# Context or Knowledge is Not Always Necessary: A Contrastive Learning Framework for Emotion Recognition in Conversations

**Geng Tu**[1,3] **Bin Liang**[1,3] **Ruibin Mao**[4] **Min Yang**[5] **Ruifeng Xu**[1,2,3] *
[1]Harbin Institute of Technology, Shenzhen, China
[2]Peng Cheng Laboratory, Shenzhen, China
[3]Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies
[4]Shenzhen Securities Information Co., Ltd, China
[5]SIAT, Chinese Academy of Sciences, Shenzhen, China

## Abstract

Emotion recognition in conversations (ERC) aims to detect the emotion of utterances in conversations. Existing efforts generally focus on modeling context- and knowledge-sensitive dependencies. However, it is observed that the emotions of many utterances can be correctly detected without context or external knowledge. In such cases, blindly leveraging the context and external knowledge may impede model training. Based on this, we propose a novel framework based on contrastive learning (CL), called CKCL (including the contrastive learning scenarios among **C**ontext and **K**nowledge), to distinguish the above utterances for better vector representations. The CKCL framework defines context- and knowledge-independent utterances, as the positive sample, whose predicted results are unchanged even masking context and knowledge representations, otherwise, the negative sample. This can obtain a latent feature reflecting the impact degree of context and external knowledge on predicted results, thus effectively denoising irrelevant context and knowledge during training. Experimental results on four datasets show the performance of CKCL-based models is significantly boosted and outperforms state-of-the-art methods.

## 1 Introduction

Emotion recognition in conversations (ERC) has received an active research attention (Tu et al., 2022c; Li et al., 2022b; Xie et al., 2021; Mao et al., 2021; Lian et al., 2021; Xiao et al., 2020) because of its wide applications in many fields such as opinion mining (Cortis and Davis, 2021) and recommender systems (Zheng et al., 2020). Existing works in ERC conventionally need to model context-sensitive dependencies (Wang et al., 2020; Jiao et al., 2020) and knowledge-sensitive dependencies (Li et al., 2021a,b; Ghosal et al., 2020; Zhong et al., 2019).

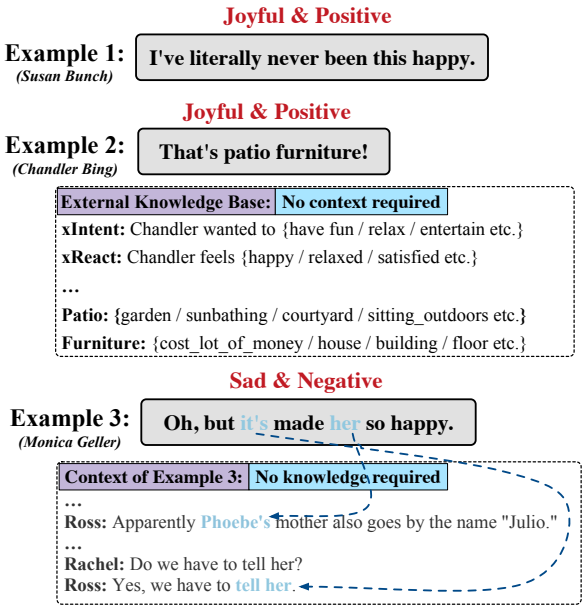



Figure 1: Examples of utterances, reflecting the context and knowledge are not always necessary in ERC.

Especially, in knowledge-sensitive ERC models, there are mainly two kinds of external knowledge: One is concepts retrieved from external knowledge bases ConceptNet (Speer et al., 2017) or SenticNet (Cambria et al., 2020). The other is generated by the pre-training commonsense transformers (COMET) (Bosselut et al., 2019). Additionally, in context-sensitive ERC models, recent studies proposed various methods, including memory networks (Kumar et al., 2022b; Xing et al., 2020; Hazarika et al., 2018) and graph-based models (Nie et al., 2021; Li et al., 2020; Ghosal et al., 2019; Zhang et al., 2019). However, these works do not follow whether models need context and external knowledge for the current utterance, but rather on improving modeling methods. In Fig. 1, intuitively, the emotion of example 1 can be recognized even without leveraging context and external knowledge. Example 2 (context-independent) is the first utterance in a conversation, lacking

* Corresponding author: xuruifeng@hit.edu.cn



*Findings of the Association for Computational Linguistics: ACL 2023*, pages 14054–14067
July 9-14, 2023 

any context. Without establishing a relationship between patio furniture and 'joyful', it becomes challenging to detect emotions. For example 3 (knowledge-independent), the literal meaning of the utterance is opposite to the conveyed emotion. So, it is difficult to correctly detect the emotion of the utterance without context. Although the above example exists, the removal of context or external knowledge will lead to a semantic gap between utterances representations. Therefore, how to differentiate context- and knowledge-independent utterances from other utterances is a challenge.

Based on the above, we proposed a framework based on contrastive learning (CL), CKCL, to distinguish context- and knowledge-independent utterances during training. Concretely, the context-independent and knowledge-independent (or context-dependent and knowledge-dependent) utterances are labeled as '1' (or labeled '0'). Then, the CKCL pulls utterances with the same or different labels together or further apart. For ERC models, this can alleviate the performance degradation of context- and knowledge-independent utterance representations during training. And the CKCL can also denoise irrelevant context and knowledge to improve the robustness ability of models. In addition, inspired by Li et al. (2022a), we introduce a weighted supervised CL (SCL) named Emotion SCL into CKCL, to further distinguish similar emotions, which takes into account the uneven distribution of classes in ERC. To summarize, our main contributions can be summarized as follows:

- We are the first to explore self-supervised CL in the ERC task.
- We propose a CKCL framework to differentiate context- and knowledge-independent utterances, which promotes the robustness of ERC models against irrelevant context and knowledge during training.
- Experimental results demonstrate that our proposed method can boost various baselines and outperforms state-of-the-art ERC methods.

