments are conducted on a single RTX 4090 GPU.

Table 1: Statistics of two conversational datasets.

Dataset	Dialogues			Utterances			Classes
	train	val	test	train	val	test	
MELD	1039	114	280	9,989	1,109	2610	7
IEMOCAP	120		31	5,810		1,623	6

Table 2: Percentage (%) of uncertain samples in the test set. $\mathbb{M}$ represents the removal of any single modality (the union of results from A, V, and T), $\mathbb{C}$ denotes the removal of any single context (the union of results from X and O), and $\mathbb{M} \cup \mathbb{C}$ signifies the union of results from $\mathbb{M}$ and $\mathbb{C}$.

Methods	IEMOCAP							
	A	V	T	X	O	$\mathbb{M}$	$\mathbb{C}$	$\mathbb{M} \cup \mathbb{C}$
DialogueRNN	43.30	45.16	56.01	12.20	27.79	83.24	31.05	88.17
DialogueGCN	41.47	41.10	54.16	27.05	27.85	80.22	29.39	85.52
MMGCN	36.66	35.67	53.91	30.81	49.35	79.85	66.36	95.32
MMDFN	39.13	55.27	51.69	25.26	52.06	85.52	64.70	95.44
M3Net	36.78	47.13	33.46	35.06	49.85	77.26	74.55	91.93

Methods	MELD							
	A	V	T	X	O	$\mathbb{M}$	$\mathbb{C}$	$\mathbb{M} \cup \mathbb{C}$
DialogueRNN	53.56	-	29.96	27.89	64.56	70.19	85.10	93.49
DialogueGCN	57.20	-	35.86	46.01	57.24	79.00	79.08	95.17
MMGCN	33.95	26.05	28.39	51.99	55.86	60.34	79.16	92.61
MMDFN	47.32	60.65	27.43	30.15	55.98	84.10	70.54	93.30
M3Net	30.65	42.11	26.28	60.69	46.09	67.85	90.04	97.43

Evaluation metric: Following Zhang and Li [66], we utilize the weighted F1 score (w-F1) as evaluation metrics and we also report F1 scores per class. To evaluate model prediction reliability, aside from various typically used confidence estimation metrics such as Expected Calibration Error (ECE), Maximum Calibration Error (MCE), Root Mean Square Calibration Error (RMSCE) [11], Area Under the Receiver Operating Characteristic Curve (AUROC), and Area Under the Precision-Recall Curve (AUPRC) [69], we suggest a novel metric called Confidence Enhancement Level (CEL) to measure the degree to which predictive confidence improves for test samples when certain modalities or contexts are removed. A lower CEL denotes more reliable predictions. Except for CEL, all metrics are represented in percentages (%).

$$\mathbf{CEL}(\mathbf{d}_i) = \mathbf{dF}_{\mathbb{C}}(\mathbf{d}_i) + \mathbf{dF}_{\mathbb{M}}(\mathbf{d}_i) \tag{10}$$

4.3 Comparison Models

Aggregation-based fusion: Concatenation: DialogueRNN [34] and DialogueGCN [10]; Attention networks: CTNet [65] and SCMM [63].
Graph-based fusion: MMDFN [18], MMGCN [20], M3Net [4], CMCF-SRNet [66], and CORECT [40].

Furthermore, we also consider other model-agnostic confidence calibration methods suitable for multi-modal scenarios: T-Scaling [11], Ensemble [28], CRL [37], FMFP [69], and CML [33]. These can be seamlessly integrated into the MERC task for a fair comparison.

In our ablation study, we present variations of our proposed CMERC: "**w/o** CL" denotes without CL for MERC; "**w/o** HSCL" denotes without Hybrid SCL (HSCL); "**w/o** CC" denotes without Confidence Constraint (CC).

Table 3: W-F1 scores (%) of the Baseline‡ under various missing information rates across different datasets.

Missing Rate	0	0.001	0.005	0.01	0.05	0.1	0.2	0.3	0.4
IEMOCAP	69.61	69.45	69.26	68.57	66.83	65.79	63.17	60.46	55.09
MELD	65.37	65.09	64.88	64.55	63.82	62.55	60.36	56.64	53.24

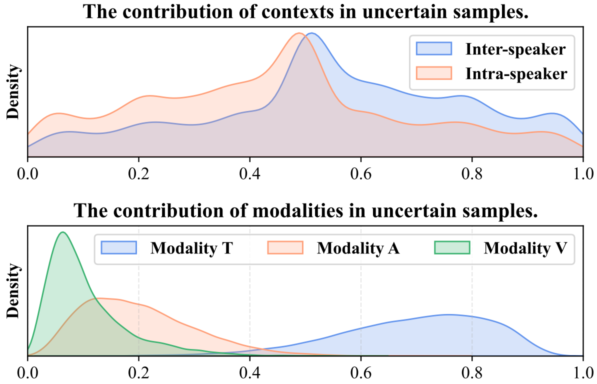

Figure 3: Contribution of various modalities or contexts in uncertain samples in the IEMOCAP testing set.

