A Dual Contrastive Learning Framework for Enhanced Multimodal Conversational Emotion Recognition

Yunhe Xie **Chengjie Sun*** **Ziyi Cao** **Bingquan Liu**
Zhenzhou Ji **Yuanchao Liu** **Lili Shan**
Faculty of Computing, Harbin Institute of Technology
{xieyh,sunchengjie,liubq,jizhenzhou,shanlili}@hit.edu.cn
zyc@stu.hit.edu.cn, ycliuharbin@163.com

Abstract

Multimodal Emotion Recognition in Conversations (MERC) identifies utterance emotions by integrating both contextual and multimodal information from dialogue videos. Existing methods struggle to capture emotion shifts due to label replication and fail to preserve positive independent modality contributions during fusion. To address these issues, we propose a Dual Contrastive Learning Framework (DCLF) that enhances current MERC models without additional data. Specifically, to mitigate label replication effects, we construct context-aware contrastive pairs. Additionally, we assign pseudo-labels to distinguish modality-specific contributions. DCLF works alongside basic models to introduce semantic constraints at the utterance, context, and modality levels. Our experiments on two MERC benchmark datasets demonstrate performance gains of 4.67%-4.98% on IEMOCAP and 5.52%-5.89% on MELD, outperforming state-of-the-art approaches. Perturbation tests further validate DCLF's ability to reduce label dependence. Additionally, DCLF incorporates emotion-sensitive independent modality features and multimodal fusion representations into final decisions, unlocking the potential contributions of individual modalities.

1 Introduction

Multimodal Emotion Recognition in Conversations (MERC) aims to integrate various modalities from dialogue data to track the emotional trajectories of interlocutors. This field has gained significant attention due to its broad applicability in human-centered conversational intelligence (Li et al., 2023c; Ji et al., 2023; Anand et al., 2023).

Recent studies concentrate on modeling the intricate conversational information flow, primarily employing recurrence-based (Ju et al., 2023; Liang et al., 2024; Guo et al., 2024) or graph-based methods (Li et al., 2023a,b, 2024). Additionally, research on multimodal fusion strategies explores early fusion (Zhang et al., 2021; Shou et al., 2022; Wen et al., 2023) or a hybrid approach combining graph-based and late fusion (Hu et al., 2022; Fan et al., 2024; Ai et al., 2024). However, challenges include *sensing emotion shifts due to label replication* and the *dilution of individual modality contributions* in fusion processes remain unresolved, constraining MERC models' potential.

Model	Raw	ECCS	EICS
Unimodal Setting			
AGHMN(2020)	59.1	54.0(↓5.1)	25.7(↓33.4)
DialogueRNN(2019)	62.2	60.0(↓2.2)	30.5(↓31.7)
Multimodal Setting			
DDIN(2020)	66.7	64.3(↓2.4)	47.8(↓18.9)
MMGCN(2021)	67.4	63.7(↓3.7)	53.3(↓14.1)

Table 1: Preliminary experimental results of basic models: weighted-F1 performance on IEMOCAP (2008).

Ghosal et al. (2021) observe that existing models often replicate dominant labels frequently found in the context or mimic emotion transition patterns from the training data, rather than genuinely understanding the contextual semantics. To verify this label replication effect, Zhang and Song (2022) introduce a perturbation test. This test replaces the original context with different utterances from the same dataset that share the same emotion, termed Emotion-Consistent Context Substitution (ECCS). Another extreme setting involves replacing the context with utterances bearing entirely different emotions, referred to as Emotion-Inconsistent Context Substitution (EICS). We extend this test to a multimodal setting, with results shown in Table 1. Our findings indicate that ECCS slightly impacts model performance, while the EICS setting leads to a significant performance drop. This confirms that these models rely heavily on emotion labels, failing to capture the deeper context semantics. This

*Corresponding author.

Proceedings of the 31st International Conference on Computational Linguistics, pages 4055–4065
January 19–24, 2025.

Conversation Record		T	V	A	T+V+A	Ground Truth
Joey	u_1:Ross is planning your birthday party.	Neutral	Anger	Neutral	Neutral	Neutral
Monica	u_2:Oh my God! I love him!	Joy	Joy	Surprise	Joy	Joy
Joey	u_3:You'd better act surprised.	Surprise	Neutral	Neutral	Neutral	Neutral
Phoebe	u_4:About what?	Neutral	Joy	Surprise	Neutral	Surprise
Monica	u_5:My surprise party!	Surprise	Joy	Surprise	Surprise	Joy
Phoebe	u_6:Well, he didn't tell me.	Neutral	Sad	Neutral	Neutral	Sad

Figure 1: A MELD (2019) dialogue snippet with misclassifications highlighted in red across settings.

overreliance prevents the models from effectively handling sudden emotion shifts (Shen et al., 2021; Tu et al., 2023a; Kang and Cho, 2024).

On the other hand, theoretically, complementary information across modalities (e.g., u_1-u_3 in Figure 1) should lead to a significant performance improvement compared to single-modality settings. However, this advantage is not evident in MERC (Wang et al., 2023). Taking MMGCN's (Hu et al., 2021) performance on the MELD (Poria et al., 2019) dataset as an example, if we take the correct prediction from any single modality (textual, audio, visual) as the final judgment, the theoretical F1 score could reach 81.7. Yet, the current best performance is only around 70 (Dai et al., 2024). We attribute this discrepancy to the dilution of the accurate contributions of each modality during the fusion process. For instance, in u_6 of Figure 1, a `Sad` prediction based on visual signals is overshadowed by `Neutral` inferences from the textual and audio modalities. Existing methods decode the fused result without considering the varying contributions of each modality, resulting in a Pyrrhic victory that ultimately limits the model's potential.

To address these challenges, we propose a **D**ual **C**ontrastive **L**earning **F**ramework (DCLF) that integrates seamlessly with existing multimodal conversational discriminative models without requiring additional data. To mitigate the label replication effect, we construct contextually semantic-aware contrastive pairs. Specifically, we first employ a typical MERC model to distill the context and regard this representation as a dialogue summary. We then concatenate the historical utterances in the dialogue window with the summary to form context-consistent (positive) samples. Concurrently, we randomly sample utterances from the same dataset that share the same emotion as the historical utterances, pairing them with the dialogue summary as context-inconsistent (negative) samples. To distinguish the contributions of individual modalities, we assign pseudo-labels to the corresponding utterances, based on their ability to make accurate predictions in single-modality settings. Ultimately, DCLF operates alongside the original basic model, performing parallel contrastive learning with these newly constructed labels, thereby jointly establishing semantic constraints at the context, utterance, and modality levels, respectively.

