

# MPEG: A Multi-Perspective Enhanced Graph Attention Network for Causal Emotion Entailment in Conversations

Tiantian Chen<sup>1</sup>, Ying Shen<sup>1</sup>, Xuri Chen<sup>1</sup>, Lin Zhang<sup>1</sup>, *Senior Member, IEEE*,  
and Shengjie Zhao<sup>1</sup>, *Senior Member, IEEE*

**Abstract**—Emotion causes constitute a pivotal component in the comprehension of emotional conversations. Recently, a new task named Causal Emotion Entailment (CEE) has been proposed to identify the causal utterances for the target emotional utterance in a conversation. Although researchers have achieved some progress in solving this problem, they failed to adequately incorporate speaker characteristics and overlooked the effects of temporal relations in conversation structures. To fill such a research gap to some extent, we propose a novel causal emotion entailment framework, namely MPEG (Multi-Perspective Enhanced Graph attention network). The training of MPEG consists of three stages. First, we utilize a speaker-aware pre-trained model and two attention mechanisms to obtain the utterance representations that incorporate local contexts as well as the speaker and emotional information. Then, these representations are fed into a graph attention network to model the conversation structures and emotional dynamics from both local and global perspectives. Finally, a fully-connected network is implemented to predict the relationships between emotional utterances and causal utterances. Experimental results show that MPEG achieves state-of-the-art performance.

**Index Terms**—Conversational sentiment analysis, causal emotion entailment, graph neural network, dialogue system.

## I. INTRODUCTION

EMOTIONS play a significant role in human communication and understanding [1], [2], [3]. To facilitate emotional communications, current research primarily focuses on two tasks: *emotion recognition* and *emotional response generation*. Emotion recognition refers to identifying emotions conveyed

through utterances in a conversation [4], [5], [6], [7], [8], [9], while emotional response generation aims to generate appropriate emotional responses during conversations [10], [11], [12], [13], [14], [15].

Emotion causes, which refer to the events or situations that trigger or elicit emotions, are essential in both emotion recognition and emotional response generation tasks. They help to infer the speaker's situation, thoughts and emotional states, and thereby facilitate recognizing speaker's emotions and generating responses to the speaker's emotions. However, despite their importance in emotional conversation tasks, there has been a lack of research focused on the extraction of emotion causes. In many studies, emotion cause extraction is often only considered as a supplement to other emotion-related tasks [16], [17], [18], [19], [20], [21]. For example, Zhao et al. [17] utilized the internal and external emotion causes extracted from conversations to predict the emotions of target utterances. If the emotion cause extraction problem cannot be specifically studied, emotional dialogue systems may incorrectly identify the underlying causes of the user's emotions and generate inappropriate responses. To address this research gap, Poria et al. [22] introduced a new task called *Recognizing Emotion Cause in Conversations* (RECCON) in 2021, which aims to extract emotion stimuli during conversations. To support the RECCON task, they created an annotated dataset named RECCON-DD and developed two transformer-based baseline models. According to the granularity of the identified causes, the RECCON task can be further divided into the *Causal Emotion Entailment* (CEE) task and the *Causal Span Extraction* (CSE) task. CEE aims to identify which causal utterances trigger the non-neutral emotions in the target utterances, whereas CSE aims to extract causal spans, i.e., specific events in the form of phrases, from the identified causal utterances. An example of the RECCON task has been shown in Fig. 1.

As depicted in Fig. 1, Speaker A initiates the conversation by felicitating Speaker B's engagement in utterance 1, and triggers the happy emotion in the following utterances. Therefore, utterance 1 can be regarded as one of the causal utterances for emotions in utterances 1 to 4. The goal of CEE is to identify all the causal utterances that trigger the happy emotions in utterances 1 to 4 respectively. More precisely, the happy emotion in utterance 2 is caused by both facts of "engagement" mentioned

Manuscript received 4 July 2023; revised 6 September 2023; accepted 12 September 2023. Date of publication 15 September 2023; date of current version 2 September 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 61972285, in part by the Fundamental Research Funds for the Central Universities, and in part by the Shuguang Program of Shanghai Education Development Foundation and Shanghai Municipal Education Commission under Grant 21SG23. Recommended for acceptance by K. Qian. (*Corresponding author: Ying Shen.*)

Tiantian Chen, Ying Shen, Lin Zhang, and Shengjie Zhao are with the School of Software Engineering, Tongji University, Shanghai 200082, China (e-mail: 2111287@tongji.edu.cn; yingshen@tongji.edu.cn; cslinzhang@tongji.edu.cn; shengjiezhao@tongji.edu.cn).

Xuri Chen is with the School of Humanities, Tongji University, Shanghai 200082, China (e-mail: xurichen@tongji.edu.cn).

The source code is available at <https://github.com/slptongji/MPEG>.

This article has supplementary downloadable material available at <https://doi.org/10.1109/TAFFC.2023.3315752>, provided by the authors.

Digital Object Identifier 10.1109/TAFFC.2023.3315752

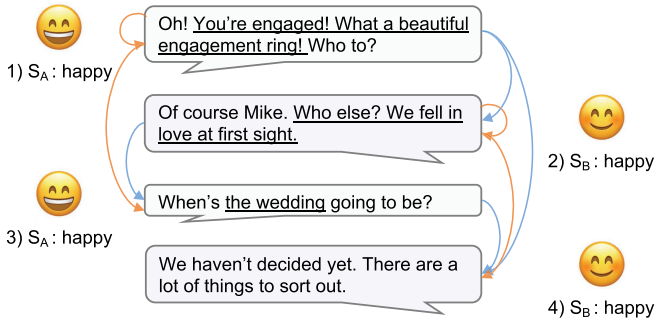


Fig. 1. Conversation example retrieved from the RECCON-DD dataset [22].

in utterance 1 and “love at first sight” mentioned in utterance 2 itself. These facts are causal events for the happy emotion in utterance 2. CSE aims to identify all these causal events for the target utterances.