4.4 Reliability Analysis of MERC Models

In Table 2, a significant proportion of uncertain samples is observed in various MERC models, exceeding 90% in some cases, especially when modalities and contexts are removed. Additionally, the impact of each modality or context cannot be underestimated. Interestingly, more advanced models like M3Net show an even higher proportion, despite enhancements in performance, highlighting the formidable difficulty in reliable confidence prediction for MERC models.
Further reliability analysis: We investigate the reasons behind the generation of uncertain samples, focusing on the Baseline as a representative case. In Fig. 3, we visualize the contribution of each modality and context to uncertain samples, as measured by MM-SHAP [41]. This reveals that on the one hand, a significant portion of uncertain samples arises from the model overly prioritizing the textual modality. On the other hand, uncertain samples can also be attributed to the model struggling to appropriately weigh the influence of disparate contexts, treating them indiscriminately.

4.5 Analysis of Removed Information

Because if the removed information (i.e., three modalities and two types of contexts) is noise, the increase in predictive confidence is not a bad thing. In fact, the likelihood of increased predictive confidence resulting from noise removal is extremely low for the following reasons: (1) The granularity of removed information is substantial, operating at certain modalities or specific contextual cues rather than at the feature level. (2) The determination of whether removed information is noise depends on the model. As models improve in understanding and denoising, such cases are expected to decrease, particularly as we employ advanced MERC models. (3) Assuming removed information is noise is probable, thus removing it with extremely low probability ideally shouldn't

Table 4: Comparison of results under the multimodal setting. ★ indicates the source code available. ‡ denotes our re-implementation results. $^\sharp$, $^\flat$, and $^\natural$ represent results come from [18], [51], and original papers, respectively.

Methods	IEMOCAP								MELD								
	Happy	Sad	Neutral	Angry	Excited	Frustrated	W-F1	CEL	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	W-F1	CEL
★DialogueRNN$^\flat$	33.67	72.91	52.32	61.40	74.24	56.54	59.75	**450.94**‡	75.50	48.81	0.00	18.24	52.04	0.00	45.77	57.11	2002.86‡
★DialogueGCN$^\sharp$	51.57	80.48	57.69	53.95	72.81	57.33	62.89	939.51‡	75.97	46.05	-	19.6	51.2	-	40.83	56.36	1730.18‡
CTNet$^\natural$	51.30	79.90	65.80	67.20	78.70	58.80	67.50	-	77.40	52.70	10.00	32.50	56.00	11.20	44.60	60.50	-
★MMGCN$^\sharp$	45.14	77.16	64.36	68.82	74.71	61.40	66.26	555.78‡	76.33	48.15	-	26.74	53.02	-	46.09	58.31	**499.89**‡
★MMDFN$^\sharp$	42.22	78.98	66.42	69.77	75.56	66.33	68.18	766.17‡	77.76	50.69	-	22.93	54.78	-	47.82	59.46	1275.21‡
SCMM$^\natural$	45.37	78.76	63.54	66.05	76.70	66.18	67.53	-	-	-	-	-	-	-	-	59.44	-
★M3Net$^\flat$	52.74	79.39	67.55	69.30	74.39	66.58	69.24	1772.54‡	79.31	58.76	20.51	40.46	63.21	26.17	52.53	65.47	799.58‡
CMCF-SRNet$^\natural$	52.20	80.90	68.80	**70.30**	76.70	61.60	69.60	-	-	-	-	-	-	-	-	62.30	-
CORECT$^\natural$	59.30	80.53	66.94	69.59	72.69	**68.50**	70.02	-	-	-	-	-	-	-	-	-	-
Baseline‡	57.05	76.70	70.55	66.08	77.37	64.37	69.61	1772.54	78.80	55.93	**28.89**	**40.83**	64.29	29.57	52.56	65.37	799.58
w/ CMERC (Ours)	60.73	**81.89**	**71.65**	69.51	77.45	67.02	**71.98**	1143.50	**80.18**	60.42	24.69	40.48	**65.30**	**32.31**	54.16	**66.85**	605.05
w/o CL	54.95	80.79	69.46	65.51	**78.99**	65.51	70.29	1447.87	79.82	60.04	22.78	35.50	63.46	29.27	**55.51**	66.01	659.26
w/o HSCL	59.28	80.35	70.11	66.02	77.50	66.05	70.67	1364.62	79.65	59.82	25.00	39.89	63.02	27.91	53.50	65.94	623.82
w/o CC	**61.82**	81.78	70.88	66.48	76.28	63.22	70.46	1530.88	79.86	**60.54**	27.59	40.35	62.17	30.89	54.16	66.23	611.77

Table 5: Comparison of results with other model-agnostic confidence calibration methods under the multimodal setting.

