# Knowing What and Why: Causal Emotion Entailment for Emotion Recognition in Conversations

Hao Liu[a,1], Runguo Wei[b,1], Geng Tu[b], Jiali Lin[c], Dazhi Jiang[b,*] and Erik Cambria[d,*]

[a] *School of Computer Science, Northwestern Polytechnical University, Xi'An, China*
[b] *Department of Computer Science of Shantou University, Shantou, China*
[c] *Business School of Shantou University, Shantou, China*
[d] *College of Computing and Data Science, Nanyang Technological University, Singapore*

ARTICLE INFO

*Keywords*:
Emotion recognition
Causal emotion entailment
Reasoning network

ABSTRACT

The clues for eliciting emotion deserve attention in the realm of Emotion Recognition in Conversations (ERC). In an ideal dialogue system, comprehending emotions alone is insufficient, and underlying the causes of emotion is also imperative. However, previous research overlooked the integration of causal emotion entailment for a prolonged period. Therefore, an emotion-cause hybrid framework that utilizes causal emotion entailment (CEE) is proposed to promote the ERC task. Specifically, the presented method integrates the information of the cause clause extracted through the CEE module that triggers emotions into the utterance representations obtained by the ERC model. Moreover, a Bidirectional Reasoning Network (BRN) is designed to extract emotional cues, to simulate human complex emotional cognitive behavior. Experimental results demonstrate that our framework achieves a new state-of-the-art performance on different datasets, indicating that the proposed framework can improve the model's ability to emotion understanding.

## 1. Introduction

Emotion recognition in conversations (ERC) plays a pivotal role in the field of Artificial Intelligence (AI) (Koolagudi and Rao, 2012; Cambria et al., 2023). For example, ERC can be implemented in human-computer interaction, opinion mining, sarcasm detection, etc. (Zhu et al., 2024; Lee and Hong, 2016; Hazarika et al., 2018; Liu et al., 2024a,b; Cambria et al., 2024). The emotional content of an utterance is influenced by various factors, such as the conversational context and Causal Emotion Entailment (CEE) (Majumder et al., 2019; Poria et al., 2021). Existing research on ERC mainly uses recurrent neural networks (Majumder et al., 2019; Hazarika et al., 2018) to obtain the dependencies between utterances or use graph-based structure (Shen et al., 2021; Saxena et al., 2022) to gain long-term information.

Additionally, transformer-based models are also employed in this task (Chudasama et al., 2022; Luo et al., 2024; Tu et al., 2024). However, these methods neglect to uncover the causes of emotion generation and its utilization, failing to understand and utilize the information associated with emotions entirely. Figure. 1 illustrates the connection between cause clauses and emotion clauses. Moreover, these methods tend to ignore partial context information when extracting contextual utterance representation. To address the above issues, we introduce two modules, namely Causal Emotion Entailment (CEE) and Bidirectional Reasoning Network (BRN), into the ERC model.

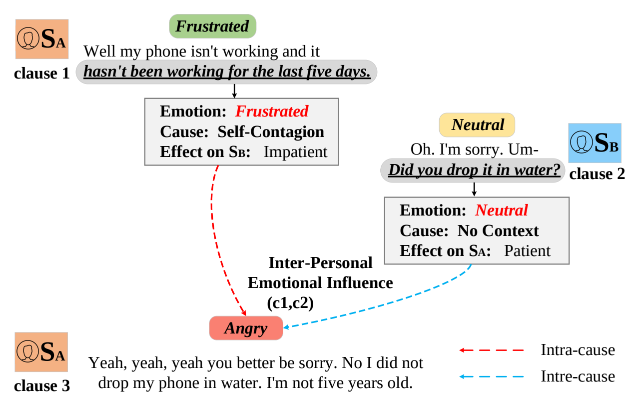



**Figure 1:** Example of causal emotion entailment. The dotted line indicates that the emotion of the specified utterance is influenced by the cause clause associated with it. The first utterance of Speaker A indicates that he has already been impatient with Speaker B. But Speaker B apologizes and misinterprets Speaker A's question, which makes Speaker A annoyed again. Because the emotion-cause pairs between utterances like Speaker A's first utterance and Speaker B's second utterance directly triggered Speaker A's anger. This contextual pairing of emotion-cause helps predict participants' emotion labels.

These modules extract causal information triggering emotions in utterances, offering a significant opportunity to address the identified issues. Recent works for CEE tasks are based on graph networks. Poria et al. (Poria et al., 2021) set some baselines for CEE tasks, such as ECPE-MLL, and RankCP, which all use graph attention networks to extract relations between utterances. zhang et al. (Zhang et al., 2022) used graph neural networks to provide interaction between utterances and integrate speaker information.

*Corresponding author
liuhao23@mail.nwpu.edu.cn ( Hao Liu); dzjiang@stu.edu.cn ( Dazhi Jiang); cambria@ntu.edu.sg ( Erik Cambria)
ORCID(s): 0000-0001-8593-9276 ( Hao Liu); 0000-0001-8453-0364 ( Runguo Wei); 0000-0003-0781-9126 ( Dazhi Jiang); 0000-0002-3030-1280 ( Erik Cambria)
[1] The authors contribute equally to this work.

In addition, some works, such as KEC and KBCIN (Li et al., 2022; Zhao et al., 2023) introduce commonsense knowledge into graph neural networks to improve the information extraction ability of the model. MPEG (Chen et al., 2023) fuses speaker and sentiment information via a heterogeneous graph attention network to capture the inter-utterances causal relationship. Unlike the above methods, in this paper, we use window transformer. Because transformers perform well in capturing the context of conversations. In addition, causal clauses often appear around emotion clauses, and window transformer can effectively interact inter-utterances semantic information with limited window size. Specifically, in CEE, the utterance is first encoded through the encoding layer. Then we model the inter-clause document through the 2D Window Transformer (Ding et al., 2020), which is proficient in effectively extracting the semantic correlation between emotion and their underlying cause clauses in conversations. Inspired by the theory of emotion perception, it posits that individuals infer the emotional states of others by observing their emotional expressions. To dynamically simulate human emotional and cognitive behavior, we utilize LSTM to capture the contextual information of the conversation and incorporate it into bidirectional reasoning. Firstly, the context of different stages of the emotion analysis model is stored in static memory nodes. We use LSTM to integrate and extract this contextual information to grasp the internal logic of conversation utterances and extract emotional cues. Concurrently, we update memory information dynamically. Finally, through multiple iterative processes, we conduct conscious emotional cognitive reasoning in the conversation, enhancing the accuracy of simulating human emotional and cognitive behavior in conversation. To summarize, this paper makes the following contributions:

1. We first combine the CEE module with the ERC task so that the model can use the information of the cause clause associated with the utterance used for emotion prediction.
2. We propose the BRN module to imitate human emotion and cognitive behavior in dynamic conversations.
3. Experiments on different conversational datasets showcase that our proposed approach enhances multiple baselines and surpasses state-of-the-art ERC methods.

## 2. Related Work

Deep learning plays a significant role in human activities (Huang et al., 2022; Jia et al., 2020; Wang et al., 2023; Fan et al., 2024). Emotion analysis around conversations is an important topic in recent years, which has attracted much attention in natural language processing. The availability of many conversation datasets partly explains this phenomenon, and the growing interest in conversational emotion-cause pairs can also explain this phenomenon. In the following paragraphs, we divide the related works into two categories according to the problems they use to model the conversation context.