We conduct experiments on two MERC benchmark datasets. Basic models utilizing different modeling strategies exhibit performance gains of 4.67%-4.98% on IEMOCAP and 5.52%-5.89% on MELD when integrated with DCLF. Our results show that context-aware contrastive learning helps reduce the model's excessive reliance on labels by controlling for emotion-related factors. Additionally, compared to baseline models, our framework consistently improves performance by effectively combining emotion-specific features from individual modalities with multimodal fusion data. This approach maximizes the unique contributions of each modality, enhancing the overall decision-making process.

Our main contributions are as follows:

1. We propose DCLF to enhance existing MERC models. This framework is compatible with existing models and requires no additional data. Basic models equipped with DCLF outperform current SOTA methods.

2. Context-aware contrastive pairs effectively mitigate the label replication effect, improving the model's ability to discriminate in emotion transition scenarios.

3. By assigning pseudo-labels based on the performance of each individual modality, DCLF enhances modality-specific contributions, reducing performance losses during fusion.

2 Related Work

2.1 Multimodal Conversational Emotion Recognition

Early MERC works explore the role of various modalities in emotion inference. Zhang et al. (2020) parallelize multiple DialogueRNN (Majumder et al., 2019), assigning a separate channel for each modality and fusing the outputs with an attention mechanism. Conversely, Ren et al. (2021) reorder the modules by applying attention to obtain a text-centered representation before dialogue modeling. Xing et al. (2020) replace CMN (Hazarika et al., 2018)'s memory module with a dynamic version for speaker state tracking, while Wen et al. (2023) expand CMN into a multimodal version using gated recurrent units for global modeling.

Recent studies introduce specialized modules to address the unique challenges in MERC. Li et al. (2024) enhance MMGCN (Hu et al., 2021) with SMOTE (Chawla et al., 2003) algorithm to improve recognition of minority classes. Dai et al. (2024) propose a consensus-aware learning module, aligning modalities through emotion consensus learning. Ai et al. (2024) incorporate event relationships by using Doc2EDAG (Zheng et al., 2021) for event extraction and constructing a weighted multi-relation graph to capture interlocutor-event dependencies.

2.2 Multimodal Fusion

MERC models can primarily be categorized based on the sequence of modality fusion into early fusion (Guo et al., 2024) and late fusion (Yang et al., 2023), with recent works often adopting a sequential graph-based and late fusion paradigm (Li et al., 2023a; Fan et al., 2024). Early fusion involves integrating data from different modalities at the feature level (Ji et al., 2023). In contrast, late fusion processes and classifies each modality's data separately, then combines the results. Self-attention mechanisms that treat different modalities as query, key, and value also gain popularity (Lian et al., 2021; Zhang et al., 2023). Additionally, some approaches treat different modalities of the same utterance as distinct languages, employing end-to-end encoder-decoder structures to explore cross-modal relationships (Wang et al., 2020; Lian et al., 2022).

2.3 Contrastive Learning

Li et al. (2022) are the first to introduce supervised contrastive learning to ERC, enhancing emotion differentiation by excluding dissimilar emotions. Nie et al. (2023) employ contrastive learning with theme-aligned utterances as positive samples to identify if pairs belong to the same conversation. Song et al. (2022) tackle emotional imbalance with a prototypical contrastive loss function that works without large batch sizes. Zhang and Song (2022) introduce a semantics-guided contrastive context-aware approach, but its perturbation testing does not align with the process of constructing positive and negative examples. Hu et al. (2023) combine contrastive-aware adversarial training and joint class propagation to extract structured representations. Gao et al. (2024) and Jian et al. (2024) refine pre-trained models by leveraging contrastive learning to create distinct representational spaces. In multimodal settings, Yang et al. (2023) model contextual dependencies and enhance discriminability through adaptive path selection and contrastive learning. Dai et al. (2024) introduce speaker-guided contrastive learning to ensure diversity and semantic consistency across modalities.

3 Methodology

3.1 Problem Definition

A dialogue can be represented as a sequence of utterances $\{u_1, ..., u_i, ..., u_N\}$, where i stands for the utterance index and N is the total number of utterances. Each utterance u_i is associated with a corresponding interlocutor $I_i \in \mathcal{I}$, where $|\mathcal{I}| \geq 2$. If we merge each I_i and u_i as a pair $U_i = (I_i, u_i)$, the sequential U_i constitutes the multimodal conversation record $\mathcal{C}$. Each utterance u_i is also assigned a discrete emotion label $y_i \in \mathcal{Y}$, where $\mathcal{Y}$ is a set of pre-defined emotion labels. The objective of MERC is to recognize y_i for u_i based on $\mathcal{C}$.

3.2 Overview

The proposed DCLF, as illustrated in Figure 2, comprises the following components. First, an original MERC model processes the conversation record through stages of feature extraction and Semantic Modeling & Modal Fusion (SMMF), leading to a final prediction by the decoder. In the Context-Aware Contrastive Learning (CACL) module, we leverage SMMF to extract contextual features from the target utterance, creating a dialogue summary. We then concatenate historical utterances with this summary as positive samples, while negative samples are generated by randomly selecting utterances with the same emotions from different dialogues within the same dataset, paired with the same summary. In the Modality Contribution Contrastive Learning (MCCL) module, we assign pseudo-labels to each modality's corresponding utterances based on whether correct predictions can be made under single-modality settings. Finally, CACL and MCCL are executed in parallel, working collaboratively with the original MERC model.