Methods	IEMOCAP							MELD						
	W-F1	CEL	ECE ↓	MCE ↓	RMSCE ↓	AUROC ↑	AUPRC ↑	W-F1	CEL	ECE ↓	MCE ↓	RMSCE ↓	AUROC ↑	AUPRC ↑
Baseline‡	69.61	1772.54	15.21	18.18	16.18	92.30	73.36	65.37	799.58	22.58	27.84	23.15	85.78	67.57
w/ ★T-Scaling‡	70.04	1850.75	16.33	19.17	17.16	92.25	73.70	65.64	624.63	23.68	30.54	24.95	86.76	69.11
w/ ★Ensemble‡	70.31	1579.66	14.06	16.71	14.39	92.93	75.03	65.54	707.75	22.51	31.77	23.87	87.19	69.85
w/ ★CRL‡	70.78	1670.28	14.55	16.58	14.91	92.53	74.35	65.09	671.79	23.54	33.37	24.85	86.91	69.11
w/ ★FMFP‡	63.32	2673.93	22.53	26.27	23.77	89.15	64.65	60.80	625.76	30.03	35.60	30.40	84.13	64.13
w/ ★CML‡	69.19	1324.93	13.68	16.79	14.46	92.82	75.12	66.11	637.73	22.52	30.99	23.72	87.33	70.01
w/ CMERC	**71.98**	**1143.50**	**11.27**	**13.33**	**11.97**	**93.02**	**77.01**	**66.85**	**605.05**	**22.13**	**26.18**	**22.74**	**87.56**	**70.07**

decrease model performance. However, the Baseline in Table 3 shows performance degradation even with a missing information rate as low as 0.001. This highlights that this assumption doesn't hold under such coarse-grained information removal operations.

4.6 Overall Results

Table 4 compares our method with others, showing its superior W-F1 score and establishing a new state-of-the-art benchmark. Specifically, W-F1 scores rose by 2.37% and 1.48% for IEMOCAP and MELD, respectively. CEL decreased by 629.04 and 194.53 for IEMOCAP and MELD. Table 5 further demonstrates our method's superiority over other confidence calibration methods across various metrics. Importantly, we observed a performance improvement in the advanced MERC model. However, its unreliable confidence estimation has led to deteriorating effects, as evident from the increasing CEL and echoed by other metrics in Table 6. Addressing this issue could further enhance the model's performance, as supported by Table 6. This demonstrates that the CMERC enhances both the reliability and the generalization of various MERC models.

4.7 Ablation Study

In this section, we analyze the impact of various components within the CMERC. Ablation experiments in Table 4 show significant improvements across all components. Statistical analysis further confirms this, with a p-value ≪ 0.05 for the paired t-test.

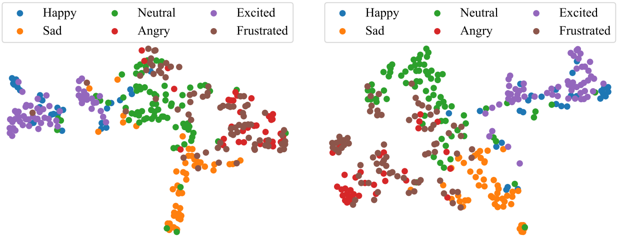

Figure 4: Visualization of intermediate embeddings of uncertain samples from the Baseline (left) and the Baseline w/ HSCL (right) in the IEMOCAP testing set.

Analysis of CL: The effectiveness of CL is evident in guiding model learning [57], potentially enhancing performance on uncertain samples, as shown in Table 4. This could be because post-CL calibration results in decreased entropy in model predictions for uncertain samples, signifying enhanced comprehension of data distribution and improved classification reliability.
Analysis of HSCL: In Table 4, HSCL demonstrates the best overall performance on CEL. In Fig. 4, we conduct t-SNE visualization on the intermediate representations of the Baseline and the Baseline with HSCL. The latter exhibits significantly clearer distinctions compared to the former, with silhouette coefficients [46] of 0.048 and 0.098, respectively. This suggests that capturing latent features

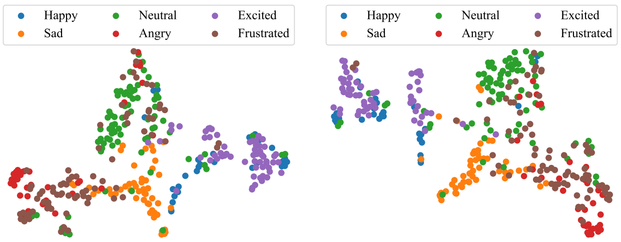

Figure 5: Visualization of intermediate embeddings of uncertain samples from the Baseline w/ Modality-specific SCL (left) and w/ Context-specific SCL (right) in the IEMOCAP testing set.