### 2.1. Emotion Recognition in Conversations

Rosalind (Picard, 2010) proposes that emotion analysis is an interdisciplinary science that involves psychology, cognitive science, and deep learning. Erik et al. has conducted a comprehensive and proactive exploration of emotional analysis(Susanto et al., 2020). With the widespread use of convolutional neural networks (Sun et al., 2021) and generative adversarial networks (Tang et al., 2021; Liu et al., 2019), deep learning is also applied in affective computing. Considering the dynamic interaction between speakers, some researchers (Hazarika et al., 2018; Majumder et al., 2019) leverage a recurrent neural network to model different speakers to obtain context information. Jiang et al. (Jiang et al., 2023) applied fuzzy neural network to emotion detection. Due to the recurrent neural network having a long-term information propagation problem, DialogueGCN (Ghosal et al., 2019), DAG-ERC (Shen et al., 2021) and HSGCF (Wang et al., 2023) employ graph convolution neural network and directed acyclic graph to model the dialogue context and simulate the information interaction between speakers, respectively. DualGATs (Zhang et al., 2023) constructs a dual graph network. To enrich the utterance representation, KET (Zhong et al., 2019), SKSEC(Tu et al., 2023a), CKCL (Tu et al., 2023b) and COSMIC (Ghosal et al., 2020) introduce external knowledge into the emotion analysis model by using Knowledge Graph, such as ConceptNet (Liu and Singh, 2004) and COMET (Bosselut et al., 2019), while TODKAT (Zhu et al., 2021) carries out topic detection, and integrates commonsense into a transformer to obtain richer context representation. Li (Li et al., 2023) proposed a Knowledge Integrated Model. To alleviate the issue of category imbalance in emotional data, Tu (Tu et al., 2023) introduced label bias. However, they cannot deal with the problems of difficulty in distinguishing similar emotions and emotion transfer. Therefore, Yang et al. (Yang et al., 2022) constructed a hybrid learning architecture to alleviate the problems of emotion transfer and confusion labeling in conversational emotion. SACL (Hu et al., 2023) propose the Supervised Adversarial Contrastive Learning to learn structured representations between classes. Multitask learning (Jiang et al., 2021; Tu et al., 2022), self-supervised learning (Jiang et al., 2024) and contrastive learning (Tu et al., 2023b) are also applied in emotion identification.

### 2.2. Causal Emotion Entailment

To explore the causes of emotion expression, early researchers proposed a task called emotion cause extraction (ECE) (Lee et al., 2010), which aims to extract the reasons behind a certain emotional expression in text. ECE task typically requires emotional expression in advance. Correspondingly, Xia (Xia and Ding, 2019) proposes the emotion-cause pair extraction task (ECPE) to extract potential emotion-cause pairs in documents and formulate a two-step solution. These tasks all process document data. Unlike emotion cause extraction and emotion-cause pair extraction tasks, the goal of emotion cause entailment is to identify the utterances that trigger the emotion of a specific utterance in a conversation.

**Table 1**
Comparison of existing research methods for CEE and ERC. The commonsense denotes the models inject the commonsense knowledge into network to improve the utterance representation learning. ♯ denotes the model includes both the encoder and decoder. The code of this table is available, which can be searched at github.

| | Model | Methods (Utterance encoder) |
|---|---|---|
| CEE | RECCON (Poria et al., 2021) | Transformer; GCN |
| | TSAM (Zhang et al., 2022) | Attention, GNN |
| | KEC (Li et al., 2022) | Commonsense; GNN |
| | KBCIN (Zhao et al., 2023) | Commonsense, GAT |
| | Ours (Jiang et al., 2023) | 2D Window Transformer |
| ERC | DialogueRNN (Majumder et al., 2019) | RNN |
| | DialogueGCN (Ghosal et al., 2019) | GCN |
| | DAG-ERC (Shen et al., 2021) | Directed Acyclic Graph |
| | KET (Zhong et al., 2019) | Commonsense; Transformer |
| | COSMIC (Ghosal et al., 2020) | Commonsense; RNN |
| | CKCL (Tu et al., 2023b) | Commonsense; Transformer |
| | TODKAT ♯ (Zhu et al., 2021) | Commonsense; Transformer |
| | MM-DFN (Hu et al., 2022) | Multimodal GCN |
| | DualGATs (Zhang et al., 2023) | dual graph network |
| | SACL (Hu et al., 2023) | Contrastive Learning |
| | Ours | BRN; CEE(Transformer) |

Previous works, such as (Li et al., 2019; Li and Xu, 2014), use ECE to solve text-based emotion classification from the perspective of finding emotion-cause and achieved excellent results. Meanwhile, graph construction and transformer (Li et al., 2022; Jiang et al., 2023) are used in this task. For example, TSAM (Zhang et al., 2022) uses a graph network to model speakers. Since the dependencies relationships between cause utterances with different emotions from the target utterances are difficult to extract, KEC (Li et al., 2022) introduces social commonsense knowledge into graph convolution networks to improve the model's reasoning ability for cause utterances. In addition, KBCIN (Zhao et al., 2023) uses commonsense knowledge to build a bridge-interaction network to enhance the understanding of the conversational context. Compared with other conversational tasks, the CEE task is most closely related to conversational emotion recognition at utterance-level. However, no researchers combined CEE with ERC.

To clearly demonstrate the differences between existing works and our works, the mainstream models and our proposed model are shown in Table 1.

# 3. Methodology

## 3.1. Task Definition

Let $U = \{u_1, u_2, \cdots, u_N\}$ be a conversation, where $N$ denotes the utterance quantity. And there is a set $S = \{S_1, S_2, \cdots, S_M\}$ consists of $M$ speakers. Each utterance $u_i$ is spoken by the speaker $S_{\varphi(u_i)}$, where $\varphi$ maps the index of the utterance into that of the corresponding its speaker. We also represent $u_i \in \mathbb{R}^{D_m}$ as the feature representation of the utterance. The task of ERC aims to predict the emotion labels of each constituent utterance $u_i$ from the pre-defined emotion labels ($happy$, $excited$, $neutral$, $angry$, $sad$, $frustrated$, $disgust$, $fear$). The CEE aims to extract all potential pairs comprised of emotion and corresponding cause clauses from the document annotated with emotion and cause labels in the conversational context. Given a document $d = [u_1, u_2, \cdots, u_i, \cdots, u_{|d|}]$, the purpose of the CEE task is to obtain a series of emotion-cause pairs $= \{\cdots, (u^e, u^{c_1}), \cdots, (u^e, u^{c_k}), \cdots\}$ where $u^e$ is an emotion clause and $u^{c_k}$ is the corresponding $k-th$ cause clause.

## 3.2. The Overall Framework

In this section, we present the overall framework. The framework consists of the Causal Emotion Entailment (CEE) module, the ERC model, and the Bidirectional Reasoning Network (BRN) module. Six mainstream ERC models are used to test the performance of this framework. Figure 2 exhibits the whole structure of the presented framework.

## 3.3. Causal Emotion Entailment

On account of CEE can correlate emotion-cause with the contextual conversation, we apply the CEE module to the ERC task. Current research focuses more on models based on graph neural networks. For example, TSAM (Zhang et al., 2022) and MPEG (Chen et al., 2023) use attention mechanisms and graph networks to fuse speaker and sentiment information. KEC (Li et al., 2022) and KBCIN (Zhao et al., 2023) introduce commonsense knowledge and use graph neural networks. However, they perform poorly in information exchange between sentences at short distances. Unlike current research, we use transformers to extract contextual information when handling CEE tasks. Additionally, rather than modeling the entire conversation or injecting external information to enhance utterance representations, we focus on the impact of inter-utterance interaction in a short span, and use 2D window transformer to interact inter-utterances semantic information with limited window size. Specifically, the 2D Window Transformer is used in the pre-trained process. The given utterances are divided into several windows according to window size. The 2D Window Transformer models the relationship between clauses to get better clause representation. 2D Window Transformer has $N$ encoder layers. Each layer comprises a window attention and a feed-forward layer.