3.3 Typical MERC Model

A typical neural MERC model generally consists of three components: a feature extractor, a semantic modeling & feature fusion module, and a decoder. In this study, we focus on the commonly used vi-

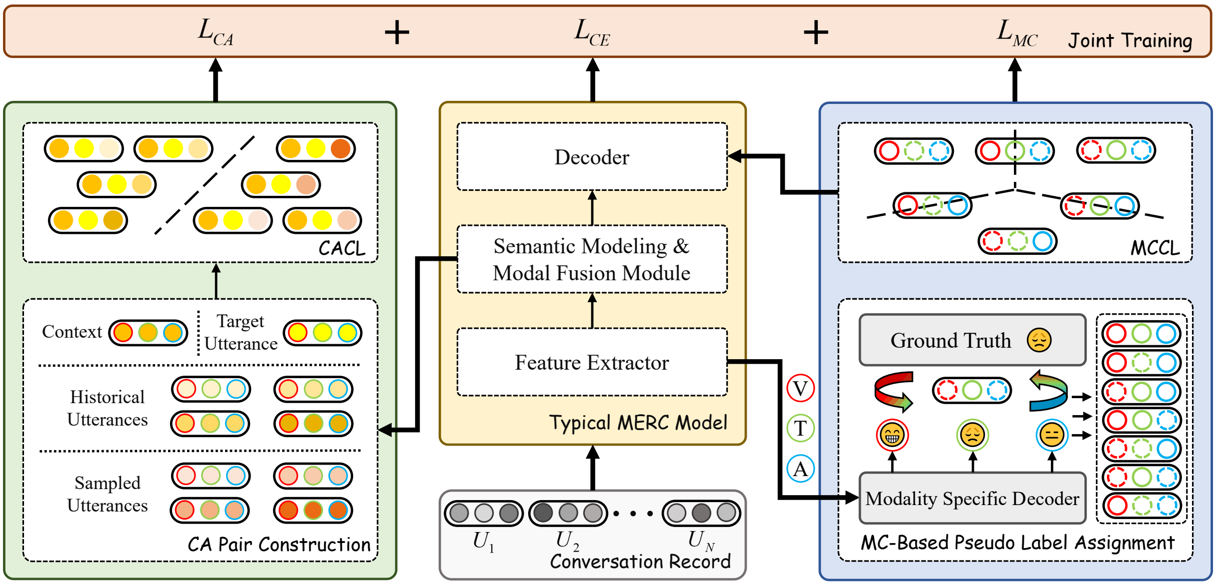


Figure 2: The overall framework of DCLF. Outer ring colors represent visual, textual, and audio modalities. CA, MC, CE and CL stand for context-aware, modality contribution, cross-entropy and contrastive learning, respectively.

sual, textual, and audio modalities. The feature extractor processes the multimodal conversation records as input and derives modality-specific representations $\boldsymbol{u}_i^m$ for each utterance u_i:

$$\boldsymbol{u}_i^m = \text{FeatureExtractor}\left(u_i\right), \tag{1}$$

where $m \in \{v, t, a\}$.

The SMMF module typically utilizes a combination of sequence modeling networks to manage the intricate streams of dialogue and modality information. Formally, it takes initial modality-specific representations as input and outputs the emotional hidden state $\boldsymbol{h}_i \in \mathbb{R}^d$ for each utterance u_i:

$$\boldsymbol{h}_i = \text{SMMF}\left(\mathcal{C}, \boldsymbol{u}_i^v, \boldsymbol{u}_i^t, \boldsymbol{u}_i^a\right). \tag{2}$$

Finally, the classification decoder, comprising fully connected layers and a softmax function, predicts the emotion label of the target utterance u_i:

$$\hat{\boldsymbol{y}}_i = \text{softmax}\left(\boldsymbol{W}\boldsymbol{h}_i + \boldsymbol{b}\right). \tag{3}$$

where $\boldsymbol{W} \in \mathbb{R}^{|\mathcal{Y}| \times d}$ and $\boldsymbol{b} \in \mathbb{R}^{|\mathcal{Y}|}$ are learnable parameters. Equations (1)-(3) outline the typical execution process of a MERC model:

$$\hat{\boldsymbol{y}}_i = \text{MERC}\left(\mathcal{C}, u_i\right), \tag{4}$$

which employs cross-entropy as the loss function:

$$L_{CE} = -\sum_{i=1}^{N}\sum_{e=1}^{|\mathcal{Y}|} y_{i,e} \log \hat{y}_{i,e}, \tag{5}$$

where $y_{i,e}$ and $\hat{y}_{i,e}$ are the components of $\boldsymbol{y}_i$ and $\hat{\boldsymbol{y}}_i$ for the emotion class e, respectively.

3.4 Context-Aware Contrastive Learning

To align with real-world applications, this study focuses exclusively on real-time emotion recognition. In this setting, the dialogue history $u_{1:i-1}$ serves as the context for the target utterance u_i.

The core of the CACL module lies in constructing context-aware contrastive pairs. The fundamental idea is to exclude the influence of emotion labels, enabling the target utterance to genuinely capture the contextual semantics. Specifically, constructing contrastive samples requires the contextual extract $\boldsymbol{c}_i \in \mathbb{R}^d$ (obtained from the SMMF module), and (pseudo) contextual utterances.

Context-Consistent (Positive) Pairs: It is assumed that the most relevant information for understanding the target utterance comes from its preceding dialogue window $u_{i-W:i-1}$, where W denotes the window size. Therefore, we sequentially concatenate the context $\boldsymbol{c}_i$ with these relevant utterances to form positive pairs, which are then aligned within the hidden state space, formalized as:

$$\boldsymbol{g}_p = \boldsymbol{W}_g\left[\boldsymbol{c}_i, \boldsymbol{h}_{i-p}\right] + \boldsymbol{b}_g, \tag{6}$$

where $p \in [1, W]$, $\boldsymbol{W}_g \in \mathbb{R}^{d \times 2d}$, $\boldsymbol{b}_g \in \mathbb{R}^d$ and $\boldsymbol{g}_p$ forms the context-consistent set $P_{CC}\left(i\right)$.