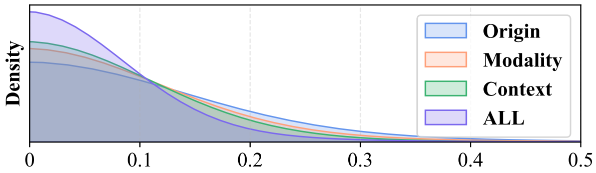

Figure 6: The distribution of CF values under CC.

contributing to uncertain samples aids in enhancing utterance representations. In Fig. 5, we also visualize the intermediate representations of the Baseline with modality-specific or context-specific SCL. Their silhouette coefficients of 0.063 and 0.067 surpass those derived from the Baseline, highlighting the efficacy of HSCL and the complementary nature of its constituent components.

Analysis of CC: Table 4 shows that CC significantly reduces CEL in the MELD dataset, but it falls short in lengthy conversations like IEMOCAP, possibly due to inefficiencies in the later conversation stages, as supported by Fig. 8. In Fig. 6, considering modalities and contexts in CC leads to CF value (Formula 4) distributions where smaller values correspond to higher density. This underscores CC’s effectiveness on uncertain samples and improves model reliability across modalities and contexts in the IEMOCAP testing set.

4.8 Hyperparameter Analysis

In Fig. 7, we demonstrate improved performance on the IEMOCAP validation set through the adjustment of hyperparameters, including γ_s, γ_m, and γ_c. These hyperparameters exhibit an initial increase, followed by a decrease and eventual stabilization with minor fluctuations. Importantly, performance consistently surpasses the case where these parameters are set to zero, showcasing the CMERC’s effectiveness across different hyperparameter settings.

4.9 Complementarity Analysis

In this section, we delve into the complementarity among three pivotal calibrations to elucidate the rationale of the CMERC framework. In Fig. 8, CC performs well in mid-conversation, particularly in the ‘Excited’ emotion and utterances with ES types, where two consecutive utterances exhibit different emotions. CL demonstrates strength in ‘Happy’ and ‘Angry’ emotions, especially in utterances without ES types. HSCL exhibits proficiency in handling ‘Neutral’ emotions. Collectively, each module showcases strengths in different aspects, underscoring their complementary performance, as also evidenced by Table 4.

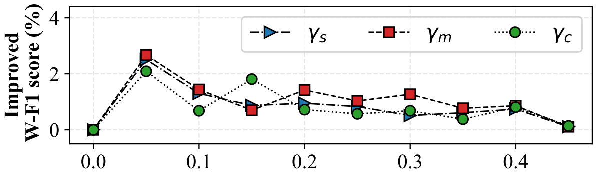

Figure 7: Improved W-F1 score of CMERC across various hyperparameters in the IEMOCAP validation set.

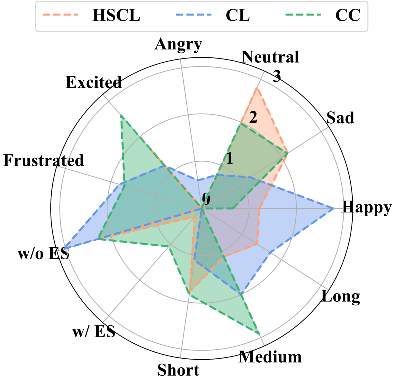

Figure 8: Improved W-F1 score (%) across various emotional categories, types (ES: Emotion shift), and lengths (Short: first 1/3, Medium: middle 1/3, and Long: last 1/3 of a conversation) in the test set of the IEMOCAP dataset.

4.10 Generalizability Analysis

To evaluate the generalizability of our CMERC, we conduct experiments with MERC models, as presented in Table 6. Noticeably, we observe a consistent decrease in CEL and an improvement in W-F1 scores across all methods. Additionally, similar enhancements are reflected in other confidence estimation metrics. These findings demonstrate the generalization capability of our CMERC across different MERC models.

4.11 Error Analysis

Many errors in our method stem from class imbalance, as evidenced by the low F1 scores of 24.69% and 32.31% for the ‘Fear’ and ‘Disgust’ emotions, respectively, in the MELD dataset. This phenomenon also constitutes a primary constraint on the performance of the MERC task, a fact supported by the results in Table 4. Furthermore, we are also investigating cases where the CMERC-enhanced MERC model misclassifies samples correctly predicted by Baseline‡, totaling 89 samples in the IEMOCAP dataset. Notably, it struggles particularly in Short (48 samples) and Medium (34 samples) positions of conversations, with relatively better performance observed in Long positions (7 samples). This could be due to the benefits of CL, as CL exhibits outstanding performance in utterances at long positions

Table 6: Performance of various MERC methods based on the CMERC framework for generalizability analysis.