Identifying emotion causes in conversations is a challenging task due to the complexity of conversation structures and emotional dynamics. Conversation structures should consider not only the content of utterances but also their temporal relations and speaker characteristics. Emotional dynamics need to consider two crucial dependencies: self- and interpersonal-dependencies. These dependencies reveal that an individual’s emotions arise from their own influence as well as from the influence brought by their counterparts during conversations [5], [23]. Existing CEE models focus on modeling the dependencies of emotional dynamics or bringing in external knowledge to enhance model performance [24], [25], [26]. However, these models do not adequately incorporate speaker characteristics and overlook the effects of temporal relations.

To address the research gap to some extent, we propose a new framework to solve the CEE problem, namely “a **Multi-Perspective Enhanced Graph** attention network (MPEG)”. Our contributions are summarized as follows.

- We propose a novel causal emotion entailment framework using a heterogeneous graph attention network architecture. It consists of two different levels, i.e., the utterance-level and the conversation-level encoders and a predictor, which work together to capture the causal relationships between utterances.
- Specifically, at the utterance level, MPEG incorporates speaker characteristics through a speaker-aware pre-trained model. It then leverages attention mechanisms to capture temporal and emotional information embedded in local contexts, enabling an in-depth analysis of utterance contexts.
- At the conversation level, MPEG utilizes a heterogeneous graph attention network for propagating messages, followed by a position-wise feed-forward layer after each graph attention layer, aiming to model the intricate interactions in global contexts.
- To validate the effectiveness of MPEG, extensive experiments are conducted on two publicly available datasets. Results have shown that MPEG yields the best performance, outperforming the SOTA methods for the CEE task and

other related tasks by a large margin (at least +1.02% in macro F1 measure).

The remainder of this paper is organized as follows. Section II introduces the existing work on CEE and two related tasks. Section III describes the proposed architecture of MPEG with details. Section IV outlines the datasets and metrics used in the experiments as well as the implementation details, and demonstrates the experimental results of the evaluated methods. Section V discusses the results of the ablation study and the case study and the impacts of hyperparameters to validate the effectiveness of MPEG. Finally, Section VI concludes the paper.

## II. RELATED WORK

### A. Causal Emotion Entailment

Poria et al. [22] introduced the RECCON task to identify the causes that trigger the speaker’s emotions in the whole conversation. According to the granularity of the identified causes, the RECCON task can be further classified into the CEE task which identifies causes at the utterance level and the CSE task which identifies causes at the phrase level. To address these challenges, Poria et al. [22] formulated CEE as a text classification problem and CSE as a reading comprehension problem, and utilized RoBERTa base/large models [27] to solve both tasks. To facilitate the research on the RECCON problem, the emotion for each utterance and the locations of emotion causes in terms of utterance indices were annotated in the DiallyDialog dataset [28], thereby a new dataset named RECCON-DD was created.

The CSE task can be solved by extracting causal phrases from the identified causal utterances obtained by CEE methods. Therefore, the CEE task receives more attention currently. Zhang et al. [24] proposed a two-stream attention model that can interchange emotion and speaker information to model the speaker’s emotional dynamics during conversations. Li et al. [25] introduced a knowledge enhanced conversation graph (KEC) and proposed a knowledge enhanced directed acyclic graph network to process the graph. Zhao et al. [26] implemented a knowledge-bridged causal interaction network which utilized common sense knowledge (CSK) as three bridges, including the semantics-level bridge, the emotion-level bridge, and the action-level bridge, to capture the inter-utterance dependencies. Bhat and Modi [29] proposed an end-to-end multi-task learning framework for parallel extraction of emotions, causal spans, and causal utterances during conversations, where the emotions of utterances should be predicted beforehand.

Inspired by the RECCON task, Li et al. [30] proposed a new task named Emotion-Cause Pair Extraction in Conversations (ECPEC). They annotated a dataset named ConvECPE based on the IEMOCAP dataset [31] and proposed a two-step framework for the new ECPEC task.

### B. Emotion Recognition in Conversations

Emotion recognition in conversations (ERC) is a highly relevant task to CEE, the problem of which has been defined much earlier. Both tasks involve modeling conversation structures and uncovering emotional dynamics. Despite the relevance of

their solutions, ERC aims to identify the emotion types for the target utterance instead of the emotion causes for the target utterance, which is the main difference between the two tasks. Besides, there is a notable difference between their solutions. ERC methods have to infer unknown emotions for the target utterance. As a comparison, CEE methods directly utilize the emotional information of each utterance which is already known in the CEE task.

ERC methods can be generally categorized into two types, i.e., RNN-based and GNN-based methods, based on their way of conversational context construction. Early studies on the ERC problem utilized recurrent neural networks (RNNs) to model conversation sequence information. Among these methods, CMN [32] was one of the pioneers, which constructed distinct memory networks to store speaker information for speaker modeling. Subsequently, ICON [33] improved upon CMN by interconnecting memory networks to simulate interpersonal- and self-dependencies. DialogueRNN [34] incorporated speaker information and simulated the two dependencies through a hierarchical RNN network with an attention mechanism. COSMIC [35] enhanced DialogueRNN's performance by integrating CSK into the model. While RNN-based approaches can effectively capture the temporal information of conversations, they pay more attention to the closest contextual utterances to the targeted ones. Such tendencies can make it difficult to model long-distance contexts and may compromise their performance.