Context-Inconsistent (Negative) Pairs: We sample utterances consecutively from different dialogues within the same dataset as negative examples, aligning them as closely as possible to the emotions present in $u_{i-W:i-1}$. If an exact match is not available, we gradually relax the alignment cri-

Algorithm 1: Calculation of L_{MC} for each mini-batch $\mathcal{B}$

Input: $\mathcal{B}=\{\boldsymbol{u}_i^v,\boldsymbol{u}_i^t,\boldsymbol{u}_i^a\}_{i=1}^{N_b}, \ell_v, \ell_t, \ell_a, \leftarrow 0$
Output: $\mathcal{L}_{MC}$
// MC-Based Pseudo Label Assignment
1 **for** $i=1$ **to** N_b **do**
2   **for** $m \in \{v,t,a\}$ **do**
3     $p_i^m \leftarrow 0$;
4     $p_i^m \leftarrow \mathbb{I}(\mathcal{U}^m(\mathbf{u}_i^m) == y_i)$;
5 **for** $i=1$ **to** N_b **do**
6   $\ell_v^+,\ell_v^-,\ell_t^+,\ell_t^-,\ell_a^+,\ell_a^- \leftarrow 0$;
7   $n_v, n_t, n_a \leftarrow 0$;
  // Contrastive loss for u_i
8   **for** $j=1$ **to** N_b ***and*** $j \neq i$ **do**
9   **for** $m \in \{v,t,a\}$ **do**
10   **if** $p_j^m == p_i^m$ **then**
11         $\ell_m^+ \mathrel{+}= \mathcal{F}(\boldsymbol{u}_i^m, \boldsymbol{u}_j^m, \tau)$;
12         $n_m \mathrel{+}= 1$;
13   **else**
14         $\ell_m^- \mathrel{+}= \mathcal{F}(\boldsymbol{u}_i^m, \boldsymbol{u}_j^m, \tau)$;
15   **for** $m \in \{v,t,a\}$ **do**
16   **if** $n_m > 0$ **then**
17       $\ell_m + = -\log \frac{\ell_m^+}{n_m \times \ell_m^-}$;
18 $\mathcal{L}_{MC} \leftarrow \ell_v + \ell_t + \ell_a$

teria until a suitable match is found. This approach minimizes the influence of emotion labels and their transition patterns. Similarly, we concatenate the context $\boldsymbol{c}_i$ with these negative examples, converting them into negative pairs $\boldsymbol{g}_n$ ($n \in [1, W]$), forming the context-inconsistent set $P_{CI}(i)$.

$P_{CC}(i)$ and $P_{CI}(i)$ constitute the contrastive pair for u_i. We apply supervised contrastive learning (Khosla et al., 2020), treating $\boldsymbol{g}_p$ as the positive example and $\boldsymbol{g}_n$ as the negative example. The total loss L_{CA} for the CACL module is computed as:

$$\mathcal{F}(\boldsymbol{h}_i, \boldsymbol{g}_j) = \exp(\mathcal{G}(\boldsymbol{h}_i, \boldsymbol{g}_j)/\tau), \quad (7)$$

$$\mathcal{P}_{CA}(i) = \sum_{\boldsymbol{g}_p \in P_{CC}(i)} \mathcal{F}(\boldsymbol{h}_i, \boldsymbol{g}_p), \quad (8)$$

$$\mathcal{N}_{CA}(i) = \sum_{\boldsymbol{g}_n \in P_{CI}(i)} \mathcal{F}(\boldsymbol{h}_i, \boldsymbol{g}_n), \quad (9)$$

$$L_{CA} = -\sum_{i=1}^{N} \log \frac{1}{|P_{CC}(i)|} \frac{\mathcal{P}_{CA}(i)}{\mathcal{N}_{CA}(i)}, \quad (10)$$

where $\mathcal{G}(\cdot)$ is a score function, here using cosine similarity, and $\tau \in \mathbb{R}^+$ is a temperature parameter.

Dataset	#Dialogue			#Uterance			#Ut./Dia.
	Train	Val	Test	Train	Val	Test	
IEMOCAP	100	20	31	5146	664	1623	49.2
MELD	1039	114	280	9889	1109	2610	9.5

Table 2: Data distribution of IEMOCAP and MELD.

3.5 Modality Contribution Contrastive Learning

Effectively utilizing multimodal information is crucial in MERC. While some methods intuitively prioritize single modality as primary (Zhang et al., 2022), Song et al. (2022) demonstrate that textual information may fail to distinguish between emotions. Mao et al. (2021) reveal that textual information depends heavily on context, unlike visual and audio signals. Although modality fusion enhances MERC models, low-quality unimodal information can disrupt accuracy. In some cases, MERC models even underperform in comparison to single-modality settings, underscoring the need to isolate and understand individual modality contributions.

We design a modality contribution contrastive learning approach to capture both the correlations and differences in recognition tendencies across modalities. In the MCCL module, we connect the feature extractor directly to the modality-specific decoder, forming the element model $\mathcal{U}$. Initially, we conduct self-supervised modality-level pseudo-labeling, as detailed in *Lines* 1 to 4 of Algorithm 1. Then, we compute the contrastive loss for each utterance based on the pseudo-labels, following the steps outlined in *Lines* 8 to 14 of Algorithm 1, leading to the overall MCCL module loss L_{MC}.

Moreover, we combine the strengths of both feature- and decision-level fusion by concatenating each modality's contribution-aware representation with $\boldsymbol{h}_i$ before feeding it into the decoder. This ensures that high-confidence single-modality features are incorporated into the decision-making process.

3.6 Joint Training

The total loss of DCLF consists of two main categories: the original MERC model loss and the contrastive loss. We jointly train our proposed DCLF by minimizing the sum of the following losses:

$$L = L_{CE} + \gamma_{ca} L_{CA} + \gamma_{mc} L_{MC} + \lambda \|\boldsymbol{\theta}\|_2, \quad (11)$$

where γ_{ca} and γ_{mc} are tunable hyper-parameters. $\boldsymbol{\theta}$ is a set of learnable parameters of DCLF. λ represents the coefficient of L_2 regularization.