Methods	IEMOCAP								MELD								
	Happy	Sad	Neutral	Angry	Excited	Frustrated	W-F1	CEL	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	W-F1	CEL
DialogueRNN‡	30.71	83.71	53.37	62.57	68.06	56.92	60.44	450.94	76.19	47.66	0.00	23.28	51.99	0.00	42.65	57.30	2002.86
w/ CMERC	34.34	78.48	57.62	59.04	76.84	57.80	62.43	352.10	76.42	48.38	0.00	23.90	52.90	0.00	45.92	58.11	1758.44
DialogueGCN‡	44.88	77.97	59.56	60.62	69.69	57.18	62.46	939.51	76.19	27.50	0.00	12.04	40.32	5.71	36.86	51.82	1730.18
w/ CMERC	35.75	79.53	59.67	60.92	75.84	59.35	63.58	825.94	74.06	45.01	0.00	23.81	52.90	0.00	39.30	55.73	1593.99
MMGCN‡	41.43	77.75	60.77	70.19	74.29	63.00	65.62	555.78	75.89	43.79	0.00	21.43	54.48	0.00	46.84	57.52	499.89
w/ CMERC	43.08	78.63	62.29	69.44	76.32	61.68	66.24	435.84	76.89	47.20	0.00	27.44	54.34	0.00	45.59	58.67	474.27
MMDFN‡	43.45	80.00	63.13	71.15	74.04	66.24	67.51	766.17	76.25	48.34	0.00	24.94	52.36	0.00	46.49	58.09	1275.21
w/ CMERC	46.26	81.64	67.50	67.49	75.29	65.98	68.83	717.01	77.34	48.57	0.00	26.20	53.41	0.00	48.02	59.11	963.40
M3Net‡	57.05	76.70	70.55	66.08	77.37	64.37	69.61	1772.54	78.80	55.93	28.89	40.83	64.29	29.57	52.56	65.37	799.58
w/ CMERC	60.73	81.89	71.65	69.51	77.45	67.02	71.98	1143.50	80.18	60.42	24.69	40.48	65.30	32.31	54.16	66.85	605.05

Methods	IEMOCAP					MELD				
	ECE ↓	MCE ↓	RMSCE ↓	AUROC ↑	AUPRC ↑	ECE ↓	MCE ↓	RMSCE ↓	AUROC ↑	AUPRC ↑
DialogueRNN‡	40.20	77.12	44.48	49.75	18.45	36.92	48.00	38.64	49.78	28.85
w/ CMERC	16.71	61.78	21.45	50.80	19.97	28.32	46.08	31.83	50.27	29.24
DialogueGCN‡	10.77	15.06	11.04	86.65	59.96	11.88	19.25	13.00	75.97	50.65
w/ CMERC	4.52	14.77	7.08	88.02	63.22	8.77	14.76	10.23	78.71	56.09
MMGCN‡	7.56	16.27	8.31	90.44	69.34	21.40	31.24	22.54	81.39	59.29
w/ CMERC	4.49	13.08	5.45	91.10	70.36	16.85	20.91	17.17	82.15	60.82
MMDFN‡	11.54	14.46	12.10	91.89	72.47	12.62	16.04	13.12	81.29	60.19
w/ CMERC	6.28	10.90	6.76	92.10	73.50	6.65	10.48	7.41	81.84	60.97
M3Net‡	15.21	18.18	16.18	92.30	73.36	22.58	27.84	23.15	85.78	67.57
w/ CMERC	11.27	13.33	11.97	93.02	77.01	22.13	26.18	22.74	87.56	70.07

in conversations, as depicted in Fig 8. Misclassifications are distributed almost evenly across samples with and without ES types. This suggests that our method is not affected by the phenomenon of emotional shift, regardless of whether it is present or not.

4.12 Case Study

In Fig. 9, we examine misclassified uncertain samples, where color intensity indicates the contribution of modalities or contextual cues. The Baseline tends to favor the textual modality, overlooking other modalities. Through calibration, improvements are observed. For instance, in the first utterance, the model correctly focuses on the visual modality to identify the emotion 'Sad'. Similarly, in the second utterance, emphasizing visual cues helps identify the 'Happy' emotion. For the third and fourth utterances, The Baseline treats inter- and intra-speaker contexts alike, while calibration introduces a preference. Specifically, in the fourth utterance, the model appropriately prioritizes inter-speaker context, recognizing the influence of others on the speaker's 'Frustrated' emotion.