## 2 Related Work

**Emotion Recognition in Conversations**

**Context-sensitive Models** The emotion generation theory (Gross and Barrett, 2011) indicates the importance of contextual information for emotion identification. RNN-based models (Poria et al., 2017) are often used to model context dependencies. However, they are unable to capture the distinction between historical utterances (Lian et al., 2021) when modeling context. To solve this problem, most works began to focus on the memory network (Hazarika et al., 2018; Jiao et al., 2020). In addition, the role of participants in ERC is also important to the speaker's emotional state (Wen et al., 2023). To model the speaker-level context, researchers have a greater emphasis on speaker-specific models (Kim and Vossen, 2021), graph-based models (Nie et al., 2021), and so on. For example, Majumder et al. (2019) utilized three GRUs to track global context, speaker state, and emotional state in conversations. Ghosal et al. (2019); Shen et al. (2021) employed a graph-based model to model self- and inter-speaker dependencies.

**Knowledge-sensitive Models** Although the above works have achieved respectable performance in ERC, they are not able to work like a human because of the lack of commonsense knowledge (Zhong et al., 2019). Therefore, Ghosal et al. (2020) utilized GRUs and generated knowledge from COMET, to model the psychological states of participants in conversations. Li et al. (2021b) introduce the psychological-knowledge-aware interaction graph (SKAIG) model to further model the structural psychological states. Fu et al. (2021) proposed a graph-based model to model the knowledge-sensitive dependencies by incorporating concepts retrieved from ConceptNet (Speer et al., 2017). Subsequently, Zhao et al. (2022) proposed a causal aware model using generated knowledge to capture the context information. However, these methods, both context- and knowledge-sensitive models, have performance degradation, that is, their performance in some utterances is even worse than models without context and knowledge.

**Contrastive Learning**

Chen et al. (2020) proposed a classic comparative learning network SimCLR, which uses various image augmentation strategies to produce positive and negative samples from the same image for visual representation. In the field of NLP, motivated by the poor performance of BERT in semantic text similarity tasks, Yan et al. (2021) proposed a **self-supervised CL** for fine-tuning BERT. Additionally, Kim et al. (2021) explore a CL method without data augmentation, which employs BERT with frozen and fine-tunable parameters to produce positive and negative samples. Then, Gao et al. (2021)

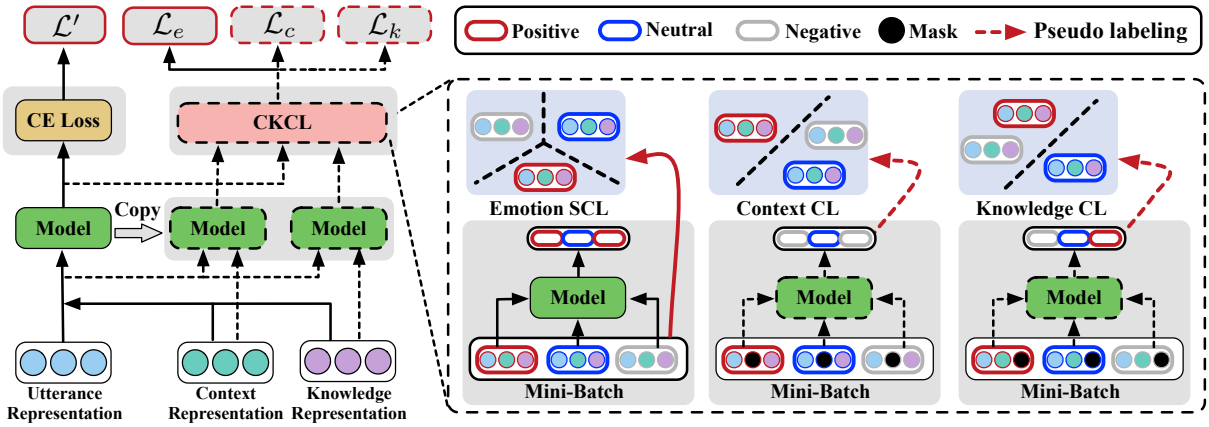



Figure 2: The proposed CKCL framework. The dotted line represents data flows without backpropagation. CE denotes cross-entropy, which is widely adopted in ERC.

employed dropout to augment data. To effectively use label information, Gunel et al. (2020) extended the self-supervised CL to a **fully-supervised CL** setting, which pulls the samples with the same or different labels together or away, respectively, which also boosts the performance of the model in few-shot learning scenarios. In ERC, Li et al. (2022a) tried to employ the supervised CL (SCL) to pull utterances with different emotions further away to better identify similar emotions. Unfortunately, there is no related work based on unsupervised CL in ERC.

## 3 Proposed CKCL Framework

### 3.1 Task Definition

Let a conversation $C$ consist of utterances $u_1, u_2..., u_n$, where $n$ is the number of utterances. Each utterance $u_i = \{w_1, w_2..., w_m\}$ consists of $m$ tokens. And, there are $g$ participants $S = \{s_1, s_2..., s_g\}, (g \geq 2)$ in $C$. Each utterance $u_i$ is uttered by one of $S$. Then, the ERC task aims to predict the pre-defined emotion label set $Y = \{y_1, y_2, ..., y_e\}$ of each utterance in $C$. However, unlike vanilla emotion recognition, ERC models need to focus on modeling context- and knowledge-sensitive dependencies because of interaction between participants. Accordingly, the above issues can be expressed as follows: $y_i = f\big((u_i, k_i), (u_{i-1}, k_{i-1}), ..., (u_{i-w}, k_{i-w})\big)$, where $w$ is the size of context, $K = \{k_1, k_2, ..., k_n\}$ denotes the external knowledge of utterances.

### 3.2 Overview

CL was first proposed to augment data for improving visual representation. Afterward, due to the poor performance of BERT in semantic text similarity tasks, researchers began to introduce self-supervised CL to capture the correlation and difference between utterances. Subsequently, more and more CL-based methods appeared in the NLP field (Kumar et al., 2022a). Unfortunately, there is no related work based on unsupervised CL in ERC. In this section, given the defect in performance degradation of ERC models in modeling context and external knowledge, we introduce a CL-based framework CKCL to refine the utterance representations during training. Additionally, inspired by Li et al. (2022a), we incorporate a weighted supervised CL into CKCL, to distinguish uneven distribution samples with similar emotions.