DialogueGCN [36] is the first algorithm that models conversations as graph structures by using a graph convolutional neural network (GCN) to propagate contextual messages among utterances. To preserve temporal information in graph structures, Ishiwatari et al. [16] proposed a relational-aware graph attention network which adopted a relative position encoding scheme. Lee et al. [37] regarded ERC as a dialogue-based relation extraction (RE) task and proposed a heterogeneous GCN network called TUCORE-GCN. Meanwhile, SKAIG [38] introduced CSK into the GNN model, employing four types of relationships to model the emotional states of speakers. These GNN-based approaches offer a solution to the disability of conveying long-distance contextual information that exists in RNNs. Additionally, they can explicitly model the interpersonal- and self-dependencies of emotions, thereby revealing speaker characteristics and emotional triggers.

### C. Emotion Cause Extraction

Emotion cause extraction (ECE) aims to identify the causes or stimuli which trigger the emotions in each sentence in a long document. In early studies, researchers attempted to extract causal words or clauses for specific emotional expressions through handcrafted rules or features [39], [40], [41], [42], [43]. However, recent studies have employed deep neural networks to solve this problem [44], [45], [46], [47]. For example, Gui et al. [48] viewed the ECE task as a question and answering (Q&A) problem and designed an ECE model based on a Q&A system. Xu et al. [49] approached the ECE problem from an information retrieval perspective and identified emotion causes through learning to rank. Fan et al. [50] utilized a hierarchical

neural network and knowledge-based regularizations to extract emotion causes, aiming to incorporate discourse context information and constrain the parameters.

Xia and Ding [51] further introduced the task of Emotion-Cause Pair Extraction (ECPE) which identifies emotion clauses and their causes in a discourse jointly. They also proposed a two-step framework to perform individual emotion extraction and cause extraction using multi-task learning [51]. However, this framework has an error accumulation problem and struggles with learning the interaction between emotion extraction and cause extraction. To overcome these limitations, Ding et al. [52] developed a novel ECPE-2D framework that utilizes a 2D Transformer to directly model clause pairs. Another recent development in this area is RankCP [53]. This model used graph attention networks to capture the content and structure of documents, which enables it to extract emotion-cause pairs from a ranking perspective.

## III. METHOD

### A. Task Definition

The objective of CEE models is to identify all the causal utterances within the conversational context for a specific target utterance. A conversation  $C$  can be denoted as  $C = \{(u_1, e_1, s_1), \dots, (u_i, e_i, s_i), \dots, (u_n, e_n, s_n)\}$ , where  $n$  is the number of utterances in  $C$ ,  $u_i$ ,  $e_i$  and  $s_i$  are the content, emotion, and speaker of the  $i$ -th utterance respectively. The contextual sequence of  $u_i$  can be represented as  $C_i = \{(u_1, e_1, s_1), \dots, (u_j, e_j, s_j), \dots, (u_i, e_i, s_i)\}$ ,  $j \leq i$ . If  $u_i$  corresponds to a non-neutral emotional utterance, a CEE model has to identify whether each  $u_j$  in  $C_i$  is the emotion cause of  $u_i$ . If yes, the pair  $(u_i, u_j)$  is labeled with 1; otherwise, it is labeled with 0.

### B. Model Overview

The framework of MPEG has been shown in Fig. 2. It consists of three main components: the *utterance encoder*, the *conversation encoder*, and the *cause predictor*. The utterance encoder utilizes a speaker-aware pre-trained model and attention mechanisms to extract the feature representation for each utterance. A heterogeneous conversation graph is then created to model the conversation structure and the emotion dynamics, with each node associated with the feature representation of the corresponding utterance. Based on the heterogeneous conversation graph, the conversation encoder utilizes a combination of graph attention layers and feed-forward layers to propagate conversational messages from both local and global perspectives, resulting in a comprehensive feature representation of each node. This representation is then used to predict the emotional causality. The final component, the cause predictor is implemented as a fully-connected network to predict causal utterances which trigger the emotion in the target utterance.

### C. Utterance Encoder

1) *Input Embedding Layer*: First, a sequence  $I$  is constructed by concatenating the contextual sequence  $C_i$ , target utterance