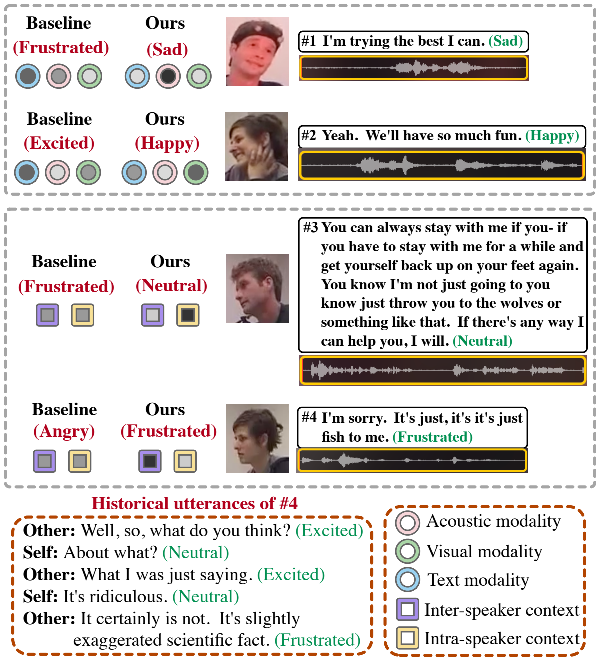
Baseline (Frustrated)
Ours (Sad)
#1 I'm trying the best I can. (Sad)
Baseline (Excited)
Ours (Happy)
#2 Yeah. We'll have so much fun. (Happy)
Baseline (Frustrated)
Ours (Neutral)
#3 You can always stay with me if you- if you have to stay with me for a while and get yourself back up on your feet again. You know I'm not just going to you know just throw you to the wolves or something like that. If there's any way I can help you, I will. (Neutral)
Baseline (Angry)
Ours (Frustrated)
#4 I'm sorry. It's just, it's it's just fish to me. (Frustrated)
Historical utterances of #4
Other: Well, so, what do you think? (Excited)
Self: About what? (Neutral)
Other: What I was just saying. (Excited)
Self: It's ridiculous. (Neutral)
Other: It certainly is not. It's slightly exaggerated scientific fact. (Frustrated)
Acoustic modality
Visual modality
Text modality
Inter-speaker context
Intra-speaker context

Figure 9: Examples of utterances in the IEMOCAP testing set for the case study. Predicted and golden labels are highlighted in red and green fonts, respectively.

5 CONCLUSION

In this paper, we introduce a novel calibration framework CMERC designed to tackle the issue of unreliable predictive confidence in MERC models without altering their structures. It integrates CL to guide the model to learn progressively uncertain samples in an entropy-increasing environment; Hybrid SCL to separate uncertain samples and reinforce the model's emphasis on factors causing uncertainty; and Confidence Constraint to penalize unexpected confidence surges on uncertain samples. These modules complement each other to improve the reliability of model predictions, particularly on uncertain samples that may arise due to the model's tendency to overly prioritize textual information or struggle to effectively balance different contexts. Experimental results on two conversational datasets demonstrate the effectiveness and generalization capabilities of our CMERC across various MERC models, surpassing state-of-the-art methods.

ACKNOWLEDGMENTS

We thank all anonymous reviewers for their helpful comments. This work was supported in part by the National Natural Science Foundation of China under Grants 62176076, the Natural Science Foundation of Guangdong under Grant 2023A1515012922, Shenzhen Foundational Research Funding under Grant JCYJ20220818102415032, The Major Key Project of PCL under Project PCL2023A09, and Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies under Grant 2022B1212010005k.

REFERENCES

[1] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*. 41–48.

[2] M Burgin. 2002. The essence of information: Paradoxes, contradictions, and solutions. In *Electronic Conference on Foundations of Information Science: The nature of information: Conceptions, misconceptions, and paradoxes (FIS 2002). Retrieved September*, Vol. 13. Citeseer, 2013.

[3] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *LREC* 42, 4 (2008), 335–359.

[4] Feiyu Chen, Jie Shao, Shuyuan Zhu, and Heng Tao Shen. 2023. Multivariate, Multi-Frequency and Multimodal: Rethinking Graph Neural Networks for Emotion Recognition in Conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10761–10770.

[5] Adam D Cobb and Brian Jalaian. 2021. Scaling Hamiltonian Monte Carlo inference for Bayesian neural networks with symmetric splitting. In *Uncertainty in Artificial Intelligence*. PMLR, 675–685.

[6] Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. 2019. Addressing failure prediction by learning model confidence. *Advances in Neural Information Processing Systems* 32 (2019).

[7] John Denker and Yann LeCun. 1990. Transforming neural-net output levels to probability distributions. *Advances in neural information processing systems* 3 (1990).

[8] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. 1459–1462.

[9] Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: COmmonSense knowledge for eMotion identification in conversations. In *Findings of EMNLP*. 2470–2481.

[10] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2020. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *EMNLP-IJCNLP*.

[11] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*. PMLR, 1321–1330.