### 3.3 Context CL

Context is the core of NLP-related tasks and significantly improves the performance of NLP systems. In ERC, surrounding utterances (at time < t) of the current utterance (at time t) are treated as a context (Poria et al., 2017). However, it is challenging to model context, mainly because of (1) emotional dynamics: self- and inter-personal dependency modeling (Poria et al., 2019), and (2) differences between local and distant historical utterances (Ghosal et al., 2021). Although existing ERC models aid classification performance by modeling context, there is also a marked degeneration behind this because of low-quality context. Specifically, the performance of a model in certain utterances is even worse than the model that does not consider contextual information, which highlights the significance of denoising irrelevant context in ERC. Furthermore, the efficacy of denoising low-quality context has been demonstrated

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

**Input:** $\mathcal{B} = \{x_i, \widehat{x}_i, k_i\}_{i=1}^{N_b}, \ell_c, \ell_k \leftarrow 0, 0$
**Output:** $\mathcal{L}_c$, $\mathcal{L}_k$

1 **for** $i = 1$ **to** $N_b$ **do**
2 $\quad o_i \leftarrow \mathcal{M}(\{x_i, \widehat{x}_i, k_i\})$
3 $\quad o_i^c \leftarrow \mathcal{M}^\dagger(\{x_i, [MASK], k_i\})$
4 $\quad o_i^k \leftarrow \mathcal{M}^\dagger(\{u_i, \widehat{x}_i, [MASK]\})$
5 $\quad \triangleright$ Pseudo labeling for each $x_i$
6 $\quad$ **if** $o_i \neq o_i^c$ *or* $o_i \neq o_i^k$ **then**
7 $\quad\quad z_i^c, z_i^k \leftarrow 0, 0$
8 $\quad\quad$ **if** $o_i == o_i^c$ *and* $o_i \neq o_i^k$ **then**
9 $\quad\quad\quad z_i^c, z_i^k \leftarrow 1, 0$
10 $\quad\quad$ **if** $o_i \neq o_i^c$ *and* $o_i == o_i^k$ **then**
11 $\quad\quad\quad z_i^c, z_i^k \leftarrow 0, 1$
12 $\quad$ **else**
13 $\quad\quad z_i^c, z_i^k \leftarrow 1, 1$
14 $\quad \ell_c^+, \ell_c^-, \ell_k^+, \ell_k^- \leftarrow [\,], [\,], [\,], [\,]$
15 $\quad$ **for** $j = 1$ **to** $N_b$ *and* $i \neq j$ **do**
16 $\quad\quad$ **if** $z_i^c == z_j^c$ **then**
17 $\quad\quad\quad \ell_c^+(x_i) \mathrel{+}= \mathcal{F}(x_i, x_j, \tau)$
18 $\quad\quad$ **if** $z_i^k == z_j^k$ **then**
19 $\quad\quad\quad \ell_k^+(x_i) \mathrel{+}= \mathcal{F}(x_i, x_j, \tau)$
20 $\quad\quad \ell_c^-(x_i) \mathrel{+}= \mathcal{F}(x_i, x_j, \tau)$
21 $\quad\quad \ell_k^-(x_i) \mathrel{+}= \mathcal{F}(x_i, x_j, \tau)$
22 $\quad \ell_c \mathrel{+}= \ell_c^+(x_i) \,/\, \ell_c^-(x_i)$
23 $\quad \ell_k \mathrel{+}= \ell_k^+(x_i) \,/\, \ell_k^-(x_i)$
24 $\triangleright$ Computing contrastive loss for each $x_i$
25 $\mathcal{L}_c \leftarrow -\ell_c / N_b$
26 $\mathcal{L}_k \leftarrow -\ell_k / N_b$

in other NLP tasks (Zhang et al., 2021).

Based on this, we design a context CL to capture the correlation and difference between context-independent and context-dependent utterances. We first copy the model $\mathcal{M}$ and feed the input data $\{x_i, [MASK], k_i\}_{i=1}^{N_b}$ masking context representation $\widehat{x}_i$ of $u_i$ with 0, into the replica model $\mathcal{M}^\dagger$ for each mini-batch $\mathcal{B}$. Especially, the context representation of an utterance is conventionally in the hidden layer of ERC models (Ghosal et al., 2020; Majumder et al., 2019), but there are also models to use context as input (Zhong et al., 2019). Then, we conduct self-supervised pseudo labeling, represented as *Line* 6 - *Line* 12 in Algorithm 1. Finally, we can calculate contrastive loss item $\mathcal{L}_c$ according to the pseudo labels $z^c = \{z_i^c\}_{i=1}^{N_b}$, described as *Line* 14 - *Line* 24 in Algorithm 1.

### 3.4 Knowledge CL

In conversations, humans usually rely on commonsense knowledge to convey emotions (Zhong et al., 2019). However, in knowledge-sensitive ERC models, irrelevant knowledge for understanding the utterance might be absorbed as noise (Tu et al., 2022b). Although there are some works Jiang et al. (2022); Tu et al. (2022a); Zhu et al. (2021) striving for knowledge selection, they are still limited in knowledge-independent utterances as identifying the emotions of these utterances does not necessitate external knowledge. To distinguish between knowledge-independent and knowledge-dependent utterances and denoise irrelevant knowledge, we also design a CL-based method, Knowledge CL. The process of Knowledge CL is similar to that of Context CL, but the difference is that Knowledge CL is masking the knowledge representation, rather than the context representation. As a result, we can obtain another loss item $\mathcal{L}_k$, described in Algorithm 1.

### 3.5 Emotion SCL

Considering the ERC task characteristics, that is, the class distribution is extremely uneven and emotional labels have heightened similarity, we proposed a class-weighted SCL, Emotion SCL, to clarify the representation of utterances with similar emotions. The Emotion SCL can pull samples with different emotional labels further apart and alleviates the impact of the class imbalance problem to a certain extent. The process of Emotion SCL for each mini-batch $\mathcal{B}$ is as follows:

$$\mathcal{L}_e = -\frac{1}{N_b} \sum_{x_i \in \mathcal{B}} \log \ell_e \tag{1}$$

$$x_i = EmbeddingLayer(u_i) \tag{2}$$

$$\ell_e = \frac{\sum_{j=1}^{N_b} \mathbb{1}_{[i \neq j]} \mathbb{1}_{[y_i = y_j]} \alpha_j \cdot \mathcal{F}(x_i, x_j, \tau)}{\sum_{k=1}^{N_b} \mathbb{1}_{[i \neq k]} \mathcal{F}(x_i, x_k, \tau)} \tag{3}$$