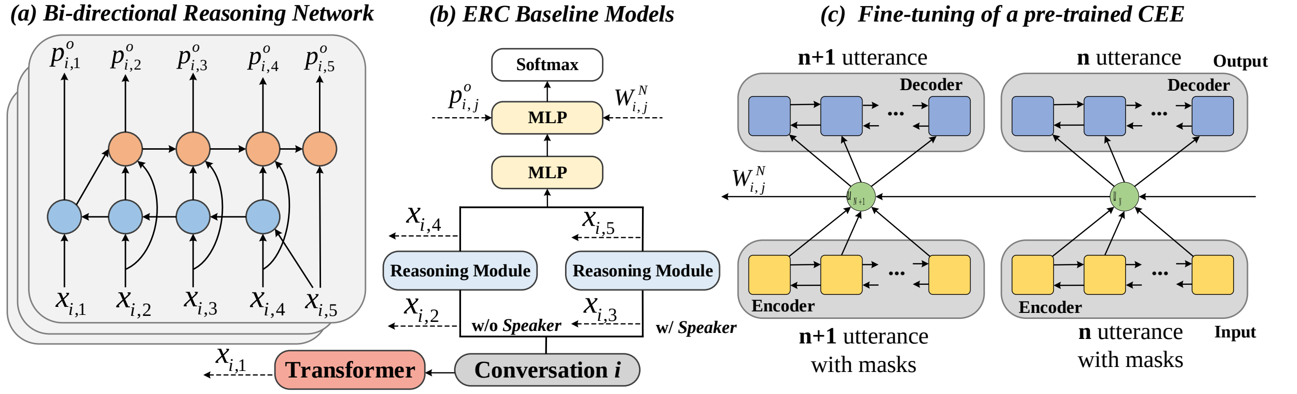



**Figure 2:** The overall framework. The part (a) is the BRN module. For a given conversation, we encode the utterances using the Transformer to obtain $x_{i,1}$ and feed it into the BRN module with the intermediate state vectors $x_{i,2}$, $x_{i,3}$, $x_{i,4}$, $x_{i,5}$ obtained from the ERC model. Part (c) is a pre-trained model. Part (b) is the ERC model, and the part of the ERC model that extracts the utterance context representation is used as a whole as the contextual reasoning module. We concatenate $x_{i,4}$ with $x_{i,5}$ and feed it into the first MLP. The dotted lines indicate the direction of data propagation from different modules to each other. The solid lines mean the direction of information propagation between different nodes within each module, whereas the solid lines in part (c) indicate the process of encoding and decoding the words in the utterance.

The 2D Window Transformer is utilized as the encoder layer of the CEE module. Each utterance pair $(u_i, u_j)$ is fed into embedding layer to get the representation $W_{i,j}$. Firstly, $W_{i,j}$ is calculated by window attention which is multi-head self-attention. The $W_{i,j}$ is fed into three linear layers to calculate the query vector $q_{i,j}$, key vector $k_{i,j}$ and the value vector $v_{i,j}$.

$$q_{i,j} = W_{i,j} W_q \tag{1}$$
$$k_{i,j} = W_{i,j} W_k \tag{2}$$
$$v_{i,j} = W_{i,j} W_v \tag{3}$$

where $W_q \in \mathbb{R}^{n \times n}$, $W_k \in \mathbb{R}^{n \times n}$ and $W_v \in \mathbb{R}^{n \times n}$ are learned parameters. For the three vectors $q_{i,j}$, $k_{i,j}$ and $v_{i,j}$, the weight $\beta_{i,j}$ and the output of window attention is calculated as follows:

$$\beta_{i,j} = softmax\left(\frac{k_{i,j}^T \cdot q_{i,j}}{\sqrt{n}}\right) \tag{4}$$
$$z_{i,j} = v_{i,j} \beta_{i,j}^T \tag{5}$$

where $z_{i,j}$ is the output of window attention. The input for the feed-forward layer is $z_{i,j}$ entered into a layer that has two identical constructions followed by a normalization layer at its output:

$$o_{i,j,1} = dropout\left(z_{i,j} W_1 + b_1\right) \tag{6}$$
$$o_{i,j,2} = o_{i,j,1} + dropout\left(o_{i,j,1} W_2 + b_2\right) \tag{7}$$
$$o_{i,j} = o_{i,j,2} + norm\left(o_{i,j,2}\right) \tag{8}$$

where the *norm* denotes laynorm layer. $o_{i,j,1}$ and $o_{i,j,2}$ and are the output of the two sublayers, respectively. $o_{i,j}$ is the output of a encoder layer in 2D Window Transformer.

$$W_{i,j}^{t+1} = o_{i,j}^t \tag{9}$$

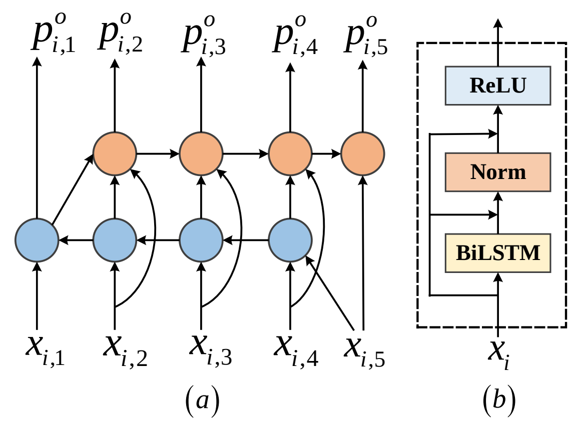



**Figure 3:** The Bidirectional Reasoning Network (BRN) module. (a) is the overall structure of BRN, and (b) is the structure of each node in BRN. The $x_i$ and $p_i$ are inputs and outputs. The input $x_{i,n}$, $(n \in [1, \cdots, 5])$ represents the utterances representation of the model at different stages.

where the output $W_{i,j}^N$ of the last layer is the representation of utterance pair $(u_i, u_j)$ extracted by 2D Window Transformer. The relative position modeling is used to learn the representation of clauses pair and ranks the candidate clauses.

By saving the pre-trained weight and transferring this model, we can convert low-level clause representation to high-level representation, which contains information about the cause evoking the clause. The utterance representation extracted by the pre-trained model is expressed as $md_i$.

### 3.4. Bidirectional Reasoning Network

To simulate human emotional and cognitive behavior in dynamic conversations, we design the BRN module to capture the context of emotional information. Instead of employing BiLSTM to extract contextual information, each node in the BRN structure contains a BiLSTM. Each node extracts the semantic information of the utterances and fuses the utterances representations of different stages of the model through a bidirectional pathway structure. Through nulti-turn iteration, BRN can simulate the human emotional reasoning process. The BRN module has $N$ layers and the structure of a layer is shown in Figure 3 (a). Specially, the input of BRN module is $(x_{i,1}, x_{i,2}, x_{i,3}, x_{i,4}, x_{i,5})$, and there are two pathways for information fusion. In the right-to-left pathway, the output $p^{rl}_{i,k}$ of each node is calculated as follows:

$$p^{rl}_{i,k} = cell\left(w_{k,1} \cdot x_{i,k} + w_{k,2} \cdot p^{rl}_{i,k+1}\right) \tag{10}$$

where $p^{rl}_{i,5} = x_{i,5}$ and $w_{k,i}$ is trainable weight that can be a scalar. $cell$ is the node of the BRN module. In the left-to-right pathway, the output of each node is calculated as follows:

$$p^{o}_{i,k} = cell\left(w'_{k,1} \cdot x_{i,k} + w'_{k,2} \cdot p^{rl}_{i,k} + w'_{k,3} \cdot p^{o}_{i,k-1}\right) \tag{11}$$

where $p^{o}_{i,1} = x_{i,1}$ and $w'_{k,i}$ is trainable weight that can be a scalar. $k$ is the index of the cell in the BRN module. The Figure 3 $(b)$ exhibits the structure of $cell$. This utterance representation is fed into BiLSTM, which is followed by the norm layer and activation layer. The output of these cells for the input $x_i$ can be computed as:

$$\begin{aligned} c_i &= norm(x_i + BiLSTM(x_i)) && (12) \\ p^{o}_i &= ReLU\left(x_i + c_i\right) && (13) \end{aligned}$$

where the norm layer is LayerNorm and the activation layer uses the ReLU function. The output of the current layer is the input of the next layer.

$$x^{l+1}_{i,k} = p^{l}_{i,k} \tag{14}$$

where $l$ is the index of the layer. Then, the output of the whole memory fusion network is obtained by concatenating the final layer's output. In general, given $(x_{i,1}, x_{i,2}, x_{i,3}, x_{i,4}, x_{i,5})$, the vector $mf_i$ extracted by of BRN module can be defined as:

$$mf_i = BRN(x_{1,i}, x_{2,i}, x_{3,i}, x_{4,i}, x_{5,i}) \tag{15}$$

where the $mf_i$ is the output of the BRN module.