Methods	IEMOCAP																MELD	
	Happy		Sad		Neutral		Angry		Excited		Frustrated		Average		Average	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	WA	WF1	WA	WF1
DSAGCN (2022)	60.10	**62.60**	84.80	82.30	44.50	47.50	63.70	59.60	69.30	71.50	54.80	62.10	63.50	61.70	60.90	58.70
MM-DFN (2022)	—	42.22	—	78.98	—	66.42	—	69.77	—	75.56	—	66.33	68.21	68.18	62.49	59.46
DIMMN (2023)	24.30	30.20	64.50	74.20	57.30	59.00	61.80	62.70	81.30	72.50	75.90	66.60	64.70	64.10	60.60	58.60
SCMM (2023)	—	45.37	—	78.76	—	63.54	—	66.05	—	76.70	—	66.18	—	67.53	—	59.44
HI-IMC (2023)	55.80	51.40	80.50	84.40	64.20	62.00	65.20	64.20	88.50	**78.90**	68.20	64.50	70.60	67.90	61.70	60.80
GraphMFT (2023b)	—	45.99	—	83.12	—	63.08	—	70.30	—	76.92	—	63.84	67.90	68.07	61.30	58.37
GraphCFC (2023a)	—	43.08	—	84.99	—	64.70	—	71.35	—	**78.86**	—	63.70	69.13	68.91	61.42	58.86
SACCMA (2024)	—	38.60	—	**86.53**	—	64.90	—	64.56	—	74.52	—	62.99	67.41	67.10	62.30	59.30
IMBA (2024)	—	41.89	—	80.62	—	64.88	—	69.69	—	75.54	—	59.60	—	68.22	—	58.94
MultiDAG (2024)	—	45.26	—	81.40	—	69.53	—	70.33	—	71.61	—	66.94	69.11	69.08	64.41	64.00
GCCL (2024)	—	54.05	—	81.10	—	70.28	—	68.21	—	72.17	—	64.00	69.87	69.29	62.82	60.28
DER-GCN (2024)	60.70	**58.80**	75.90	79.80	66.50	61.50	71.30	**72.10**	71.10	73.30	66.10	**67.80**	69.70	69.40	66.80	66.10
DDIN* (2020)	28.87	34.31	78.74	84.60	63.82	64.14	56.36	61.05	90.14	77.99	64.89	63.50	67.34	66.70	61.91	61.02
w/ *DCLF*	56.42	54.67	95.39	**92.98**	64.33	65.64	67.84	**72.23**	82.66	77.83	64.00	65.20	72.01	**71.68**	67.39	**66.91**
MMGCN* (2021)	52.76	47.41	69.16	75.47	75.06	**72.16**	63.03	64.52	62.22	68.38	68.84	65.45	67.10	67.40	62.31	61.59
w/ *DCLF*	45.61	48.01	87.10	84.93	69.53	**71.35**	75.83	72.08	73.51	70.83	72.07	**75.23**	73.27	**72.07**	68.37	**67.11**

Table 3: Performance comparison of different methods under the multimodal setting (T+A+V). * indicates our reproduced results. The best overall performance and the top two F1 scores for each emotion are highlighted in bold.

4 Experiment

4.1 Datasets & Evaluation Metrics

We evaluate DCLF on two MERC benchmark datasets, **IEMOCAP**[1] (Busso et al., 2008) and **MELD**[2] (Poria et al., 2019), both of which provide aligned visual, textual, and audio information. The dataset statistics are presented in Table 2.

IEMOCAP includes dyadic dialogues, with each utterance annotated into one of six emotion categories: `Happy`, `Sad`, `Neutral`, `Angry`, `Excited`, and `Frustrated`. Consistent with prior work (Hu et al., 2021), we use the first four sessions for training, reserving the final session for testing.

MELD involves two or more speakers, and utterances are labeled by at least five experts across seven emotion categories: `Anger`, `Disgust`, `Fear`, `Joy`, `Neutral`, `Sadness`, and `Surprise`. We adopt the predefined split provided by MELD.

We use four metrics: accuracy (Acc), F1, Weighted Acc (WA), and Weighted F1 (WF1), focusing on WA and WF1 due to data imbalances. Acc and F1 for each emotion are also reported. The significance of the model with and without DCLF on datasets is validated by a paired *t*-test ($p < 0.05$).

4.2 Baselines

We compare our proposed DCLF with *twelve* MERC baselines, including recurrence-based methods: DIMMN (Wen et al., 2023), SCMM (Yang et al., 2023), SACCMA (Guo et al., 2024); A Transformer-based method: HI-IMC (Ji et al., 2023), and graph-based methods: DSAGCN (Shou et al., 2022), MM-DFN (Hu et al., 2022), GraphMFT (Li et al., 2023b), GraphCFC (Li et al., 2023a), IMBA (Li et al., 2024), MultiDAG (Nguyen et al., 2024), GCCL (Dai et al., 2024), and DER-GCN (Ai et al., 2024). A detailed introduction to baselines is provided in Section 2.

We further incorporate the proposed DCLF into DDIN[3] (Zhang et al., 2020) and MMGCN[4] (Hu et al., 2021), two representative early-stage open-source models, to evaluate the impact of DCLF.

4.3 Implementation Setups

For a fair comparison, we replace the text features in DDIN and MMGCN with RoBERTa (Liu et al., 2019) while keeping all other settings consistent with the original configurations. The temperature parameter τ is set to 0.07, and other hyperparameters are manually tuned via hold-out validation. In IEMOCAP, we set W, γ_{ca}, and γ_{mc} to 10, 0.8, and 0.4, respectively, with a batch size of 16. For MELD, these parameters were adjusted to 4, 0.6, 0.4, and 8. The reported results are the average scores from five random runs on the test set.

5 Results and Analysis

5.1 Overall Performance

Table 3 shows the experimental results across datasets. Comparing baselines with their DCLF-enhanced versions reveals the following insights:

[1]https://sail.usc.edu/iemocap/
[2]https://affective-meld.github.io/
[3]https://github.com/MANLP-suda/BiDDIN
[4]https://github.com/hujingwen6666/MMGCN

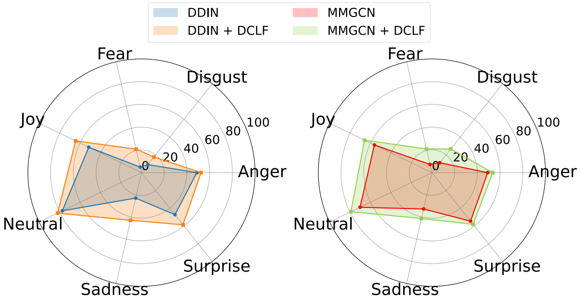


Figure 3: Specific label F1 performance on MELD.