where $\mathcal{B}$ denotes a mini-batch sample, $N_b$ is the size of $\mathcal{B}$. $\mathbb{1}_{[.]} \in \{0, 1\}$ represents an indicator function. $\alpha_j$ is the class weight of the $j$-th utterance. $EmbeddingLayer(.)$ represents the word embedding methods. ERC models usually leverage BERT (Vaswani et al., 2017), Glove (Pennington et al., 2014), or Roberta (Liu et al., 2019) to encode utterance representations. $\mathcal{F}(x_i, x_k, \tau) = e^{simi(x_i, x_j)/\tau}$, where $\tau$ is the temperature parameter, $simi(x_i, x_j) = \frac{x_i \cdot x_j}{\|x_i\| \|x_j\|}$ denotes the cosine

similarity function. And $\{y_i\}_{i=1}^{N_b}$ is the emotional label set of utterances in $\mathcal{B}$.

### 3.6 Model Training

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

$$\mathcal{L} = \mathcal{L}' + \gamma_e \mathcal{L}_e + \gamma_c \mathcal{L}_c + \gamma_k \mathcal{L}_k + \lambda \left\| \Theta \right\|^2 \quad (4)$$

where $\gamma_e$, $\gamma_c$ and $\gamma_k$ are tuned hyper-parameters. $\mathcal{L}'$ is the classification loss. $\Theta$ is a set of learnable parameters of the CKCL framework. $\lambda$ represents the coefficient of $L_2$-regularization.

## 4 EXPERIMENTS

### 4.1 Datasets

We conduct experiments on four datasets: IEMOCAP (Busso et al., 2008), Dailydialog (Li et al., 2017), MELD (Poria et al., 2019), and EmoryNLP (Zahiri and Choi, 2018). The statistics of datasets are shown in Table 1.

**IEMOCAP** consists of dyadic sessions where actors perform improvisations or scripted scenarios. And each utterance is labeled with one of the emotions: happy, angry, neutral, sad, excited, or frustrated.

**Dailydialog** is a dyadic conversation dataset from human-written daily communications. And each utterance is annotated with one of the emotions: happiness, surprise, sadness, anger, disgust, neutral, or fear, and one of the sentiments: neutral, negative, or positive.

**MELD** is a multi-party conversation dataset collected from the TV show *Friends*, which is an extension of the EmotionLines dataset (Hsu et al., 2018). Each utterance is annotated with one of the emotions: surprise, fear, disgust, anger, sadness, neutral, or joy, and one of the sentiments: neutral, negative, or positive.

**EmoryNLP** consists of multi-party sessions from the TV show *Friends*, and each utterance is labeled with one of the emotions: surprise, fear, disgust, anger, sadness, neutral, or joy, and one of the sentiments: neutral, negative or positive.

| Dataset | Dialogues | | | Utterances | | |
|---|---|---|---|---|---|---|
| | train | val | test | train | val | test |
| IEMOCAP | 120 | | 31 | 5,810 | | 1,623 |
| DailyDialog | 11,118 | 1,000 | 1,000 | 87,832 | 7,912 | 7,863 |
| MELD | 1039 | 114 | 280 | 9,989 | 1,109 | 2610 |
| EmoryNLP | 659 | 89 | 79 | 7,551 | 954 | 984 |

| Dataset | Classes | Metric |
|---|---|---|
| IEMOCAP | 6 | Weighted Avg F1 |
| DailyDialog | 7 | Macro F1 and Micro F1 |
| MELD | 3 and 7 | Weighted Avg F1 over 3 and 7 classes |
| EmoryNLP | 3 and 7 | Weighted Avg F1 over 3 and 7 classes |

Table 1: Statistics of experimental datasets. Since the IEMOCAP dataset does not provide a predefined train/validation split, we utilize 10% of the training dialogues as the validation split.

| Dataset | $\gamma_e$ | $\gamma_c$ | $\gamma_k$ | $\tau$ |
|---|---|---|---|---|
| IEMOCAP | 0.1 | 0.9 | 0.2 | 0.07 |
| DailyDialog | 0.1 | 0.1 | 0.3 | 0.07 |
| MELD | 0.2 | 0.1 | 0.3 | 0.07 |
| EmoryNLP | 0.3 | 0.3 | 0.6 | 0.07 |

Table 2: Setting of hyper-parameters.

### 4.2 Comparison Models

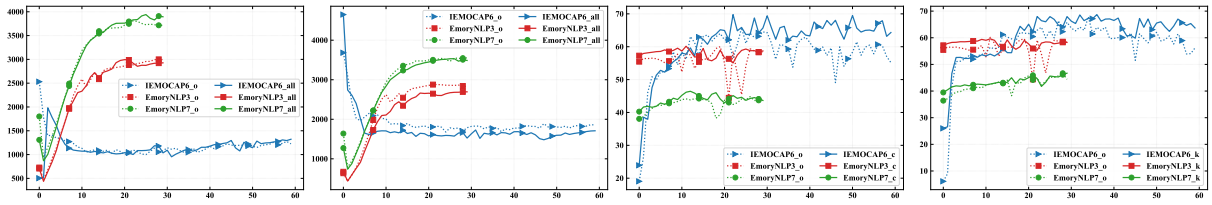
We compare our proposed framework with various ERC baselines, including RNN-based models: COSMIC (Ghosal et al., 2020), DialogueRNN (Majumder et al., 2019); Memory network: AGHMN (Jiao et al., 2020), and Graph-based models: DialogueGCN (Ghosal et al., 2019), and DAG-ERC (Shen et al., 2021); Transformer-based model: KET (Zhong et al., 2019), and state-of-the-art methods: TODKAT (Zhu et al., 2021), CoG-BART (Li et al., 2022a), COSMIC+HCL (Yang et al., 2022) and CauAIN (Zhao et al., 2022). And we also employ ERC baselines as the base model to prove the generalization ability of the CKCL framework.