### 3.5. Emotion Classifier

Based on the output vectors $md_i$, $mf_i$ obtained from the BRN module and the CEE module, respectively. We concatenate them with the vector $mc_i$ obtained from the ERC model's last layer and fuse them using MLP to gain the utterance representation $o_i$.

$$o_i = MLP\left(\left[md_i; mf_i; mc_i\right]\right) \tag{16}$$

where the $o_i$ is the final representation fed into the emotion classification layer employed for emotion prediction:

$$\hat{y}_i = softmax(W_o o_i + b_o) \tag{17}$$

The cross-entropy loss function is applied to calculate the loss value to optimize the model:

$$loss = -\frac{1}{\sum_{l=1}^{L} c(l)} \sum_{i=1}^{L} \sum_{k=1}^{c(i)} y^{l}_{i,k} \log(\hat{y}^{l}_{i,k}) \tag{18}$$

where $L$ is the number of the conversation. $c\,(l)$ denotes the number of utterance in the conversation $i$. $y^{l}_{i,k}$ and $\hat{y}^{l}_{i,k}$ are the true label of utterance $i$ in conversation $l$ and the possibility of predicting the result of category $k$, respectively.

## 4. Experimental Settings

### 4.1. Datasets and Evaluation Metrics

Our framework is evaluated on the following datasets: IEMOCAP (Busso et al., 2008), DailyDialog (Li et al., 2017). The detailed statistics of the datasets are reported in Table 2.

- **IEMOCAP** is a multimodal dataset for emotion recognition that is comprised of videos of multi-turn dialogues of ten unique speakers. The utterances are annotated with one of six emotion labels, namely *happy, excited, neutral, angry, sad, and frustrated*.
- **DailyDialog** is an emotion detection dataset that contains the conversations of our daily life and human-written daily communications. There are seven emotion labels annotated by three professional persons in this dataset: *disgust, fear, sadness, angry, neutral, joy, surprise*.

Because of the uneven distribution of the DailyDialog dataset, the percentage of utterances with the neutral label is 83%, so we adopt the Micro F1 and Macro F1, excluding the neutral samples. We follow the previous research (Majumder et al., 2019) to use average Accuracy (Acc.) and Weighted F1 on the IEMOCAP dataset. In this paper, we leverage MaF to represent the Macro F1, WF to represent the Weighted F1, and MiF to represent the Micro F1, respectively.

### 4.2. Baselines

To evaluate the performance of the proposed framework, we compare it with several baselines.

1. **KET** (Zhong et al., 2019) introduces external commonsense knowledge into a transformer architecture through self-attention and graph-attention mechanisms.
2. **VHRED** (Hazarika et al., 2021) uses a pre-trained sentence encoder and simulates the inter-sentence context through transfer learning to identify the emotion.

**Table 2**
The statistics of datasets. Statistics of splits and evaluation metrics used in different datasets.

| Datasets | Conversations | | | Utterances | | |
|---|---|---|---|---|---|---|
| | Train | Val | Test | Train | Val | Test |
| IEMOCAP | 120 | 12 | 31 | 5810 | | 1623 |
| DailyDialog | 11,118 | 1,000 | 1,000 | 87,832 | 7,912 | 7,863 |

| Datasets | Classes | Evaluation |
|---|---|---|
| IEMOCAP | 6 | Accuracy and Weighted F1 |
| DailyDialog | 7 | Macro F1 and Micro F1 |

3. **DialogueRNN** (Majumder et al., 2019) models context and speaker separately using GRU to obtain global context dependencies and speaker dependencies, meanwhile using global GRU for speaker-to-speaker interaction.
4. **DialogueGCN** (Ghosal et al., 2019) models different speakers using GCN pairs after capturing contextual information separately and classifies the emotions of utterance representations by attention mechanism.
5. **BiERU** (Li et al., 2022): construct a Bidirectional sentiment recursive unit by utilizing many GRU to detect emotion.
6. **RoBERTa** (Zhang et al., 2020) uses the pre-trained RoBERTa to obtain utterance representation and fine-turn the prediction layer.
7. **COSMIC** (Ghosal et al., 2020) employs RoBERTa (Zhang et al., 2020) to extract the data of this paper and introduce commonsense knowledge like mental state, causality, etc. Using the pre-trained model COMET (Bosselut et al., 2019) and feed them into the emotion analysis model.
8. **DialogueCRN** (Hu et al., 2021) processes utterances representation by using BiLSTM and attention mechanism to simulate the human cognitive.
9. **SKAIG** (Li et al., 2021) introduces commonsense knowledge into the graph structure.
10. **DAG-ERC** (Shen et al., 2021) combines traditional graph-based models with recursive-based neural models.
11. **MM-DFN** (Hu et al., 2022) leverages graph-base construction learning the intra- and inter-modal relationship of utterances.

In this paper, we employ the pre-trained model 840B GloVe (Pennington et al., 2014) to obtain the utterance representation with a dimension of 300. The extracted utterance representations are then fed into a network consisting of a convolutional layer, maximum pooling, and fully connected layers to extract text features. The final vector with a dimension of 100 is used as the text feature.

In addition to using GloVe (Pennington et al., 2014) as the feature extractor, the pre-trained model BERT (Kenton and Toutanova, 2019) and RoBERTa (Zhang et al., 2020) is also applied to extract context-independent text features. The output vector of the final layer of the pre-trained model is used as the text feature.

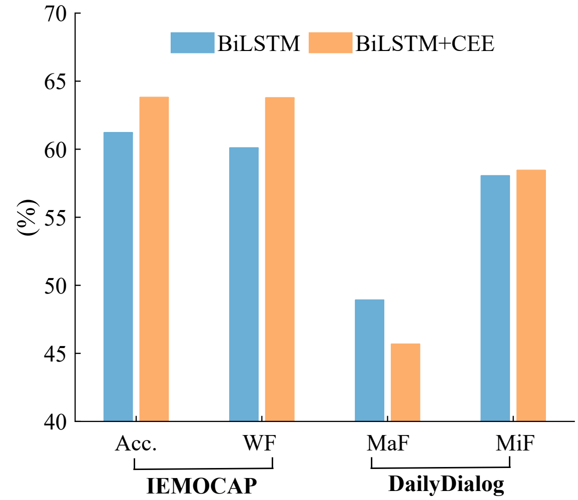



**Figure 4:** Experimental results for the verification of CEE validity on two different datasets.

### 4.3. Hyperparameters Settings

We conduct hyperparameters search for our proposed framework on IEMOCAP and DailyDialog datasets. We employ Adam optimization with a batch size of 32, epochs of 50, the learning rate of $\{1e-5, 2e-5\}$, L2 weight decay of 2e-4, and dropout of $\{0.3, 0.2\}$. The number of 2D Window Transformer's encoder layers is 3 and the window size is 4 in the CEE module. The number of layers in the BRN module is 2.

## 5. Results and Discussions

### 5.1. The Role of Causal Emotion

To prove the validity of CEE in the ERC model, we conduct one of the most classical models BiLSTM for emotion recognition, where the RoBERTa is applied to extract the textual features. The results are displayed in Figure 4. Compared with using BiLSTM only for emotion classification, the experimental results of BiLSTM+CEE are improved. This improvement underscores how incorporating CEE enhances the model's ability to capture causal emotional cues, thereby strengthening its overall effectiveness.

### 5.2. Experimental Results and Analysis

Our framework based on the DialogueCRN model is compared with the baselines in Table 3. To obtain our results in Table 2, we employ RoBERTa to extract text features. As expected, our framework outperforms all the baselines. On the IEMOCAP dataset, we achieve a new state-of-the-art Acc. of 69.01% and WF of 69.07%. Compared with the previous work, our framework gains 0.80%, and 0.89% in terms of Acc. and WF. On the DailyDialog dataset, our framework gets a 1.44% and 0.06% improvement on MaF and MiF.