(1) **Importance of contextual semantics and modality-specific contributions**: Among the baseline models, DER-GCN and MultiDAG outperform others on both datasets. DER-GCN uses an event extraction model, which *enhances contextual understanding* by extracting key information, constructing semantic networks, and analyzing causal links. While MultiDAG does not explicitly differentiate modality contributions, it *uniquely integrates unimodal features* into final decisions via residual connections, reinforcing their impact.

(2) **Effectiveness of DCLF**: DDIN and MMGCN serve as suitable basic models for testing DCLF due to their straightforward design. Both models show notable performance improvements when integrated with DCLF, surpassing the current state-of-the-art model by **0.81**% to **2.67**%. Specifically, DCLF improves DDIN by **4.98**% on IEMOCAP and **5.89**% on MELD, while MMGCN gains **4.67**% and **5.52**%. These results demonstrate the broad effectiveness of DCLF in enhancing MERC models with varied dialogue modeling approaches.

5.2 Specific Label Analysis

We compare the performance on specific emotions, as shown in Table 3. While certain methods excel in recognizing minority-class emotions in IEMOCAP, their deliberate emphasis on these emotions reduces performance on dominant emotions, leading to only marginal overall improvement. In contrast, DCLF does not simply prioritize minority-class emotion recognition but corrects dominant emotion misclassifications, leading to more balanced performance. For instance, after equipping DCLF, DDIN experiences only a 0.61% drop in `Neutral` while achieving better overall results.

On MELD, DCLF's benefits for minority-class emotions are even more pronounced. As shown in Figure 3, after integrating DCLF, DDIN and MMGCN double their initial performance on `Disgust` and improve `Fear` recognition by **2.89** to **4.25** times. Importantly, these enhancements are achieved without sacrificing performance on dominant emotions such as `Neutral` or `Joy`.

Methods	IEMOCAP		MELD	
	DDIN†(71.68)	MMGCN†(72.07)	DDIN†(66.91)	MMGCN†(67.11)
-w/o CACL	68.77(↓2.91)	69.95(↓2.12)	63.50(↓3.41)	64.55(↓2.56)
-w/o CE	70.89(↓0.79)	71.36(↓0.71)	66.25(↓0.66)	66.51(↓0.60)
-w/o MCCL	67.21(↓4.47)	68.02(↓4.05)	62.53(↓4.38)	62.45(↓4.66)
-w/o ICA	69.47(↓2.21)	70.11(↓1.96)	65.05(↓1.86)	65.83(↓1.28)

Table 4: WF1 results of ablation studies for different settings. † denotes DCLF-equipped.

Methods \ $\#W$	2	4	8	16
DDIN+DCLF	69.54	70.27	71.22	70.96
MMGCN+DCLF	70.29	71.05	71.61	71.44

(a) WF1 performance comparison on IEMOCAP.

Methods \ $\#W$	1	2	4	8
DDIN+DCLF	63.89	64.37	66.91	66.38
MMGCN+DCLF	64.82	66.29	67.11	66.84

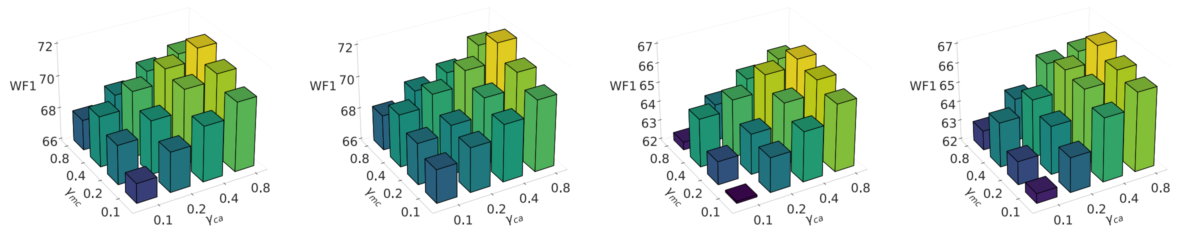
(b) WF1 performance comparison on MELD.

Table 5: Performance comparison for different dialogue window sizes (W).

5.3 Ablation Study

We conduct an ablation study to evaluate the impact of each DCLF component (Table 4). "-w/o CACL" and "-w/o MCCL" represent the removal of the CACL and MCCL modules, respectively. "-w/o CE" skips contextual extraction, using utterances directly as positive and negative samples, while "-w/o ICA" removes independent modality contribution awareness from decoding, leaving MCCL as a soft constraint during feature extraction.

The results indicate that MCCL has a greater impact than CACL, as prior methods often suppress independent modality contributions, which offer more room for improvement than contextual understanding. Tu et al. (2023b) also note that current models inherently denoise unrelated contexts. While CACL addresses the label replication effect in emotion shifts, most conversations exhibit stable emotional flows. Additionally, limiting MCCL to a soft constraint before modality fusion significantly reduces its effectiveness. Finally, combining contextual extraction with utterances strengthens the distinction between positive and negative pairs, improving contextual comprehension.



(a) DDIN† on IEMOCAP. (b) MMGCN† on IEMOCAP. (c) DDIN† on MELD. (d) MMGCN† on MELD.

Figure 4: WF1 results for different combinations of γ_{ca} and γ_{mc} values across datasets. † denotes DCLF-equipped.

Methods	IEMOCAP			MELD		
	Raw	ECCS	EICS	Raw	ECCS	EICS
DDIN	66.70	64.30(↓2.40)	47.80(↓18.90)	61.02	59.15(↓1.87)	47.74(↓13.28)
DDIN†	71.68	70.31(↓1.37)	64.15(↓7.53)	66.91	65.69(↓1.22)	61.42(↓5.49)
MMGCN	67.40	63.70(↓3.70)	53.30(↓14.10)	61.59	59.96(↓1.63)	45.62(↓15.97)
MMGCN†	72.07	69.91(↓2.16)	66.26(↓5.81)	67.11	66.27(↓0.84)	61.29(↓5.82)

Table 6: WF1 results of perturbation test. † denotes DCLF-equipped.