### 4.3 Experimental Settings

All of the baselines have released their source codes. Thus, we hold identical settings as the original papers. For CKCL, $\gamma_e$, $\gamma_c$, $\gamma_k$, and $\tau$ are tuned manually on each dataset with hold-out validation. Specifically, the hyperparameters setting of COSMIC$^\star$+CKCL reported in Table 2. $\gamma_e$, $\gamma_c$, $\gamma_k$ of baselines are 1: 1: 1 on each dataset and $\tau$ of baselines is always 0.07. The reported results are the average score of 5 random runs on the test set. Additionally, because the models are different in modeling context and knowledge representation, as shown in Table 3, thus, the mask objects are also different during our experiments.

### 4.4 Experimental Results and Analysis

Table 4 reports the experimental results on different datasets. We can observe that the performance on sentiment and emotion identification of COS-




(a) Numbers_C (b) Numbers_K (c) F1_C (d) F1_k

Figure 3: Ablation studies on CKCL. The Numbers_C and Numbers_K represent the number changes of in context- and knowledge-independent utterances. And the F1_C and F1_K denote the weighted avg F1 score of context- and knowledge-independent utterances in the validation set. Especially, the '_o', '_c', _k, and '_all' indicate the COSMIC $^*$, COSMIC $^*$+ Context CL, COSMIC $^*$+ Knowledge CL, and COSMIC $^*$+ CKCL, respectively.

| Model | Context Representation | Knowledge Representation |
|---|---|---|
| COSMIC | Context State | Commonsense Vectors |
| KET | Context Representation | Concept Embedding |
| DialogueRNN | Global State | - |
| DialogueGCN | The Output of Speaker-Level Context Encoding | - |
| DAG-ERC | The Output of DAGERC layers | - |
| AGHMN | Contextual Vectors | - |

Table 3: The mask objects are in various models, which is consistent with the original papers.

MIC based on the CKCL framework is significantly boosted. The COSMIC$^\star$ achieves the best improvement result of F1 on the DailyDialg dataset, i.e., 2.18%. And the COSMIC $^*$+CKCL outperforms all compared methods on different datasets except for the IEMOCAP dataset. This verifies the effectiveness ability of our CKCL framework.

### 4.5 Ablation Study

To investigate the impact of each component of our proposed CKCL framework, we conducted an ablation study on COSMIC$^\star$, and the results are shown in Table 4. **'w/o $\mathcal{L}_{con}$'**, **'w/o $\mathcal{L}_{kno}$'**, and **'w/o $\mathcal{L}_{emo}$'** represent without Context CL, Knowledge CL, and Emotion SCL respectively. The results suggest that all components of the CKCL framework have worked and all the improvements by Context CL, Knowledge CL, and Emotion SCL are statistically significant, as evidenced by the paired t-test results with a p-value < 0.05.

**Analysis of CKCL:** In model training, adaptively distinguishing context- or knowledge-independent utterances can capture a latent feature, reflecting the impact degree of modeling context or knowledge on the prediction results, which enriches utterance representations and the denoising ability of the model in ERC. Reported results in Table 4 also demonstrated it, and the effectiveness of Knowledge CL is superior to that of Context CL. It may be attributed to the model's inherent ability to denoise irrelevant context to some extent, but struggling to effectively handle irrelevant knowledge. In addition, considering the difference in data size between IEMOCAP and Dailydialog datasets, we analyze CKCL on IEMOCAP and EmoryNLP datasets for better visualization. The number of context- or knowledge-independent utterances is shown in Fig. 3, which converges as the training iterations and remains consistent with the model's convergence. This is primarily because the CKCL annotation heavily relies on the model's prediction results. Especially, Fig. 3.(c-d) also demonstrates the capability of CKCL to effectively enhance the performance of context- or knowledge-independent utterances. This further proves the effectiveness of the CKCL framework in denoising irrelevant contexts or knowledge.

**Analysis of Emotion SCL:** To better understand differences of utterance representations with different emotions, we show the t-SNE (van der Maaten and Hinton, 2008) visualization of the intermediate representation of COSMIC$^\star$ and COSMIC$^\star$+ Emotion SCL on IEMOCAP and Dailydialog datasets. Overall, the differences in utterance representations derived from the latter are clearer than the former, as shown in Fig. 5. Specifically, as shown in Fig. 5.(a), Emotion SCL alleviates the difficulty of distinguishing similar emotions such as 'happy' and 'excited' to some extent. Additionally, the utterance representations of the emotion "happy" is effectively differentiated from others. Similarly, in

| Methods | IEMOCAP | DailyDialog | | MELD | | EmoryNLP | |
|---|---|---|---|---|---|---|---|
| | W-Avg F1 | Macro F1 | Micro F1 | W-Avg F1 (3-cls) | W-Avg F1 (7-cls) | W-Avg F1 (3-cls) | W-Avg F1 (7-cls) |
| ♯ DialogueRNN | 62.57 | 41.80 | 55.95 | 66.10 | 57.03 | 48.93 | 31.70 |
| ♮ DialogueGCN | 64.18 | - | - | - | 58.10 | - | - |
| ♮ AGHMN | 62.70 | - | - | | 58.10 | - | - |
| ♠ ♮ KET | 59.56 | - | 53.37 | | 58.18 | - | 34.39 |
| ♠ ♮ COSMIC | 65.28 | 51.05 | 58.48 | 73.20 | 65.21 | 56.51 | 38.11 |
| ♮ DAG-ERC | 68.03 | | 59.33 | - | 63.65 | - | 39.02 |
| ♠ ♮ TODKAT | 61.33 | - | 58.47 | - | 65.47 | - | 38.69 |
| ♮ CoG-BART | 66.18 | - | 55.34 | - | 64.81 | - | 39.04 |
| ♠ ♮ COSMIC + HCL | 66.23 | | 59.54 | - | 65.85 | - | 38.96 |
| ♠ ♮ CauAIN | 67.61 | **53.85** | 58.21 | - | 65.46 | - | - |
| COSMIC⋆ | 65.10 | 51.87 | 58.87 | 73.19 | 64.91 | 56.50 | 38.69 |
| COSMIC⋆+ CKCL | 67.16(↑ *2.06%*) | 53.09(↑ *1.22%*) | **60.96**(↑ *2.18%*) | **73.90**(↑ *0.71%*) | **66.21**(↑ *1.30%*) | 58.18 (↑ *1.68%*) | **40.23**(↑ *1.54%*) |
| **w/o** $\mathcal{L}_e$ | 66.06 (↓ *1.10%*) | 52.51 (↓ *0.58%*) | 59.18 (↓ *1.78%*) | 73.37 (↓ *0.53%*) | 65.55 (↓ *0.66%*) | 57.16 (↓ *1.02%*) | 38.84 (↓ *1.39%*) |
| **w/o** $\mathcal{L}_c$ | 65.92 (↓ *1.18%*) | 52.48 (↓ *0.61%*) | 59.91 (↓ *1.05%*) | 73.44 (↓ *0.46%*) | 65.39 (↓ *0.82%*) | 57.01 (↓ *1.17%*) | 39.41 (↓ *0.82%*) |
| **w/o** $\mathcal{L}_k$ | 65.83 (↓ *1.27%*) | 52.37 (↓ *0.72%*) | 59.37 (↓ *1.59%*) | 73.51 (↓ *0.39%*) | 65.27 (↓ *0.94%*) | 56.88 (↓ *1.30%*) | 38.81 (↓ *1.42%*) |