**Table 3**
The experimental results. The results in bold are the best-performing ones under each column. The best values are highlighted in bold. All the results of the comparable baselines can be found in papers (Shen et al., 2021; Hu et al., 2022).

| | Models | IEMOCAP | | DailyDialog | |
|---|---|---|---|---|---|
| | | Acc. | WF | MaF | MiF |
| GloVe-based | 1.KET | - | 59.56 | - | 53.37 |
| | 2.VHRED | - | 58.60 | - | 48.40 |
| | 3.DialogueRNN | 63.03 | 62.50 | - | 50.56 |
| | 4.DialogueGCN | 65.25 | 64.18 | - | - |
| | 5.BiERU | 63.02 | 63.13 | - | - |
| | 6.DialogueCRN | 65.25 | 65.21 | - | - |
| RoBERTa-based | 7.RoBERTa | - | 63.38 | 48.20 | 55.16 |
| | 8.COSMIC | - | 65.28 | 51.05 | 58.48 |
| | 9.SKAIG | - | 66.98 | 51.95 | 59.75 |
| | 10.DAG-ERC | - | 68.08 | - | 59.33 |
| | 11.MM-DFN | 68.21 | 68.18 | - | - |
| | ours(RoBERTa) | **69.01** | **69.07** | **53.39** | **59.81** |

**Table 4**
The experimental results of generalization analysis. The best values are highlighted in bold.

| Models | IEMOCAP | | DailyDialog | |
|---|---|---|---|---|
| | Acc. | WF | MaF | MiF |
| BiERU | 63.22 | 63.52 | 29.35 | 52.79 |
| **ours+BiERU** | 62.57 | 62.45 | **39.30** | **56.37** |
| DialogueRNN | 64.20 | 64.21 | 39.69 | 56.19 |
| **ours+DialogueRNN** | **66.42** | **66.37** | **51.29** | **58.59** |
| DialogueCRN | 66.54 | 66.11 | 52.25 | 58.28 |
| **ours+DialogueCRN** | **69.01** | **69.07** | **53.39** | **59.81** |

**Table 5**
The results of significance tests for generalization analysis (P-Value) (Legends: P-Values < 0.05).

| Datasets | BiLSTM | BiERU | DialogueRNN | DialogueCRN |
|---|---|---|---|---|
| **IEMOCAP** | 9.74e-3 | 7.84e-3 | 9.46e-3 | 3.67e-8 |
| **DailyDialog** | 4.57e-6 | 6.58e-6 | 9.85e-3 | 1.67e-2 |

In order to explain the gaps in experimental results, it is essential to understand the logical relationships of the conversations. Previous works focus on model speakers or introducing external knowledge to enrich contextual representations. They both encode the utterances, but none of them consider the connection between utterance and its associated cause clauses. In the process of contextual information propagation, the ERC model gradually loses the information between the cause clauses associated with the current utterance. In contrast, we solve these two problems with the CEE and BRN modules.

When constructing a emotion analysis model in conversational system, it is crucial to analyze emotional changes in dynamic conversations in real-time. The CEE task is used to extract the entailed cause information contained in the clause, which utilizes the utterance information produced before the current utterance. Additionally, when identifying the emotion of an utterance, we can input this utterance into the model together with the preceding utterances, and then use the proposed framework to reason and identify the emotion labels through these utterances. Because, this method is evaluated on the datasets IEMOCAP and Dailydailogue, which are English dialogue datasets. When adapting it to new language scenarios, we can fine-tune on cross-lingual emotion recognition datasets in future work and use some domain adaptation techniques.

## 5.3. Generalization Analysis

Based on the promising results of the CEE+BiLSTM model, we extended our framework to the DialogueCRN model to evaluate the effectiveness of combining the CEE module and BRN module with the ERC model. To assess the generalizability of our framework, we compared the results of two emotion recognition models, BiERU (Li et al., 2022) and DialogueRNN (Majumder et al., 2019), as shown in Table 4. The results of Table 3 indicate that RoBERTa outperforms GloVe in extracting textual features. Therefore, in this section, we use RoBERTa to extract text features. Overall, the experimental results demonstrate the effectiveness of our framework. The results of the significance tests on the compared models are reported in Table 5, which demonstrates that our framework is significantly different from the comparison models.

Moreover, in order to elucidate the usage of RoBERTa as a feature extractor and explore the adaptability of the proposed framework to ERC tasks, we employed three pre-trained models, namely GloVe, BERT, and RoBERTa, to extract text representations. Extensive comparative experiments were conducted on frameworks utilizing these three extractors, with the DialogueCRN model employed in the experiments. The experimental results, as depicted in Figure 5, indicate that the RoBERTa-based framework outperforms the others, suggesting that the representations derived from a more powerful extractor, such as RoBERTa, yield greater benefits for emotional recognition.

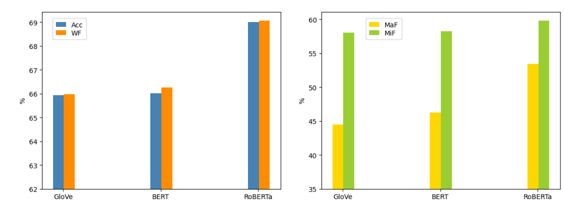
## 5.4. Ablation Study

To investigate the contribution of the proposed modules, we conducted several ablation studies on the DialogueCRN model, where each constituent component was removed individually. As shown in Table 6, the performance decreases slightly when either the BRN or CEE module is removed, indicating the significance of both modules.

**Analysis of Bidirectional Reasoning Network:** As demonstrated in Table 6, when RoBERTa is employed to extract text features, the results are enhanced on the IEMOCAP dataset, with the evaluation indicators Acc. and WF increasing by 1.36% and 2.08%, respectively.

**Table 6**
The experimental results of ablation studies on IEMOCAP and Dailydialog datasets (Legends: $\mathfrak{R}$ represents the CEE Pre-trained Model. The features used in the ERC task are extracted by RoBERTa. The best values are highlighted in bold).

| | BRN | CEE | IEMOCAP | | | | | | DailyDialog | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $\mathfrak{R}_{BERT}$ | | | $\mathfrak{R}_{RoBERTa}$ | | | $\mathfrak{R}_{BERT}$ | | | $\mathfrak{R}_{RoBERTa}$ | | |
| | | | Acc. | MaF | WF | Acc. | MaF | WF | Acc. | MaF | MiF | Acc. | MaF | MiF |
| RoBERTa-based | × | × | 66.54 | 65.81 | 66.11 | 66.54 | 65.81 | 66.11 | 83.13 | **52.35** | 58.28 | 83.13 | 52.35 | 58.28 |
| | × | ✓ | 67.71 | 66.81 | 67.43 | 68.33 | 66.07 | 67.97 | 84.33 | 50.32 | 58.37 | 85.12 | 49.01 | 58.38 |
| | ✓ | × | 67.90 | **67.11** | 68.19 | 67.90 | 67.11 | 68.19 | 80.13 | 49.78 | 56.27 | 80.13 | 49.78 | 56.27 |
| | ✓ | ✓ | **68.86** | 67.04 | **68.58** | **69.01** | **68.15** | **69.07** | **85.63** | 51.10 | **59.30** | **85.56** | **53.39** | **59.81** |




(a) (b)

**Figure 5:** Performance on different feature extractors. (a) IEMOCAP. (b) DailyDialog.

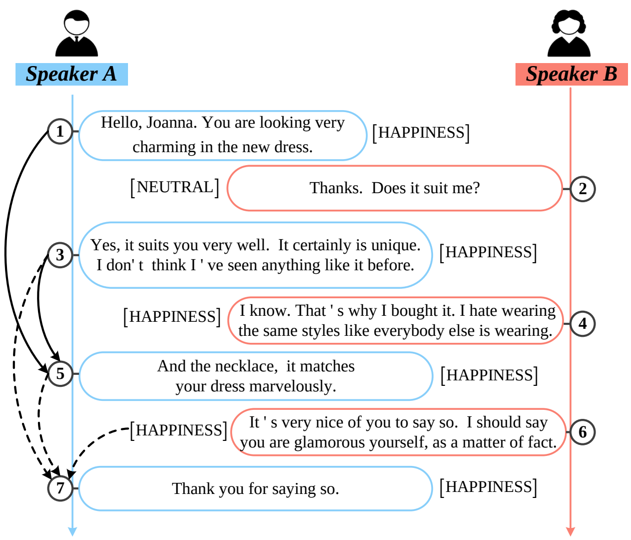
Unlike the reasoning network in DialogueCRN, which directly processes utterance representations through lstm-attention, while Bidirectional Reasoning Network (BRN) has a bidirectional structure. BRN does not simply act as a neural network at a certain layer of the model but dynamically processes different levels of utterance representation of the emotion recognition model to conduct emotional reasoning. These results showcase the effectiveness of the bidirectional reasoning network in extracting emotion cues by utilizing the representation from the intermediate layers of the ERC model. Moreover, it successfully simulates the process of human-like emotional reasoning in conversations through multiple iterations, enhancing the overall comprehension and coherence of the framework.