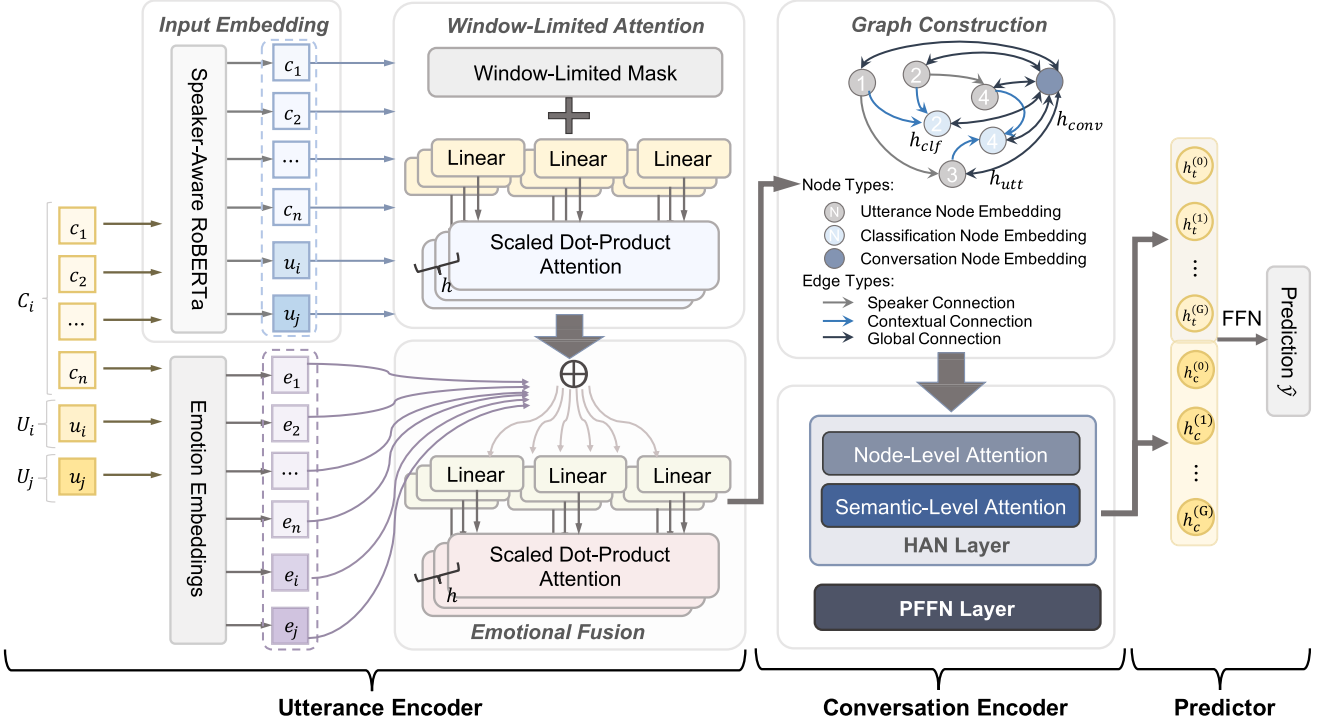


Fig. 2. Overall architecture of MPEG, which consists of three components: the utterance encoder, the conversation encoder and the cause predictor.

$U_i$ , and potential causal utterance  $U_j$ . It can be represented as  $I = \{[CLS], C_i, [SEP], U_i, [SEP], U_j, [SEP]\}$ , where  $U_i = s_i \oplus e_i \oplus u_i$ , and  $\oplus$  denotes the concatenation operation.  $[CLS]$  and  $[SEP]$  are the special tokens in the pre-trained model which mark the beginning of the sequence and the separation of the utterances, respectively. The definition of  $C_i$  has been given in Section III-A, which can also be denoted as  $C_i = \{U_1, U_2, \dots, U_k, \dots, U_i\}$ , where  $U_k$  is the  $k$ -th contextual utterance,  $k \leq i$ . If  $C_i$  exceeds the maximum input sequence length for the pre-trained model, the most distant utterances or tokens will be removed.

Then,  $I$  is fed into Speaker-Aware RoBERTa (SA-RoBERTa) [54] to obtain the corresponding sequence representation  $E$  which integrates contextual information and speaker information. The embedding layers of SA-RoBERTa are shown in Fig. 3. It adds a speaker embedding layer to the traditional embedding layers for the sake of distinguishing different speakers' utterances and modeling speaker transitions during conversations.

2) *Window-Limited Attention Layer*: Local contexts play a crucial role in identifying emotion causes. As shown in Fig. 4, in the RECON-DD dataset, up to 90% emotion causes are located within four utterances previous to the target, and around 60%–70% of the causes appear in the current and the last utterance. To capture the influence of local contexts, we employ a Masked Multi-Head Self-Attention (MHSA) mechanism [55] to the sequence representation  $E$ . The mask  $M$  in MHSA is set to be zero inside an utterance window and negative infinity outside the window, aiming to capture the influence of contextual utterances within the window size  $w$  on the current utterance. It

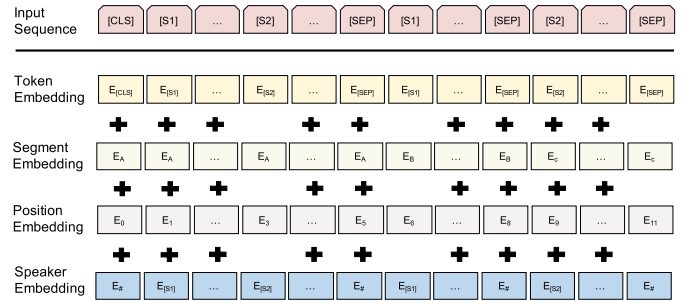


Fig. 3. Embedding layers of SA-RoBERTa [54]. The final input embedding is the sum of the token embedding, the segment embedding, the position embedding, and the speaker embedding. Note that "S1" is a special token which signifies an utterance spoken by Speaker 1.

can be expressed as

$$M[i, j] = \begin{cases} 0, & \text{if } |T_i - T_j| \leq w, \\ -\infty, & \text{otherwise,} \end{cases} \quad (1)$$

where  $i$  and  $j$  are the token positions in the input sequence  $I$ ,  $T_i$  and  $T_j$  are the indices of the utterances where  $i$ -th and  $j$ -th tokens are located.  $w$  is a hyperparameter, indicating that only the local utterances within the window size  $w$  are concerned.