Methods	IEMOCAP			MELD		
	Whole	w/o ES	w/ ES	Whole	w/o ES	w/ ES
DDIN	66.70	73.85	54.27	61.02	68.34	54.69
DDIN+DCLF	71.68	76.16	63.88	66.91	75.83	59.25
MMGCN	67.40	72.60	58.35	61.59	68.95	55.27
MMGCN+DCLF	72.07	75.86	65.47	67.11	74.30	60.94

Table 7: WF1 results comparison on emotion shift.

5.4 Quantitative Analysis

5.4.1 Impact of Contrastive Pair Quantity

The dialogue window size W controls the CACL contrastive pairs' number. We analyze how varying W affects performance, assuming equal positive and negative pairs. As shown in Table 5, performance initially improves as W increases but then slightly declines. The optimal W values, 10 for IEMOCAP and 4 for MELD, appear to correspond with the average number of utterances per dialogue.

5.4.2 Impact of Contrastive Loss Coefficient

γ_{ca} and γ_{mc} control the model's focus on contextual information and on the characteristics of each modality, respectively. We assess model performance using various combinations of γ_{ca} and γ_{mc}, each ranging from [0.1, 0.2, 0.4, 0.8]. As shown in Figure 4, the performance improves initially as γ_{mc} increases but declines after a certain point. This indicates that distinguishing modality features early on benefits the model, while overemphasizing them can hinder the integration of contextual corrections later. On IEMOCAP, the optimal performance occurs at $\gamma_{ca} = 0.8$, while on MELD, it is $\gamma_{ca} = 0.6$. This difference can be attributed to the significantly longer dialogues in IEMOCAP, which require a stronger emphasis on contextual understanding.

5.5 Perturbation Test

We conduct perturbation tests as outlined in the introduction. We report the average scores in Table 6 based on five random seeds. The results show that DCLF significantly mitigates performance drops under the ECCS setting. This is primarily due to CACL acting as a regularization module, reducing reliance on label patterns and improving stability.

5.6 Error Analysis

Section 5.2 and Section 5.5 demonstrate DCLF's effectiveness in mitigating label replication effect. In this section, we extend the evaluation to emotion-shift scenarios. As shown in Table 7, the results reveal that integrating DCLF into MERC models effectively narrows the performance gap between emotion-shift and stable-emotion contexts, which aligns with the key motivation of this work.

6 Conclusion

This paper presents a Dual Contrastive Learning Framework for MERC, designed to enhance performance in emotion-shift dialogue scenarios. DCLF also ensures that the unique characteristics of each modality are preserved and effectively utilized. It integrates seamlessly with existing MERC models by applying semantic constraints at the context, utterance, and modality levels.

Experimental results confirm the effectiveness of DCLF in improving overall model performance. DCLF addresses the issue of replicated label patterns and reduces the loss of accuracy during the fusion of different modalities. Additionally, the framework improves the effectiveness of single modalities while maintaining flexibility, enabling it to extend beyond just MERC tasks and demonstrating DCLF's broad applicability.

Limitations

We evaluate DCLF using two MERC models with distinct dialogue modeling approaches. We do not extend the evaluation to more complex MERC models due to the limited availability of open-source implementations. Furthermore, evaluating straightforward models better highlights DCLF's true impact. Additionally, this work focuses solely on real-time recognition scenarios. It is worth noting that the performance of the MCCL module is constrained by the capacity of the feature extractor, and the quality of pseudo-labels heavily depends on the model's predictions. This dependency may lead to fluctuations in performance during training, though these stabilize as the model converges.

Acknowledgments

This work was supported by the National Key R&D Program of China under grant 2023YFC3804600 and the Fundamental Research Funds for the Central Universities (project number: 2022FRFK060002).

References

Wei Ai, Yuntao Shou, Tao Meng, and Keqin Li. 2024. Der-gcn: Dialog and event relation-aware graph convolutional neural network for multimodal dialog emotion recognition. *IEEE Transactions on Neural Networks and Learning Systems*.

Sidharth Anand, Naresh Kumar Devulapally, Sreyasee Das Bhattacharjee, and Junsong Yuan. 2023. Multi-label emotion analysis in conversation via multimodal knowledge distillation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6090–6100.

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. *Language resources and evaluation*, 42:335–359.

Nitesh V Chawla, Aleksandar Lazarevic, Lawrence O Hall, and Kevin W Bowyer. 2003. Smoteboost: Improving prediction of the minority class in boosting. In *Knowledge Discovery in Databases: PKDD 2003: 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, Cavtat-Dubrovnik, Croatia, September 22-26, 2003. Proceedings 7*, pages 107–119. Springer.

Yijing Dai, Jinxing Li, Yingjian Li, and Guangming Lu. 2024. Multi-modal graph context extraction and consensus-aware learning for emotion recognition in conversation. *Knowledge-Based Systems*, 298:111954.

Chunxiao Fan, Jie Lin, Rui Mao, and Erik Cambria. 2024. Fusing pairwise modalities for emotion recognition in conversations. *Information Fusion*, 106:102306.

Qingqing Gao, Jiuxin Cao, Biwei Cao, Xin Guan, and Bo Liu. 2024. Cept: A contrast-enhanced prompt-tuning framework for emotion recognition in conversation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2947–2957.

Deepanway Ghosal, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2021. Exploring the role of context in utterance-level emotion, act and intent classification in conversations: An empirical study. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1435–1449.

Lili Guo, Yikang Song, and Shifei Ding. 2024. Speaker-aware cognitive network with cross-modal attention for multimodal emotion recognition in conversation. *Knowledge-Based Systems*, 296:111969.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2122–2132.

Dou Hu, Yinan Bao, Lingwei Wei, Wei Zhou, and Songlin Hu. 2023. Supervised adversarial contrastive learning for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10835–10852.

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 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7037–7041. IEEE.

Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 2021. Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. 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)*, pages 5666–5675.

Xiaoyue Ji, Zhekang Dong, Yifeng Han, Chun Sing Lai, and Donglian Qi. 2023. A brain-inspired hierarchical interactive in-memory computing system and its application in video sentiment analysis. *IEEE transactions on circuits and systems for video technology*, 33(12):7928–7942.