**Analysis of Causal Emotion Entailment:** Table 6 presents the results of our experiments, which show that using CEE in the DialogueCRN model improves the performance, regardless of whether BERT or RoBERTa is used as the text feature extractor. Specifically, when RoBERTa is used, the evaluation metrics Acc. and WF on the IEMOCAP dataset show an improvement of 1.79% and 1.86%, respectively, with values of 68.33% and 67.97%. However, the results of MaF and MiF on the DailyDialog dataset show weak improvement, with values of 49.01% and 58.38%, respectively. This indicates that focusing on the corresponding cause information through the CEE module can enhance the emotion reasoning ability of the conversation emotion recognition model, particularly in multi-round conversations.

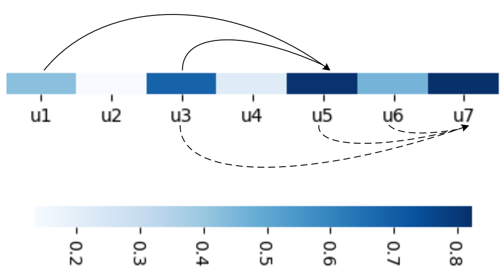
In the case of combining BRN with CEE, the result is better than that using one of them alone on IEMOCAP and DailyDialog datasets. We can draw a conclusion that ablation concerning both modules simultaneously leads to a higher drop in ERC model performance. That shows that the BRN module and the CEE module can complement each other. Although our method is evaluated on the IEMOCAP and DailyDialog datasets, it is not constrained to conversations with two participants. According to our ablation analysis, the BRN proposed in this paper utilizes the model's multiple intermediate layer outputs to extract emotional cues and simulate human emotional cognitive behavior. The CEE module enables the model to focus on relevant causal information, enhancing its performance. Importantly, both modules do not limit the number of speakers in a conversation. Table 4 also presents the results of applying our method to several emotion detection models, which can be used in dialogue systems involving multiple speakers. To better adapt to multi-speaker conversations, it is essential to annotate corresponding labels in the multi-speaker conversation dataset. This may pose some challenges, such as emotional dependency within conversations, which is the emotional interactions between speakers. We can consider fusing personalized factors into sentence representation extraction.

## 5.5. Case Study

In this section, we present a case study on a conversation example from the Dailydialog dataset in Figure 6, which shows the cause clause's role in the utterance's emotion. To validate the efficacy of our framework, two individuals were tasked with annotating emotion-cause clause labels for each utterance in the case study samples, indicating whether the utterance contains causal information. The connecting lines in the graph depict the emotion-cause relationships between the utterances. Furthermore, we visualized the attention layer of the final ERC model for enhanced comprehension. The emotion label of **utterance 5** and **utterance 7** is easily predicted to be *neutral*, while the actual label is *happiness*. As shown by the solid lines in Figure 6, the cause clauses of **utterance 5** are **utterance 1** and **utterance 3**. At the same time, the dotted lines indicate that the cause clauses of **utterance 7** are **utterance 3**, **utterance 5**, and **utterance 6**.




(a) A conversation from DailyDialog dataset.




(b) The visualization of attention weights of the utterances in conversation. Each cell represents a utterance ($u_i$).

**Figure 6:** Case study of a conversation from the DailyDialog.

When not performing the CEE, **utterance 5** and **utterance 7** obtain contextual information by using the ERC model, we will incorrectly predict their emotion as *neutral* after a few rounds of training. In contrast, with the CEE task, the model enhances the effection of the cause clause on the associated utterance emotion. During model training, the contextual representation of **utterance 5** will contain more information related to **utterance 1** and **utterance 3**, and the contextual representation of **utterance 7** will contain more information related to **utterance 3**, **utterance 5**, and **utterance 6**, which causes **utterance 5** and **utterance 7** to be correctly predicted as *happiness*. The illustrated utterances show that CEE highlights the influence of some clauses on a particular utterance.

### 5.6. Emotion Interpretability Analysis

In this section, we conduct a emotion interpretability analysis on a conversation and display the visualization of each utterance. The conversation is sampled from the conversational dataset Dailydailogue.

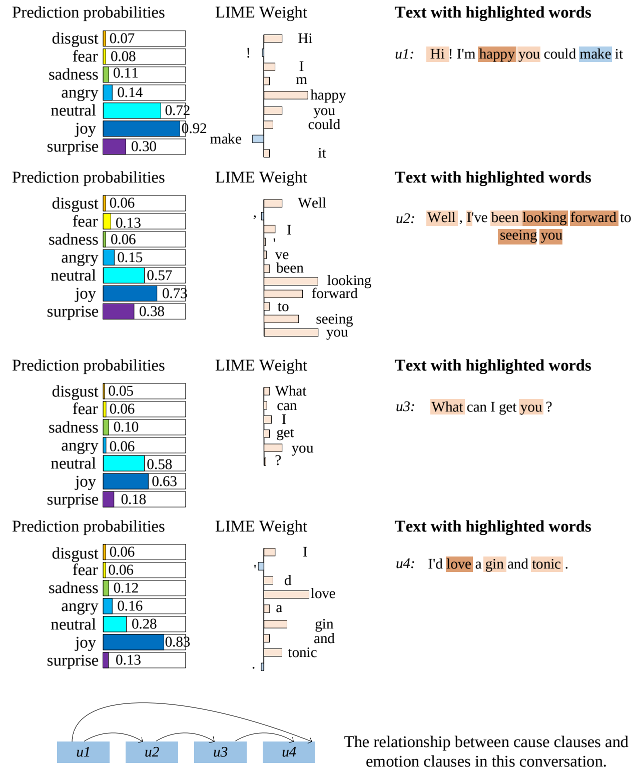
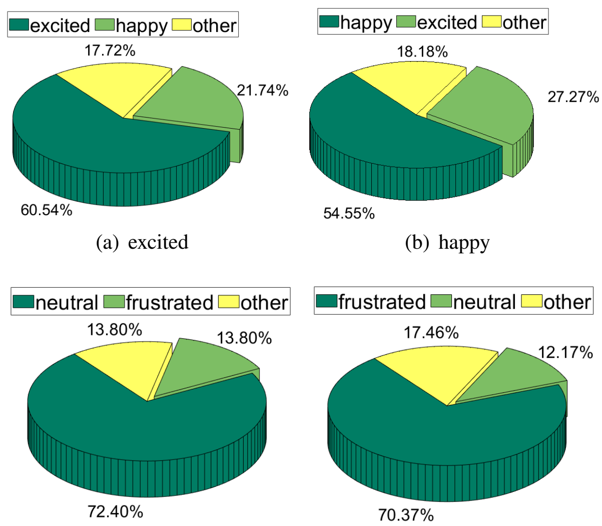



**Figure 7:** The emotion Interpretability analysis of a short conversation in DailyDialog dataset.

We employ the model interpretability tool LIME (Ribeiro et al., 2016) to conduct the emotion interpretability analysis. Figure 7 demonstrates the analysis result. The four utterances in this conversation are annotated with the emotion label of joy. The prediction probabilities for the neutral and joy categories are higher, which may be due to the large proportion of neutral samples in the DailyDialog dataset. Additionally, emotions opposite to joy, such as disgust and fear, have lower prediction probabilities, which further suggests the influence of emotional inertia in conversations on emotion analysis.

By examining the weights of the highlighted words in the text, we can observe that the words influencing the emotion of the utterances are typically among the more important ones. These words entail the cause information for the emotion clauses. For example, in utterances u1, u3, and u4, the utterances u1 and u3 plays a guiding role for u4, thereby affecting u4 emotion. Meanwhile, the highlighted words and their corresponding weights reveal a strong correlation between utterances. Each utterance in the dialogue affects each other. This indicates that the model can simulate human emotional cognition by reasoning on the contextual semantic information extracted from the conversation.