Then the window-limited mask  $M$  is applied to the dot product results between queries and keywords in MHSA, enhancing the capture of local contexts. This process can be computed as

$$\text{Attention}(Q, K, V, M) = \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} + M \right) V,$$

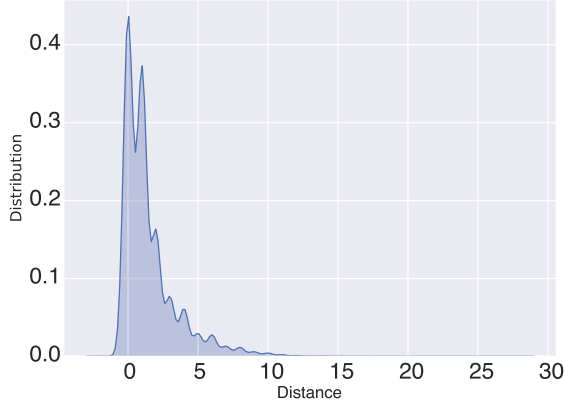


Fig. 4. Distance distribution of causal utterances in the RECCON-DD dataset [22]. “Distance” refers to the distance between the causal and emotional utterance. If the causal and emotional utterances are the same utterance, the distance is zero. “Distribution” reflects the proportion of a specific distance.

$$\text{head}_i = \text{Attention} \left( \mathbf{E}\mathbf{W}_i^Q, \mathbf{E}\mathbf{W}_i^K, \mathbf{E}\mathbf{W}_i^V, \mathbf{M} \right),$$

$$\mathbf{H}^w = \text{MHSA}(\mathbf{E}, \mathbf{M}) = \text{concat}(\text{head}_1, \dots, \text{head}_h), \quad (2)$$

where  $\mathbf{W}_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_q}$ ,  $\mathbf{W}_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$ ,  $\mathbf{W}_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$  are parameter matrices,  $d_q$ ,  $d_k$ ,  $d_v$  denote the dimensions of query vectors ( $Q$ ), key vectors ( $K$ ), and value vectors ( $V$ ),  $d_{\text{model}}$  denotes the hidden layer size of SA-RoBERTa, and  $h$  denotes the number of heads.  $\mathbf{H}^w$  is the output of the window-limited attention layer, denoted as  $\mathbf{H}^w = \{\mathbf{h}_1^w, \mathbf{h}_2^w, \dots, \mathbf{h}_n^w\}$ ,  $n$  is the sequence length of  $\mathbf{E}$ ,  $\mathbf{h}_i^w$  is the feature embedding of utterance  $u_i$ .

3) *Emotional Fusion Layer*: To simulate the emotional dynamics during conversations,  $C_i$ ,  $U_i$  and  $U_j$  are further integrated with their corresponding emotions. Each utterance  $u_i$  ( $u_i \in C_i \cup U_i \cup U_j$ ) has been associated with an emotion label  $e_i \in \mathcal{L}$ , where  $\mathcal{L}$  is an emotion label set, denoted as  $\mathcal{L} = \{l_1, \dots, l_{|\mathcal{L}|}\}$ . Each  $e_i$  can be represented as a trainable embedding  $\text{Emb}(e_i)$ , which is initialized by the weight vectors of SA-RoBERTa. Then the feature embedding of  $u_i$  is concatenated with its corresponding emotion embedding, and the emotional context  $\mathbf{H}^e$  is therefore represented as

$$\mathbf{H}^e = \{\mathbf{h}_{[cls]}^w; [\mathbf{h}_{u_1}^w, \text{Emb}(e_1)]; \dots; [\mathbf{h}_{u_k}^w, \text{Emb}(e_k)]; \mathbf{h}_{[sep]}^w; [\mathbf{h}_{tu}^w, \text{Emb}(e_{tu})]; [\mathbf{h}_{[sep]}^w; [\mathbf{h}_{cu}^w, \text{Emb}(e_{cu})]]\}. \quad (3)$$

Finally,  $\mathbf{H}^e$  is fed into the MHSA network with a linear transformation to learn the mapping relationships, which can be computed as

$$\mathbf{H}^{\text{utt}} = \text{MHSA}(\mathbf{H}^e) \mathbf{W}^e + \mathbf{b} \quad (4)$$

where  $\mathbf{W}^e$  is a parameter matrix and  $\mathbf{b}$  is a bias.  $\mathbf{H}^{\text{utt}}$  is the computed utterance-level feature representation of  $I$ . It consists of the feature embeddings of different utterances, i.e.,

$$\mathbf{H}^{\text{utt}} = \{\mathbf{h}_{[cls]}^w; \mathbf{h}_{u_1}; \mathbf{h}_{u_2}; \dots; \mathbf{h}_{u_k}; \mathbf{h}_{[sep]}^w; \mathbf{h}_{tu}; \mathbf{h}_{[sep]}^w; \mathbf{h}_{cu}\} \quad (5)$$

where  $\mathbf{h}_{u_k}$ ,  $\mathbf{h}_{tu}$  and  $\mathbf{h}_{cu}$  are the feature embeddings of the  $k$ -th contextual utterance, the target utterance  $tu$ , and the causal utterance  $cu$ .

By following the aforementioned process, MPEG is capable of deriving utterance representations that integrate both speaker features and emotional information. It also reinforces the influence of local contexts.