[12] Kartik Gupta, Amir Rahimi, Thalaiyasingam Ajanthan, Thomas Mensink, Cristian Sminchisescu, and Richard Hartley. 2020. Calibration of Neural Networks using Splines. In *International Conference on Learning Representations*.

[13] Danijar Hafner, Dustin Tran, Timothy Lillicrap, Alex Irpan, and James Davidson. 2020. Noise contrastive priors for functional uncertainty. In *Uncertainty in Artificial Intelligence*. PMLR, 905–914.

[14] Marton Havasi, Rodolphe Jenatton, Stanislav Fort, Jeremiah Zhe Liu, Jasper Snoek, Balaji Lakshminarayanan, Andrew Mingbo Dai, and Dustin Tran. 2020. Training independent subnetworks for robust prediction. In *International Conference on Learning Representations*.

[15] Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018. Icon: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 conference on empirical methods in natural language processing*. 2594–2604.

[16] Ramya Hebbalaguppe, Jatin Prakash, Neelabh Madan, and Chetan Arora. 2022. A stitch in time saves nine: A train-time regularizing loss for improved neural network calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16081–16090.

[17] Dan Hendrycks and Kevin Gimpel. 2016. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *International Conference on Learning Representations*.

[18] Dou Hu, Xiaolong Hou, Lingwei Wei, Lianxin Jiang, and Yang Mo. 2022. MM-DFN: Multimodal Dynamic Fusion Network for Emotion Recognition in Conversations. In *ICASSP*. IEEE, 7037–7041.

[19] Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021. DialogueCRN: Contextual Reasoning Networks for Emotion Recognition in Conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 7042–7052.

[20] Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 2021. MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation. In *ACL*. 5666–5675.

[21] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *CVPR*. 4700–4708.

[22] Dazhi Jiang, Runguo Wei, Jintao Wen, Geng Tu, and Erik Cambria. 2022. AutoML-Emo: Automatic Knowledge Selection using Congruent Effect for Emotion Identification in Conversations. *IEEE Transactions on Affective Computing* (2022).

[23] Theofanis Karaletsos and Thang D Bui. 2020. Hierarchical gaussian process priors for bayesian neural network weights. *Advances in Neural Information Processing Systems* 33 (2020), 17141–17152.

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

[25] Meelis Kull, Telmo Silva Filho, and Peter Flach. 2017. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial intelligence and statistics*. PMLR, 623–631.

[26] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems* 30 (2017).

[27] Wei Li, Wei Shao, Shaoxiong Ji, and Erik Cambria. 2022. BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis. *Neurocomputing* 467 (2022), 73–82.

[28] Zhizhong Li and Derek Hoiem. 2020. Improving confidence estimates for unfamiliar examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2686–2695.

[29] Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. 2023. GCNet: Graph Completion Network for Incomplete Multimodal Learning in Conversation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[30] Zheng Lian, Bin Liu, and Jianhua Tao. 2021. CTNet: Conversational transformer network for emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 985–1000.

[31] Bingyuan Liu, Ismail Ben Ayed, Adrian Galdran, and Jose Dolz. 2022. The devil is in the margin: Margin-based label smoothing for network calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 80–88.

[32] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[33] Huan Ma, Qingyang Zhang, Changqing Zhang, Bingzhe Wu, Huazhu Fu, Joey Tianyi Zhou, and Qinghua Hu. 2023. Calibrating multimodal learning. In *International Conference on Machine Learning*. PMLR, 23429–23450.

[34] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 6818–6825.

[35] Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. 2021. Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems* 34 (2021), 15682–15694.

[36] Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. 2017. Variational dropout sparsifies deep neural networks. In *International conference on machine learning*. PMLR, 2498–2507.

[37] Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. 2020. Confidence-aware learning for deep neural networks. In *international conference on machine learning*. PMLR, 7034–7044.

[38] Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. 2020. Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems* 33 (2020), 15288–15299.

[39] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? *Advances in neural information processing systems* 32 (2019).

[40] Cam Van Thi Nguyen, Tuan Mai, Dang Kieu, Duc Trong Le, et al. 2023. Conversation Understanding using Relational Temporal Graph Neural Networks with Auxiliary Cross-Modality Interaction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 15154–15167.

[41] Letitia Parcalabescu and Anette Frank. 2023. MM-SHAP: A Performance-agnostic Metric for Measuring Multimodal Contributions in Vision and Language Models & Tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 4032–4059. https://doi.org/10.18653/v1/2023.acl-long.223

[42] Kanil Patel, William H Beluch, Bin Yang, Michael Pfeiffer, and Dan Zhang. 2020. Multi-Class Uncertainty Calibration via Mutual Information Maximization-based Binning. In *International Conference on Learning Representations*.

[43] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508* (2018).