(a) excited (b) happy (c) neutral (d) frustrated

**Figure 8:** Experimental results of error analysis. Four similar emotion prediction results on the IEMOCAP dataset. For example, excited is misclassified as happy and other. Its percentages are as shown in the subgraph (a) above.

### 5.7. Error Analysis

Although our framework has shown strong performance, it still has some limitations. The analysis of our experimental results indicates that our model struggles to effectively distinguish between similar emotion categories such as excited, happy, neutral, and frustrated.Figure 8 illustrates the classification results of our experiments on these four emotion categories in the IEMOCAP dataset. A similar situation is observed in the DailyDialog dataset. We suspect that this difficulty arises because utterances with similar emotions have similar semantic information in the extracted features.

Furthermore, our experiments are limited to text data, whereas multimodal data can provide additional information for non-neutral emotions in utterances. For example, videos may show a disappointed expression for utterances with sad emotions, and utterances with angry emotions may have a higher pitch. However, due to the limitations of the CEE task, not all utterances used for emotion analysis can obtain information about their corresponding cause clauses, which ultimately limits the performance of our framework.

## 6. Conclusions and Future Work

This paper proposes a framework that combines CEE and BRN to enhance the ability of emotion analysis in ERC. Specifically, our framework emphasizes the causal clause that triggers emotions via the CEE module and addresses the issue of disregarding the context of other clauses when the CEE module is integrated with the ERC model, with the help of the BRN module. The proposed framework achieves state-of-the-art result on two public conversational emotion recognition datasets.

Nevertheless, there are still some shortcomings, such as the limitation of datasets in CEE tasks. Therefore, we plan to annotate the dataset by taking context and personal relationships into account in future work. And apply the proposed method to more conversational scenarios. Besides, we will pay attention to incorporating multimodal information into this framework and effectively fusing commonsense information.

## Acknowledgment

The authors would like to respect and thank all reviewers for their constructive and helpful review. This research is funded by the National Natural Science Foundation of China (62372283, 62206163), Science and Technology Major Project of Guangdong Province(STKJ2021005, STKJ202209002, STKJ2023076), Natural Science Foundation of Guangdong Province (2024A1515010239), and Key Project of Colleges and Universities in Guangdong Province (Natural Science) (2024ZDZX1019).

## Declarations

**Data availability** The datasets used in this paper are public and available.

**Ethical Approval** This article does not contain any studies with human participants or animals performed by any of the authors.

**Conflict of Interest** The authors declare that they have no conflict of interest.

## References

S. G. Koolagudi, K. S. Rao, Emotion recognition from speech: a review, International journal of speech technology 15 (2012) 99–117.

E. Cambria, R. Mao, M. Chen, Z. Wang, S.-B. Ho, Seven pillars for the future of artificial intelligence, IEEE Intelligent Systems 38 (2023) 62–69.

L. Zhu, R. Mao, E. Cambria, B. J. Jansen, Neurosymbolic AI for personalized sentiment analysis, in: Proceedings of the International Conference on Human-Computer Interaction (HCII), Washington DC, USA, 2024.

J. Lee, I. B. Hong, Predicting positive user responses to social media advertising: The roles of emotional appeal, informativeness, and creativity, International Journal of Information Management 36 (2016) 360–373.

D. Hazarika, S. Poria, S. Gorantla, E. Cambria, R. Zimmermann, R. Mihalcea, CASCADE: Contextual sarcasm detection in online discussion forums, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 1837–1848.

H. Liu, R. Wei, G. Tu, J. Lin, C. Liu, D. Jiang, Sarcasm driven by sentiment: A sentiment-aware hierarchical fusion network for multimodal sarcasm detection, Information Fusion 108 (2024a) 102353.

H. Liu, B. Yang, Z. Yu, A multi-view interactive approach for multimodal sarcasm detection in social internet of things with knowledge enhancement, Applied Sciences 14 (2024b).

E. Cambria, X. Zhang, R. Mao, M. Chen, K. Kwok, SenticNet 8: Fusing emotion AI and commonsense AI for interpretable, trustworthy, and explainable affective computing, in: Proceedings of the International Conference on Human-Computer Interaction (HCII), Washington DC, USA, 2024.

N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, E. Cambria, DialogueRNN: An attentive RNN for emotion detection in conversations, in: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, 2019, pp. 6818–6825.

S. Poria, N. Majumder, D. Hazarika, D. Ghosal, R. Bhardwaj, S. Y. B. Jian, P. Hong, R. Ghosh, A. Roy, N. Chhaya, et al., Recognizing emotion cause in conversations, Cognitive Computation 13 (2021) 1317–1332.

D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, R. Zimmermann, ICON: Interactive conversational memory network for multimodal emotion detection, in: Proceedings of the 2018 conference on empirical methods in natural language processing, 2018, pp. 2594–2604.

W. Shen, S. Wu, Y. Yang, X. Quan, 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), 2021, pp. 1551–1560.

P. Saxena, Y. J. Huang, S. Kurohashi, Static and dynamic speaker modeling based on graph neural network for emotion recognition in conversation, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop, 2022, pp. 247–253.

V. Chudasama, P. Kar, A. Gudmalwar, N. Shah, P. Wasnik, N. Onoe, M2fnet: Multi-modal fusion network for emotion recognition in conversation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4652–4661.

M. Luo, H. Fei, B. Li, S. Wu, Q. Liu, S. Poria, E. Cambria, M.-L. Lee, W. Hsu, Panosent: A panoptic sextuple extraction benchmark for multimodal conversational aspect-based sentiment analysis, in: Proceedings of ACM Multimedia, 2024, pp. 7667–7676.

G. Tu, T. Niu, R. Xu, B. B. Liang, E. Cambria, Adaclf: An adaptive curriculum learning framework for emotional support conversation, IEEE Intelligent Systems 39 (2024) 5–11.

D. Zhang, Z. Yang, F. Meng, X. Chen, J. Zhou, Tsam: A two-stream attention model for causal emotion entailment, in: Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 6762–6772. URL: https://aclanthology.org/2022.coling-1.588.

J. Li, F. Meng, Z. Lin, R. Liu, P. Fu, Y. Cao, W. Wang, J. Zhou, Neutral utterances are also causes: Enhancing conversational causal emotion entailment with social commonsense knowledge, in: L. D. Raedt (Ed.), Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, International Joint Conferences on Artificial Intelligence Organization, 2022, pp. 4209–4215. URL: https://doi.org/10.24963/ijcai.2022/584. doi:10.24963/ijcai.2022/584, main Track.

W. Zhao, Y. Zhao, Z. Li, B. Qin, Knowledge-bridged causal interaction network for causal emotion entailment, in: Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence, 2023, pp. 14020–14028.

T. Chen, Y. Shen, X. Chen, L. Zhang, S. Zhao, Mpeg: A multi-perspective enhanced graph attention network for causal emotion entailment in conversations, IEEE Transactions on Affective Computing (2023) 1–14.

Z. Ding, R. Xia, J. Yu, Ecpe-2d: emotion-cause pair extraction based on joint two-dimensional representation, interaction and prediction, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 3161–3170.

S. Huang, Y. Liu, I. W. Tsang, Z. Xu, J. Lv, Multi-view subspace clustering by joint measuring of consistency and diversity, IEEE Transactions on Knowledge and Data Engineering (2022).

B. Jia, J. Lv, X. Peng, Y. Chen, S. Yang, Hierarchical regulated iterative network for joint task of music detection and music relative loudness estimation, IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2020) 1–13.

L. Wang, Z. Yu, H. Yu, S. Liu, Y. Xie, B. Guo, Y. Liu, Adaevo: Edge-assisted continuous and timely dnn model evolution for mobile devices, IEEE Transactions on Mobile Computing (2023).

C. Fan, J. Lin, R. Mao, E. Cambria, Fusing pairwise modalities for emotion recognition in conversations, Information Fusion 106 (2024) 102306.

R. W. Picard, Affective computing: from laughter to ieee, IEEE Transactions on Affective Computing 1 (2010) 11–17.

Y. Susanto, A. G. Livingstone, B. C. Ng, E. Cambria, The hourglass model revisited, IEEE Intelligent Systems 35 (2020) 96–102.