#### D. Conversation Encoder

1) *Graph Construction*: To integrate both global and local contextual information, we construct a heterogeneous conversation graph for each pair of emotion-cause utterances ( $tu$ ,  $cu$ ). The heterogeneous graph can be represented as  $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R})$ , where  $\mathcal{V}$  is the node set consisting of all the graph nodes  $v_i$ ,  $\mathcal{E}$  is the edge set which comprises labeled edges  $(v_i, r, v_j) \in \mathcal{E}$ , and  $r \in \mathcal{R}$  is a relation between two nodes. The graph  $\mathcal{G}$  can be constructed from the conversation history as follows.

*Nodes*: There are three types of nodes in graph  $\mathcal{G}$ : the *conversation node*, the *utterance node*, and the *classification node*. There is only one *conversation node* which is initialized with the feature embedding of token  $[CLS]$ , i.e.,  $\mathbf{h}_{[cls]}$  and captures the global semantics of the conversation. On the other hand, the *utterance node* is initialized with the feature embedding  $\mathbf{h}_i$  and represents utterance  $u_i$  in the conversation. There are two special *classification nodes* in graph  $\mathcal{G}$ , i.e., the *target node* and the *cause node*, representing the target utterance  $tu$  and causal utterance  $cu$ . They are initialized with  $\mathbf{h}_{tu}$  and  $\mathbf{h}_{cu}$ , respectively. These two classification nodes are deliberately proposed to preserve the representations of  $tu$  and  $cu$ , such that later the cause predictor can predict whether  $cu$  is the cause of  $tu$  based on their final representations. In addition, as an *utterance node*,  $cu$  may be truncated if it is too far away from  $tu$ . In such a situation, the introduction of *classification nodes* guarantees that both the target and causal utterances are included in graph  $\mathcal{G}$ .

*Edges and Relations*: There are three types of edges in graph  $\mathcal{G}$ : the *global edge*, the *speaker edge*, and the *context edge*. The *global edge* serves to establish a two-way relationship between the *conversation node* and all the other nodes, facilitating the propagation of global contextual semantics. In contrast, the *speaker edge* connects nodes corresponding to the same speaker by pointing from historical nodes to future nodes. By doing so, it models speakers' emotional dynamics and aggregates information from both nearby and faraway sources. The *context edge* is designed to capture local contextual information about the target utterance  $tu$  and the causal utterance  $cu$ . It establishes a connection between the target/causal utterances and their preceding utterances, aiming to capture the influence of these utterances on  $tu$  and  $cu$ . Initially, the *classification nodes* are connected to their corresponding *utterance nodes* through the *context edge*. As the corresponding *utterance nodes* have preceding *utterance nodes*, the *classification nodes* are then connected to these nodes within a given window size  $w'$ , where  $w'$  is a hyperparameter. Note that the direction of the *context edges* is from *utterance nodes* to *classification nodes*.

2) *Heterogeneous Graph Attention Layer*: The Heterogeneous Graph Attention Network (HAN) [56] is utilized to

aggregate messages from each node. The network comprises two sub-layers: node-level and semantic-level attention layers. In the following, some concepts about HAN will be introduced.

In a heterogeneous graph  $\mathcal{G}$ , two nodes can be connected by several semantic paths (relations) which are referred to as meta-paths [57]. A meta-path  $\Phi$  is defined as  $v_1 \xrightarrow{R_1} v_2 \xrightarrow{R_2} \dots \xrightarrow{R_l} v_{l+1}$  (abbreviated as  $v_1 v_2 \dots v_{l+1}$ ), which reveals a composite relation  $R = R_1 \circ R_2 \circ \dots \circ R_l$  between nodes  $v_1$  and  $v_{l+1}$ , where  $\circ$  denotes the composition operator on relations. Given a meta-path  $\Phi$ , each node  $v_i$  can be connected to a set of nodes via  $\Phi$ , which are termed as meta-path based neighbors of  $v_i$ . To model the speaker's self-dependency and the conversation's local and global information, three types of meta-paths are set up: speaker-speaker, context-context, and global-global. The meta-path set which consists of the three types of meta-paths is denoted as  $\mathcal{P} = \{\Phi_{sp}, \Phi_{ct}, \Phi_{gl}\}$ . These meta-paths propagate information in graph  $\mathcal{G}$  which includes speaker characteristics, local context, and global context.

**Node-Level Attention Layer:** Similar to the graph attention mechanism [58], node-level attention is used to evaluate the significance of the relationship between the target node and its meta-path based neighbors. In the node-level attention layer, information is gathered from the target node's neighbors on each meta-path and the meta-path based feature representations of the target node are obtained. Given a node pair  $(i, j)$  which is connected via a meta-path  $\Phi$ , the importance of the neighbor  $j$  to the target node  $i$  can be computed as

$$e_{ij}^{\Phi} = \text{att}_{utt}(\mathbf{h}_i, \mathbf{h}_j; \Phi) \quad (6)$$

where  $\text{att}_{utt}$  represents the node-level self-attention network [55],  $\mathbf{h}_i, \mathbf{h}_j \in \mathbf{H}^{utt}$  are the feature embeddings of node  $i$  and node  $j$  as described in Section III-D1. It should be noted that all the node pairs  $(i, j)$  on the given meta-path  $\Phi$  share the same attention network  $\text{att}_{utt}$ . Then, the meta-path based attention coefficients  $e_{ij}^{\Phi}$  are normalized by the softmax function to make them more comparable across different nodes:

$$\begin{aligned} \alpha_{ij}^{\Phi} &= \text{softmax}_j(e_{ij}^{\Phi}) \\ &= \frac{\exp(\text{LeakyReLU}(\mathbf{a}_{\Phi}^T \cdot [\mathbf{h}_i \| \mathbf{h}_j]))}{\sum_{k \in N_i^{\Phi}} \exp(\text{LeakyReLU}(\mathbf{a}_{\Phi}^T \cdot [\mathbf{h}_i \| \mathbf{h}_k]))} \end{aligned} \quad (7)$$

where  $\mathbf{a}_{\Phi}$  is the node-level attention vector for meta-path  $\Phi$ ,  $N_i^{\Phi}$  is the meta-path based neighbors of node  $i$  (including itself),  $(\cdot)^T$  denotes the transpose operation, and  $\|$  denotes concatenation. Finally, the meta-path based feature representation of node  $i$  can be computed as

$$\mathbf{z}_i^{\Phi} = \sigma \left( \sum_{j \in N_i^{\Phi}} \alpha_{ij}^{\Phi} \cdot \mathbf{h}_j \right) \quad (8)$$

where  $\sigma$  denotes the activation function. The above mechanism can be extended to a multi-head attention mechanism to enhance the stability of the learning process in self-attention. Specifically, the node-level attention module is iterated  $K$  times, and the resulting output features are concatenated, thereby generating

the final feature representation for each node  $i$ :

$$\mathbf{z}_i^{\Phi} = \big\|_{k=1}^K \sigma \left( \sum_{j \in N_i^{\Phi}} \alpha_{ij}^{\Phi} \cdot \mathbf{h}_j \right) \quad (9)$$

In the end,  $|\mathcal{P}|$  groups of meta-path based node embeddings are obtained, denoted as  $\{\mathbf{Z}_{\Phi_0}, \mathbf{Z}_{\Phi_1}, \dots, \mathbf{Z}_{\Phi_{|\mathcal{P}|}}\}$ , where  $\mathbf{Z}_{\Phi_k} = \{\mathbf{z}_1^{\Phi_k}, \dots, \mathbf{z}_n^{\Phi_k}\}$ ,  $\Phi_k \in \mathcal{P}$ ,  $n$  is the number of nodes.

**Semantic-Level Attention Layer:** Semantic-level attention aims to fuse the semantic information from different meta-paths and computes a more comprehensive feature representation for each node  $i$ . Given  $|\mathcal{P}|$  groups of node embeddings obtained from the node-level attention layer, the weights for each meta-path  $(\beta_{\Phi_0}, \beta_{\Phi_1}, \dots, \beta_{\Phi_{|\mathcal{P}|}})$  can be computed as

$$(\beta_{\Phi_0}, \beta_{\Phi_1}, \dots, \beta_{\Phi_{|\mathcal{P}|}}) = \text{att}_{sem}(\mathbf{Z}_{\Phi_0}, \mathbf{Z}_{\Phi_1}, \dots, \mathbf{Z}_{\Phi_{|\mathcal{P}|}}) \quad (10)$$

$\text{att}_{sem}$  denotes the semantic-level self-attention network and its computation will be illustrated in the following.

To learn the importance of different meta-paths, the meta-path based node embeddings are first transformed by a single-layer MLP and then multiplied by a learnable semantic-level attention vector  $\mathbf{q}$ . Consequently, the importance weights for each meta-path at each node can be learned. These weights are then averaged across all the nodes to obtain the overall importance of each meta-path,  $w_{\Phi_l}$  using (11),

$$w_{\Phi_l} = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \mathbf{q}^T \cdot \tanh(\mathbf{W} \cdot \mathbf{z}_i^{\Phi_l} + \mathbf{b}) \quad (11)$$

where  $\mathbf{W}$  is a parameter matrix,  $\mathbf{b}$  is a bias. Then the importance of each meta-path  $\Phi_l$  is normalized by the softmax function to obtain its weight, which is denoted as  $\beta_{\Phi_l}$ :

$$\beta_{\Phi_l} = \text{softmax}(w_{\Phi_l}) = \frac{\exp(w_{\Phi_l})}{\sum_{\Phi_k \in \mathcal{P}} \exp(w_{\Phi_k})} \quad (12)$$

Finally, the meta-path based node embeddings are weighted by the computed weights to derive the final representation for each node  $i$  as

$$\mathbf{z}_i = \sum_{k=1}^{|\mathcal{P}|} \beta_{\Phi_k} \cdot \mathbf{z}_i^{\Phi_k} \quad (13)$$

**3) Position-Wise Feed-Forward Layer:** Inspired by the work [59], a position-wise feed-forward network (PFFN) [55] is established for the utterance nodes after each round of graph message passing. The PFFN layer provides a non-linear transformation for the hidden states of each location, thereby enhancing the representation of local contexts. Specifically, the representation of node  $i$  is updated using the aforementioned HAN and PFFN layers, which incorporate information from meta-path based neighbors:

$$\mathbf{h}_i^{t+1} = \text{PFFN}(\text{HAN}(\mathbf{h}_i^t)) \quad (14)$$

where  $\mathbf{h}_i^t$  and  $\mathbf{h}_i^{t+1}$  are the node embeddings of node  $i$  at time  $t$  and  $t+1$ , i.e.,  $\mathbf{h}_i^t = \mathbf{z}_i^t$ ,  $\mathbf{h}_i^{t+1} = \mathbf{z}_i^{t+1}$ .