Table 4: Comparison results on different methods. The best scores are in bold. ⋆ is our replication results, ♮ and ♯ represents results from the original papers and (Ghosal et al., 2020), respectively. ♠ denotes knowledge-sensitive models. W-Avg F1 denotes the weighted avg F1 score. The depth of color symbolizes the declining or rising value.

| Methods | IEMOCAP | DailyDialog | | MELD | | EmoryNLP | |
|---|---|---|---|---|---|---|---|
| | W-Avg F1 | Macro F1 | Micro F1 | W-Avg F1 (3-cls) | W-Avg F1 (7-cls) | W-Avg F1 (3-cls) | W-Avg F1 (7-cls) |
| DialogueRNN⋆ | 62.02 | 38.67 | 52.73 | 66.13 | 57.16 | - | - |
| **w/** CKCL | 63.15(↑ *1.13%*) | 39.38 (↑ *0.71%*) | 53.09 (↑ *0.36%*) | 66.58(↑ *0.45%*) | 57.65(↑ *0.49%*) | - | - |
| DialogueGCN⋆ | 63.83 | 37.89 | 51.53 | - | - | - | - |
| **w/** CKCL | 64.45 (↑ *0.62%*) | 38.66 (↑ *0.77%*) | 51.84 (↑ *0.31%*) | | | - | - |
| DAG-ERC⋆ | 68.03 | 53.29 | 59.16 | - | 63.59 | 59.54 | 39.10 |
| **w/** CKCL | **68.78**(↑ *0.75%*) | 53.82(↑ *0.53%*) | 59.44(↑ *0.28%*) | - | 64.02 (↑ *0.43%*) | **60.65** (↑ *1.11%*) | 39.55 (↑ *0.45%*) |
| AGHMN⋆ | 61.66 | - | - | - | 57.24 | - | - |
| **w/** CKCL | 62.60(↑ *0.94%*) | - | - | - | 57.70 (↑ *0.46%*) | - | - |
| KET⋆ | 57.90 | 48.18 | 53.46 | 63.66 | 57.01 | 51.42 | 34.41 |
| **w/** CKCL | 59.54 (↑ *1.64%*) | 49.27 (↑ *1.09%*) | 54.30 (↑ *0.84%*) | 65.24 (↑ *1.58%*) | 58.32 (↑ *1.31%*) | 52.95 (↑ *1.53%*) | 36.17 (↑ *1.76%*) |

Table 5: Experimental results of generalizability analysis on different baselines and datasets.

Fig. 5.(b), the differentiation between similar emotions like 'happiness' and 'surprise' is also alleviated and the differences among the other emotions become more pronounced.

### 4.6 Analysis of Performance Degradation

Although modeling context and knowledge can enhance performance, it also leads to performance degradation in certain utterances. As shown in Fig. 4, the theoretical performance suggests that the model will not experience performance degradation, meaning that after modeling context and knowledge, the model can also correctly identify utterances that it could correctly identify previously (i.e. when the model lacked modeling context and knowledge). The difference between theoretical and actual performance implies existing ERC systems can not achieve a resultful denoising effect for irrelevant context and knowledge. This is also the primary motivation behind this paper.

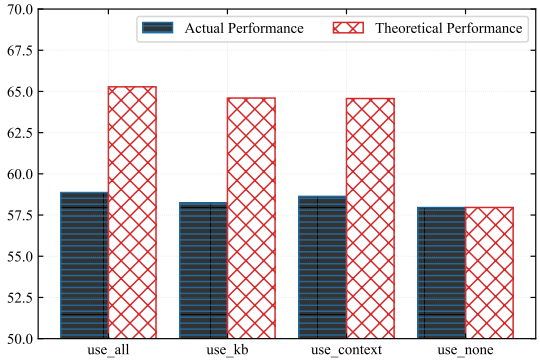
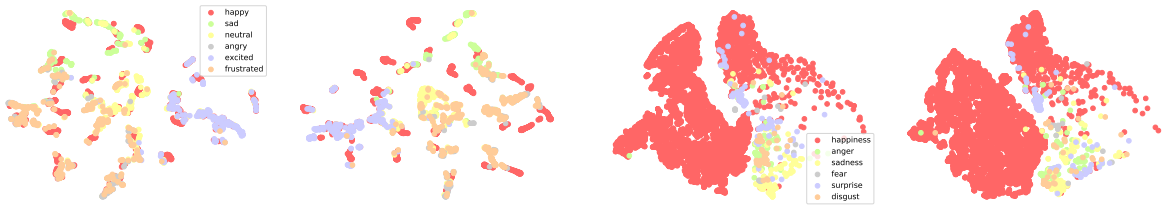



Figure 4: Performance Degradation on Dailydialog dataset. The use_none, use_context, use_kb, and use_all represents the Micro F1 score of COSMIC⋆ masking context and knowledge, masking knowledge, masking context, and masking none, respectively.