[44] Amir Rahimi, Amirreza Shaban, Ching-An Cheng, Richard Hartley, and Byron Boots. 2020. Intra order-preserving functions for calibration of multi-class neural networks. *Advances in Neural Information Processing Systems* 33 (2020), 13456–13467.

[45] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *ACL*, Vol. 2020. NIH Public Access, 2359.

[46] Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* 20 (1987), 53–65.

[47] Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2021. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 13789–13797.

[48] Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021. Directed Acyclic Graph Network for Conversational Emotion Recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 1551–1560.

[49] Xudong Shen, Xianying Huang, Shihao Zou, and Xinyi Gan. 2024. Multimodal Knowledge-enhanced Interactive Network with Mixed Contrastive Learning for Emotion Recognition in Conversation. *Neurocomputing* 582 (2024), 127550.

[50] Tao Shi and Shao-Lun Huang. 2023. 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)*. 14752–14766.

[51] Tao Shi, Xiao Liang, Yaoyuan Liang, Xinyi Tong, and Shao-Lun Huang. 2023. SSLCL: An Efficient Model-Agnostic Supervised Contrastive Learning Framework for Emotion Recognition in Conversations. *arXiv preprint arXiv:2310.16676* (2023).

[52] Valentin I Spitkovsky, Hiyan Alshawi, and Dan Jurafsky. 2010. From baby steps to leapfrog: How "less is more" in unsupervised dependency parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 751–759.

[53] Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. 2019. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in neural information processing systems* 32 (2019).

[54] Geng Tu, Bin Liang, Dazhi Jiang, and Ruifeng Xu. 2022. Sentiment-Emotion-and Context-guided Knowledge Selection Framework for Emotion Recognition in Conversations. *IEEE Transactions on Affective Computing* (2022).

[55] Geng Tu, Bin Liang, Xiucheng Lyu, Lin Gui, and Ruifeng Xu. 2023. Do topic and causal consistency affect emotion cognition? a graph interactive network for conversational emotion detection. In *In The 26th European Conference on Artificial Intelligence (ECAI'23)*. 2362–2369.

[56] Geng Tu, Jintao Wen, Hao Liu, Sentao Chen, Lin Zheng, and Dazhi Jiang. 2022. Exploration meets exploitation: Multitask learning for emotion recognition based on discrete and dimensional models. *Knowledge-Based Systems* 235 (2022), 107598.

[57] Xin Wang, Yudong Chen, and Wenwu Zhu. 2021. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 9 (2021), 4555–4576.

[58] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *AAAI*, Vol. 33. 7216–7223.

[59] Jintao Wen, Dazhi Jiang, Geng Tu, Cheng Liu, and Erik Cambria. 2023. Dynamic interactive multiview memory network for emotion recognition in conversation. *Information Fusion* 91 (2023), 123–133.

[60] Yeming Wen, Ghassen Jerfel, Rafael Muller, Michael W Dusenberry, Jasper Snoek, Balaji Lakshminarayanan, and Dustin Tran. 2020. Combining Ensembles and Data Augmentation Can Harm Your Calibration. In *International Conference on Learning Representations*.

[61] Yang Wu, Zijie Lin, Yanyan Zhao, Bing Qin, and Li-Nan Zhu. 2021. A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 4730–4738.

[62] Chen Xing, Sercan Arik, Zizhao Zhang, and Tomas Pfister. 2019. Distance-Based Learning from Errors for Confidence Calibration. In *International Conference on Learning Representations*.

[63] Haozhe Yang, Xianqiang Gao, Jianlong Wu, Tian Gan, Ning Ding, Feijun Jiang, and Liqiang Nie. 2023. Self-adaptive Context and Modal-interaction Modeling For Multimodal Emotion Recognition. In *Findings of ACL*. 6267–6281.

[64] Sukmin Yun, Jongjin Park, Kimin Lee, and Jinwoo Shin. 2020. Regularizing class-wise predictions via self-knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13876–13885.

[65] Dong Zhang, Weisheng Zhang, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2020. Modeling both intra-and inter-modal influence for real-time emotion detection in conversations. In *Proceedings of the 28th ACM International Conference on Multimedia*. 503–511.

[66] Xiaoheng Zhang and Yang Li. 2023. A Cross-Modality Context Fusion and Semantic Refinement Network for Emotion Recognition in Conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 13099–13110.

[67] Wenjie Zheng, Jianfei Yu, Rui Xia, and Shijin Wang. 2023. A Facial Expression-Aware Multimodal Multi-task Learning Framework for Emotion Recognition in Multi-party Conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 15445–15459.

[68] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. 2021. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16489–16498.

[69] Fei Zhu, Zhen Cheng, Xu-Yao Zhang, and Cheng-Lin Liu. 2022. Rethinking confidence calibration for failure prediction. In *European Conference on Computer Vision*. Springer, 518–536.