Y. Sun, G. G. Yen, B. Xue, M. Zhang, J. Lv, Arctext: A unified text approach to describing convolutional neural network architectures, IEEE Transactions on Artificial Intelligence 3 (2021) 526–540.

C. Tang, Z. He, Y. Li, J. Lv, Zero-shot learning via structure-aligned generative adversarial network, IEEE transactions on neural networks and learning systems 33 (2021) 6749–6762.

D. Liu, J. Fu, Q. Qu, J. Lv, Bfgan: backward and forward generative adversarial networks for lexically constrained sentence generation, IEEE/ACM Transactions on Audio, Speech, and Language Processing 27 (2019) 2350–2361.

D. Jiang, H. Liu, R. Wei, G. Tu, Csat-ftcn: A fuzzy-oriented model with contextual self-attention network for multimodal emotion recognition, Cognitive Computation (2023) 1–10.

D. Ghosal, N. Majumder, S. Poria, N. Chhaya, A. Gelbukh, Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 154–164.

B. Wang, G. Dong, Y. Zhao, R. Li, Q. Cao, K. Hu, D. Jiang, Hierarchically stacked graph convolution for emotion recognition in conversation, Knowledge-Based Systems (2023) 110285.

D. Zhang, F. Chen, X. Chen, DualGATs: Dual graph attention networks for emotion recognition in conversations, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 7395–7408. URL: https://aclanthology.org/2023.acl-long.408. doi:10.18653/v1/2023.acl-long.408.

P. Zhong, D. Wang, C. Miao, 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 (EMNLP-IJCNLP), 2019, pp. 165–176.

G. Tu, B. Liang, D. Jiang, R. Xu, Sentiment- emotion- and context-guided knowledge selection framework for emotion recognition in conversations, IEEE Transactions on Affective Computing 14 (2023a) 1803–1816.

G. Tu, B. Liang, R. Mao, M. Yang, R. Xu, 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, Association for Computational Linguistics, Toronto, Canada, 2023b, pp. 14054–14067. URL: https://aclanthology.org/2023.findings-acl.883. doi:10.18653/v1/2023.findings-acl.883.

D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, S. Poria, Cosmic: Commonsense knowledge for emotion identification in conversations, in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 2470–2481.

H. Liu, P. Singh, Conceptnet—a practical commonsense reasoning tool-kit, BT technology journal 22 (2004) 211–226.

A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, Y. Choi, Comet: Commonsense transformers for automatic knowledge graph construction, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 4762–4779.

L. Zhu, G. Pergola, L. Gui, D. Zhou, Y. He, Topic-driven and knowledge-aware transformer for dialogue emotion detection, in: Proceedings of ACL-IJCNLP, 2021, pp. 1571–1582.

W. Li, L. Zhu, R. Mao, E. Cambria, SKIER: A symbolic knowledge integrated model for conversational emotion recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, 2023, pp. 13121–13129.

G. Tu, R. Jing, B. Liang, M. Yang, K.-F. Wong, R. Xu, A training-free debiasing framework with counterfactual reasoning for conversational emotion detection, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 15639–15650. URL: https://aclanthology.org/2023.emnlp-main.967. doi:10.18653/v1/2023.emnlp-main.967.

L. Yang, Y. Shen, Y. Mao, L. Cai, Hybrid curriculum learning for emotion recognition in conversation, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 11595–11603.

D. Hu, Y. Bao, L. Wei, W. Zhou, S. Hu, Supervised adversarial contrastive learning for emotion recognition in conversations, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 10835–10852. URL: https://aclanthology.org/2023.acl-long.606. doi:10.18653/v1/2023.acl-long.606.

D. Jiang, R. Wei, H. Liu, J. Wen, G. Tu, L. Zheng, E. Cambria, A multitask learning framework for multimodal sentiment analysis, in: 2021 International Conference on Data Mining Workshops (ICDMW), 2021, pp. 151–157. doi:10.1109/ICDMW53433.2021.00025.

G. Tu, J. Wen, H. Liu, S. Chen, L. Zheng, D. Jiang, Exploration meets exploitation: Multitask learning for emotion recognition based on discrete and dimensional models, Knowledge-Based Systems 235 (2022) 107598.

D. Jiang, H. Liu, G. Tu, R. Wei, E. Cambria, Self-supervised utterance order prediction for emotion recognition in conversations, Neurocomputing 577 (2024) 127370.

D. Jiang, H. Liu, G. Tu, R. Wei, Window transformer for dialogue document: a joint framework for causal emotion entailment, International Journal of Machine Learning and Cybernetics (2023) 1–11.

D. Hu, X. Hou, L. Wei, L. Jiang, Y. Mo, 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), IEEE, 2022, pp. 7037–7041.

Y. M. Lee, Y. Chen, S. Li, C.-r. Huang, Emotion cause events: Corpus construction and analysis, in: 7th International Conference on Language Resources and Evaluation, LREC 2010, European Language Resources Association (ELRA), 2010, pp. 1121–1128.

R. Xia, Z. Ding, Emotion-cause pair extraction: A new task to emotion analysis in texts, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 1003–1012.

X. Li, S. Feng, D. Wang, Y. Zhang, Context-aware emotion cause analysis with multi-attention-based neural network, Knowledge-Based Systems 174 (2019) 205–218.

W. Li, H. Xu, Text-based emotion classification using emotion cause extraction, Expert Systems with Applications 41 (2014) 1742–1749.

W. Li, Y. Li, V. Pandelea, M. Ge, L. Zhu, E. Cambria, ECPEC: Emotion-cause pair extraction in conversations, IEEE Transactions on Affective Computing 14 (2022) 1754–1765.

C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, S. S. Narayanan, Iemocap: Interactive emotional dyadic motion capture database, Language resources and evaluation 42 (2008) 335–359.

Y. Li, H. Su, X. Shen, W. Li, Z. Cao, S. Niu, Dailydialog: A manually labelled multi-turn dialogue dataset, in: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2017, pp. 986–995.

D. Hazarika, S. Poria, R. Zimmermann, R. Mihalcea, Conversational transfer learning for emotion recognition, Information Fusion 65 (2021) 1–12.

W. Li, W. Shao, S. Ji, E. Cambria, BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis, Neurocomputing 467 (2022) 73–82.

C. Zhang, Y. Xu, Z. Y. Dong, Robustly coordinated operation of a multi-energy micro-grid in grid-connected and islanded modes under uncertainties, IEEE Transactions on Sustainable Energy 11 (2020) 640–651.

D. Hu, L. Wei, X. Huai, Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 7042–7052.

J. Li, Z. Lin, P. Fu, W. Wang, Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 1204–1214.

J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

J. D. M.-W. C. Kenton, L. K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, 2019, pp. 4171–4186.

M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.

**Hao Liu** has graduated from the Department of Computer Science at Shantou University. Now Now he is pursuing a doctoral degree at Northwestern Polytechnical University, China. His current research interests include affective computing and deep learning.

**Runguo Wei** is currently pursuing the master's degree with the Department of Computer Science at Shantou University, China. His current research focuses on affective computing and machine learning, etc.

**Geng Tu** has graduated from the Department of Computer Science at Shantou University. His current research interests include affective computing and deep learning.

**Jiali lin** is working with the Business School of Shantou University, Shantou University, China, where she is a associate Professor.

**Dazhi Jiang** received his BA in Computer Science from the China University of Geoscience (Wuhan) in 2004. He obtained his PhD from the State Key Laboratory of Software Engineering, Wuhan University, China in 2009. Since then, he has been with the Department of Computer Science, Shantou University, China, where he was a Professor. His research interests include affective computing, deep learning, data mining and applications of artificial intelligence.

**Erik Cambria** is a Professor at Nanyang Technological University, where he also holds the appointment of Provost Chair in Computer Science and Engineering. His research focuses on neurosymbolic AI for interpretable, trustworthy, and explainable affective computing in domains like social media monitoring, financial forecasting, and AI for social good. He is ranked in Clarivate's Highly Cited Researchers List of World's Top 1% Scientists, is recipient of many awards, e.g., IEEE Outstanding Early Career, was listed among the AI's 10 to Watch, and was featured in Forbes as one of the 5 People Building Our AI Future.