(a) IEMOCAP (b) Dailydialog

Figure 5: Visualization of intermediate embeddings of COSMIC$^\star$ (left) and COSMIC$^\star$ + Emotion SCL (right). Because neutral classes for 83% of the DailyDialog dataset, are excluded during visualization.

| ID | Utterances for Prediction | **w/o** CKCL | **w/** CKCL | Golden Label |
|---|---|---|---|---|
| 1 | Exactly, "laugh". | excited | happy | happy |
| 2 | I'm sorry, I'm sorry, I'm sorry. You're right. | neutral | sad | sad |
| 3 | I don't feel so bad then. Well, I'm excited for you. | happy | excited | excited |

Table 6: Examples of utterances from the IEMOCAP and Dailydialog datasets for the case study.

### 4.7 Case Study

Table 6 shows three utterances sampled from the IEMOCAP and Dailydialog datasets. These utterances were initially recognized correctly by COSMIC$^\star$ without modeling context or external knowledge, but upon considering the context and external knowledge, they were actually recognized incorrectly. Intuitively, the emotions of these utterances can be recognized even without context and knowledge, but the model showed disappointing performance, because blindly modeling context and knowledge may deteriorate utterance representation. Fortunately, CKCL can effectively distinguish these utterances, thus absorbing irrelevant context and knowledge as noise to improve the robustness ability of ERC models. As a result, the CKCL-based model can correctly identify these cases as expected.

### 4.8 Generalizability Analysis

To evaluate the generalizability of our CKCL framework, following Yang et al. (2022), we conduct the experiment on various ERC baselines as shown in Table 4. We can see that the improvement effect on the Dailydialog dataset is not in line with expectations, which shows that the influence of CKCL on different models is quite different. Additionally, the CKCL without hyperparameter adjustment can still boost the performance of all models on emotion or sentiment classification. It verifies the effectiveness and generalizability of CKCL in ERC.

### 4.9 Analysis of Static Pseudo Labels

Because the CKCL needs to reason three times for adaptively annotating dynamic pseudo labels, it causes the growth in time complexity. Therefore, we explored a low-time complexity method, that is, in the first epoch, using a trained model to annotate static pseudo labels that remain unchanged during subsequent training, which means there is no need for additional reasoning in the following training. As a price, the model performance has declined to some extent, but fortunately, the CKCL is still adequate for the model, as shown in Table 7.

| Methods | IEMOCAP | DailyDialog | |
|---|---|---|---|
| | W-Avg F1 | Macro F1 | Micro F1 |
| COSMIC$^\star$ | 65.10 | 51.87 | 58.87 |
| COSMIC$^\star$+ CKCL | 67.16(↑ *2.06%*) | 53.09(↑ *1.22%*) | **60.96**(↑ *2.18%*) |
| COSMIC$^\star$+ CKCL★ | 66.22(↑ *1.12%*) | 52.54(↑ *0.67%*) | 59.79(↑ *0.92%*) |

Table 7: Comparison results of dynamic and static pseudo labels on different methods. ★ denotes the CKCL framework with static pseudo labels.

### 4.10 Error Analysis

In this section, we conduct an error analysis on the reported results per dataset and found that most errors are attributed to the following three points:

• **Class imbalance problem:** The unbalanced distribution of classes is the primary cause of errors. In the training set of the MELD dataset, the number of samples is as follows: 'fear': 268, 'disgust': 271, 'sadness': 683, 'anger': 1109, 'surprise': 1205, 'joy': 1743, 'neutral': 4710, which causes the F1 score of emotion 'fear' is as low as 0.0806.

• **Diversity of context modeling:** Unlike knowledge representation, masking the context representation as demonstrated in Table 3 does not completely eliminate the influence of contextual information. This is mainly due to the fact that even RNNs or their variations can capture potential contextual information. As a result, generating pseudo labels for Context CL has become challenging.

• **Limitation of pseudo labeling:** The quality of labeled results in pseudo-label annotation primarily relies on the model's predictions. Thus, the model's performance directly impacts the quality of the labeled samples. Consequently, there is a possibility of leveraging incorrectly labeled samples, which can hinder the model's training. For example, in a specific epoch of the training process, Example 3 in Fig. 1 might be mistakenly labeled as a knowledge-dependent utterance or other. Consequently, such situations can lead to fluctuations in the model's performance during training, potentially even lower than the original model. Nevertheless, the benefits of this approach still outweigh the drawbacks as shown in Fig. 3, because as the model converges, the annotations tend to stabilize.

## 5 Conclusion

In this paper, we propose a novel CKCL framework to enhance utterance representations in ERC. More concretely, we employ Context (or Knowledge) CL to capture the correlation and difference between context-independent and context-dependent (or knowledge-independent and knowledge-dependent) utterances representations, which also enhances the ability of models to denoising irrelevant context or knowledge. Additionally, the Emotion SCL can pull utterances with different labels further apart, and then obtain clearer differences in utterances with similar emotions. Experimental results show that our CKCL framework significantly boosted various ERC models and outperformed state-of-the-art methods.

## Acknowledgements

We thank the anonymous reviewers for their valuable suggestions to improve the overall quality of this manuscript. This work was partially supported by the National Natural Science Foundation of China (62006062, 62176076), Natural Science Foundation of Guangdong 2023A1515012922, the Shenzhen Foundational Research Funding (JCYJ20220818102415032, JCYJ20210324115614039), the Major Key Project of PCL2021A06, Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies 2022B1212010005, Key Technologies Research and Development Program of Shenzhen JSGG20210802154400001.

## Limitations

Although the CKCL performs satisfactorily in ERC, there are still some limitations. Because CKCL primarily concentrates on the effect of modeling context and external knowledge on the prediction results, when met some tasks that do not rely on context and external knowledge, pseudo labels can not be annotated, which causes the paralysis of the CKCL. In addition, when the class distribution of the sample is not uneven, the improvement of Emotion SCL will be weakened.

## References

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779.

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(4):335–359.

Erik Cambria, Yang Li, Frank Z Xing, Soujanya Poria, and Kenneth Kwok. 2020. Senticnet 6: Ensemble application of symbolic and subsymbolic ai for sentiment analysis. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 105–114.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607.

Keith Cortis and Brian Davis. 2021. Over a decade of social opinion mining: a systematic review. *Artificial intelligence review*, 54(7):4873–4965.

Yahui Fu, Shogo Okada, Longbiao Wang, Lili Guo, Yaodong Song, Jiaxing Liu, and Jianwu Dang. 2021. Consk-gcn: conversational semantic-and knowledge-oriented graph convolutional network for multimodal emotion recognition. In *2021 IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Cosmic: Commonsense knowledge for emotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2470–2481.