Zhongquan Jian, Ante Wang, Jinsong Su, Junfeng Yao, Meihong Wang, and Qingqiang Wu. 2024. Emotrans: Emotional transition-based model for emotion recognition in conversation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5723–5733.

Wenxiang Jiao, Michael Lyu, and Irwin King. 2020. Real-time emotion recognition via attention gated hierarchical memory network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8002–8009.

Xincheng Ju, Dong Zhang, Suyang Zhu, Junhui Li, Shoushan Li, and Guodong Zhou. 2023. Real-time emotion pre-recognition in conversations with contrastive multi-modal dialogue pre-training. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 1045–1055.

Yujin Kang and Yoon-Sik Cho. 2024. Improving contrastive learning in emotion recognition in conversation via data augmentation and decoupled neutral emotion. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2194–2208.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Jiang Li, Xiaoping Wang, Guoqing Lv, and Zhigang Zeng. 2023a. Graphcfc: A directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition. *IEEE Transactions on Multimedia*, 26:77–89.

Jiang Li, Xiaoping Wang, Guoqing Lv, and Zhigang Zeng. 2023b. Graphmft: A graph network based multimodal fusion technique for emotion recognition in conversation. *Neurocomputing*, 550:126427.

Qianer Li, Peijie Huang, Yuhong Xu, Jiawei Chen, Yuyang Deng, and Shangjian Yin. 2024. Generating and encouraging: An effective framework for solving class imbalance in multimodal emotion recognition conversation. *Engineering Applications of Artificial Intelligence*, 133:108523.

Shimin Li, Hang Yan, and Xipeng Qiu. 2022. Contrast and generation make bart a good dialogue emotion recognizer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11002–11010.

Ziming Li, Yan Zhou, Yaxin Liu, Fuqing Zhu, Chuanpeng Yang, and Songlin Hu. 2023c. Qap: A quantum-inspired adaptive-priority-learning model for multimodal emotion recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12191–12204.

Zheng Lian, Bin Liu, and Jianhua Tao. 2021. Ctnet: Conversational transformer network for emotion recognition. *IEEE/ACM Trans. Audio, Speech, Language Process.*, 29:985–1000.

Zheng Lian, Bin Liu, and Jianhua Tao. 2022. Smin: Semi-supervised multi-modal interaction network for conversational emotion recognition. *IEEE Trans. Affect. Comput.*, 14(3):2415–2429.

Xingwei Liang, Geng Tu, Jiachen Du, and Ruifeng Xu. 2024. Multi-modal attentive prompt learning for few-shot emotion recognition in conversations. *Journal of Artificial Intelligence Research*, 79:825–863.

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*, abs/1907.11692.

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*, volume 33, pages 6818–6825.

Yuzhao Mao, Guang Liu, Xiaojie Wang, Weiguo Gao, and Xuan Li. 2021. Dialoguetrm: Exploring multi-modal emotional dynamics in a conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2694–2704.

Cao-Bach Nguyen, Duc-Trong Le, Quang Thuy Ha, et al. 2024. 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)*, pages 4259–4265.

Weizhi Nie, Yuru Bao, Yue Zhao, and Anan Liu. 2023. Long dialogue emotion detection based on commonsense knowledge graph guidance. *IEEE Transactions on Multimedia*, 26:514–528.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

Minjie Ren, Xiangdong Huang, Xiaoqi Shi, and Weizhi Nie. 2021. Interactive multimodal attention network for emotion recognition in conversation. *IEEE Signal Processing Letters*, 28:1046–1050.

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)*, pages 1551–1560.

Yuntao Shou, Tao Meng, Wei Ai, Sihan Yang, and Keqin Li. 2022. Conversational emotion recognition studies based on graph convolutional neural networks and a dependent syntactic analysis. *Neurocomputing*, 501:629–639.

Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. 2022. Supervised prototypical contrastive learning for emotion recognition in conversation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5197–5206.

Geng Tu, Ran Jing, Bin Liang, Min Yang, Kam-Fai Wong, and Ruifeng Xu. 2023a. A training-free debiasing framework with counterfactual reasoning for conversational emotion detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15639–15650.

Geng Tu, Bin Liang, Ruibin Mao, Min Yang, and Ruifeng Xu. 2023b. Context or knowledge is not always necessary: A contrastive learning framework for emotion recognition in conversations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14054–14067.

Yunxiao Wang, Meng Liu, Zhe Li, Yupeng Hu, Xin Luo, and Liqiang Nie. 2023. Unlocking the power of multimodal learning for emotion recognition in conversation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5947–5955.

Zilong Wang, Zhaohong Wan, and Xiaojun Wan. 2020. Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis. In *Proc. of WWW*, pages 2514–2520.

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

Songlong Xing, Sijie Mai, and Haifeng Hu. 2020. Adapted dynamic memory network for emotion recognition in conversation. *IEEE Transactions on Affective Computing*, 13(3):1426–1439.

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 the Association for Computational Linguistics: ACL 2023*, pages 6267–6281.

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*, pages 503–511.

Hanqing Zhang and Dawei Song. 2022. Towards contrastive context-aware conversational emotion recognition. *IEEE Transactions on Affective Computing*, 13(4):1879–1891.

Ke Zhang, Yuanqing Li, Jingyu Wang, Erik Cambria, and Xuelong Li. 2021. Real-time video emotion recognition based on reinforcement learning and domain knowledge. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1034–1047.

Tong Zhang, Shuzhen Li, Bianna Chen, Haozhang Yuan, and CL Philip Chen. 2022. Aia-net: Adaptive interactive attention network for text–audio emotion recognition. *IEEE Transactions on Cybernetics*, 53(12):7659–7671.

Yazhou Zhang, Yang Yu, Dongming Zhao, Zuhe Li, Bo Wang, Yuexian Hou, Prayag Tiwari, and Jing Qin. 2023. Learning multi-task commonness and uniqueness for multi-modal sarcasm detection and sentiment analysis in conversation. *IEEE Transactions on Artificial Intelligence*.

Shun Zheng, Wei Cao, Wei Xu, and Jiang Bian. 2021. Revisiting the evaluation of end-to-end event extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4609–4617.