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 *In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1435–1449.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 154–164.

James J Gross and Lisa Feldman Barrett. 2011. Emotion generation and emotion regulation: One or two depends on your point of view. *Emotion Review*, 3(1).

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2020. Supervised contrastive learning for pretrained language model fine-tuning. *arXiv preprint arXiv:2011.01403*.

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*, pages 2594–2604.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. Emotionlines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*.

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*, pages 1–12.

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.

Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2021. Self-guided contrastive learning for bert sentence representations. 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 2528–2540.

Taewoon Kim and Piek Vossen. 2021. Emoberta: Speaker-aware emotion recognition in conversation with roberta. *arXiv preprint arXiv:2108.12009*.

Pranjal Kumar, Piyush Rawat, and Siddhartha Chauhan. 2022a. Contrastive self-supervised learning: review, progress, challenges and future research directions. *International Journal of Multimedia Information Retrieval*, pages 1–28.

Shivani Kumar, Anubhav Shrimal, Md Shad Akhtar, and Tanmoy Chakraborty. 2022b. Discovering emotion and reasoning its flip in multi-party conversations using masked memory network and transformer. *Knowledge-Based Systems*, pages 108–112.

Dayu Li, Xiaodan Zhu, Yang Li, Suge Wang, Deyu Li, Jian Liao, and Jianxing Zheng. 2021a. Enhancing emotion inference in conversations with commonsense knowledge. *Knowledge-Based Systems*, 232:107449.

Jiangnan Li, Zheng Lin, Peng Fu, and Weiping Wang. 2021b. Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge. In *In Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1204–1214.

Jingye Li, Donghong Ji, Fei Li, Meishan Zhang, and Yijiang Liu. 2020. Hitrans: A transformer-based context-and speaker-sensitive model for emotion detection in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4190–4200.

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

Wei Li, Wei Shao, Shaoxiong Ji, and Erik Cambria. 2022b. Bieru: Bidirectional emotional recurrent unit for conversational sentiment analysis. *Neurocomputing*, 467:73–82.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the 8th International Joint Conference on Natural Language Processing*, pages 986–995.

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

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*.

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 multimodal emotional dynamics in a conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2694–2704.

Weizhi Nie, Rihao Chang, Minjie Ren, Yuting Su, and Anan Liu. 2021. I-gcn: Incremental graph convolution network for conversation emotion detection. *IEEE Transactions on Multimedia*, pages 4471–4481.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing*, pages 1532–1543.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics*, pages 873–883.

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.

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*, pages 1551–1560.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the 31st AAAI conference on artificial intelligence*, pages 4444–4451.

Geng Tu, Bin Liang, Dazhi Jiang, and Ruifeng Xu. 2022a. Sentiment- emotion- and context-guided knowledge selection framework for emotion recognition in conversations. *IEEE Transactions on Affective Computing*, pages 1–14.

Geng Tu, Bin Liang, Dazhi Jiang, and Ruifeng Xu. 2022b. Sentiment-emotion-and context-guided knowledge selection framework for emotion recognition in conversations. *IEEE Transactions on Affective Computing*, (01):1–14.

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

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. page 6000–6010.

Yan Wang, Jiayu Zhang, Jun Ma, Shaojun Wang, and Jing Xiao. 2020. Contextualized emotion recognition in conversation as sequence tagging. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 186–195.

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.

Guorong Xiao, Geng Tu, Lin Zheng, Teng Zhou, Xin Li, Syed Hassan Ahmed, and Dazhi Jiang. 2020. Multimodality sentiment analysis in social internet of things based on hierarchical attentions and csat-tcn with mbm network. *IEEE Internet of Things Journal*, 8(16):12748–12757.

Yunhe Xie, Kailai Yang, Cheng-Jie Sun, Bingquan Liu, and Zhenzhou Ji. 2021. Knowledge-interactive network with sentiment polarity intensity-aware multitask learning for emotion recognition in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2879–2889.

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

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 5065–5075.

Lin Yang, Yi Shen, Yue Mao, and Longjun Cai. 2022. Hybrid curriculum learning for emotion recognition in conversation. 36(10):11595–11603.

Sayyed M Zahiri and Jinho D Choi. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *Workshops at the thirty-second AAAI conference on artificial intelligence*, pages 44–52.

Chenwei Zhang, Yaliang Li, Nan Du, Wei Fan, and Philip S Yu. 2021. Entity synonym discovery via multipiece bilateral context matching. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 1431–1437.

Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019. Modeling both context-and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 5415–5421.

Weixiang Zhao, Yanyan Zhao, and Xin Lu. 2022. Cauain: Causal aware interaction network for emotion recognition in conversations. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence*, pages 4524–4530.

Lin Zheng, Naicheng Guo, Weihao Chen, Jin Yu, and Dazhi Jiang. 2020. Sentiment-guided sequential recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1957–1960.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 165–176.

Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-driven and knowledge-aware transformer for dialogue emotion detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 1571–1582.

## ACL 2023 Responsible NLP Checklist

### A For every submission:

☑ A1. Did you describe the limitations of your work?
*Section 4.8, Section 4.9 and Limitations*

☑ A2. Did you discuss any potential risks of your work?
*Limitations*

☑ A3. Do the abstract and introduction summarize the paper's main claims?
*Abstract and Section 1*

☒ A4. Have you used AI writing assistants when working on this paper?
*Not used*

### B ☒ Did you use or create scientific artifacts?

*Left blank.*

☐ B1. Did you cite the creators of artifacts you used?
*No response.*

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
*No response.*

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
*No response.*

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
*No response.*

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
*No response.*

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
*No response.*

### C ☑ Did you run computational experiments?

*Section 4.4 - 4.10*

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
*The data scale is not large, the running time is not very long, and the experiment can be done on ordinary equipment. In addition, we also optimized the time complexity in section 4.9.*

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*Section 4.3*

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
*Section 4.4 - 4.5*

☒ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
*The models used in this framework are open source and can be downloaded directly*

**D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?**

*Left blank.*

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
*No response.*

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
*No response.*

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
*No response.*

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*No response.*

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
*No response.*