### E. Causal Utterance Predictor

To determine whether causal utterance  $cu$  is the emotion cause of target utterance  $tu$ , the complete representation  $\tilde{H}$  of two classification nodes, i.e., the target and the causal node, should be obtained first. Specifically,  $\tilde{H}$  is computed by concatenating the hidden states of the classification nodes in each HAN layer:

$$\tilde{H} = [h_{tu}^{(0)}; h_{cu}^{(0)}; \dots; h_{tu}^{(G)}; h_{cu}^{(G)}] \quad (15)$$

where  $h_{tu}^{(i)}$  and  $h_{cu}^{(i)}$  are the hidden states of  $tu$  and  $cu$  at  $i$ -th HAN layer, respectively,  $G$  is the number of HAN layers.

Finally,  $\tilde{H}$  is fed into a fully-connected network with sigmoid activation to predict whether  $cu$  is the emotion cause of  $tu$ . And the cross-entropy loss is utilized as the objective function during the training process.

## IV. EVALUATION

### A. Experimental Settings

1) *Datasets*: Two datasets, namely RECCON-DD and ConvECPE, which were released by Poria et al. [22] and Li et al. [30] respectively, were adopted for conducting the experiments. The data samples used for the experiments were constructed by pairing each non-neutral emotional utterance with its historical utterances, including itself, one by one. If a historical utterance was found to be the cause of an emotional utterance, the historical-emotional utterance pair was labeled as positive; otherwise, the pair was labeled as negative.

However, in the RECCON-DD dataset, several future utterances were marked as emotion causes by the authors. Such causes were removed in our experiments because intuitively only historical utterances can trigger the speaker's emotion in the target utterance. In addition, the conversations in the ConvECPE dataset consist of too many utterances, and pairing historical utterances with each target utterance would lead to a class imbalance problem with a positive-to-negative ratio of 1:7.8. To overcome this issue, negative sampling is performed on ConvECPE to ensure that the numbers of negative and positive samples are equivalent. The statistics of two processed datasets are shown in Table I.

2) *Evaluation Metrics*: Referring to Poria et al. [22], macro-averaged F1 (macro F1) score, positive F1 (pos. F1) score, and negative F1 (neg. F1) score were used as the evaluation metrics in this work. Pos. F1 and neg. F1 are two evaluation metrics that measure the accuracy of binary classification models with respect to predicting the positive and negative classes, respectively [60]. Specifically, pos. F1 and neg. F1 are harmonic means of precision and recall for the positive class and negative class, which are computed as

$$\begin{aligned} \text{pos. F1} &= \frac{2 \times P_{pos} \times R_{pos}}{P_{pos} + R_{pos}}, \\ \text{neg. F1} &= \frac{2 \times P_{neg} \times R_{neg}}{P_{neg} + R_{neg}} \end{aligned} \quad (16)$$

where  $P_{pos/neg}$  and  $R_{pos/neg}$  are the precision and the recall values computed for the positive and negative classes. Macro F1

TABLE I

STATISTICS OF THE RECCON-DD AND CONVECPE DATASETS AFTER PREPROCESSING. POS. AND NEG. REFER TO THE NUMBER OF POSITIVE AND NEGATIVE CAUSAL-TARGET UTTERANCE PAIRS, RESPECTIVELY. CONV. AND UTT. ARE SHORT FOR CONVERSATION AND UTTERANCE, RESPECTIVELY

Statistics		RECCON-DD	ConvECPE
Train	Pos.	7269	5279
	Neg.	20646	5279
Valid	Pos.	347	1335
	Neg.	838	1335
Test	Pos.	1894	1824
	Neg.	5330	1824
Num.	Conv.	1106	151
	Utt.	11104	7433
Avg. Len.	Conv.	11 (utt.)	49 (utt.)
	Utt.	60 (words)	61 (words)

is an evaluation metric for the binary classification task which is defined as the average of the positive and negative F1 scores. It is computed as

$$\text{macro F1} = \frac{\text{pos. F1} + \text{neg. F1}}{2} \quad (17)$$

Macro F1 ranges from 0 to 1. A higher macro F1 score indicates a better overall performance in predicting both positive and negative classes.

3) *Baselines*: To demonstrate its effectiveness, MEPG was compared with 13 competitive baseline models in the experiments. The baseline models can be categorized into 4 groups. Following Poria et al. [22], the first group consists of three SOTA methods for the ECPE task, which include ECPE-2D [52], ECPE-MLL [30], and RankCP [53]. Considering the similarity between the CEE and ERC tasks, the second group consists of three SOTA methods for the ERC task, which include TUCORE-BERT/RoBERTa [37] and DAG-ERC [61]. The third group consists of the baseline model for the ECPEC task, i.e., Joint-EC [30]. Finally, the last group consists of six SOTA methods for the CEE task, which include RoBERTa-base/large [22], KEC [25], MuTEC<sub>CEE</sub> [29], KBCIN [26] and PAGE [62].

- *ECPE-2D* An end-to-end framework for the ECPE task that utilizes a 2D representation scheme to encode emotion-cause pairs and simulates their interactions through a 2D transformer module.
- *ECPE-MLL* An improved version of ECPE-2D that incorporates multi-label learning to extract emotion clauses and cause clauses, and combines them to predict the final results.
- *RankCP* A method that treats the ECPE task as a ranking problem and proposes a neural approach that focuses on inter-clause modeling to achieve end-to-end extraction in a single step.
- *TUCORE-BERT/RoBERTa* A method that regards the ERC task as a dialogue-based relation extraction problem and learns contextual representations of utterances through graph convolutional networks [63].
- *DAG-ERC* A SOTA method for the ERC task that treats conversations as directed acyclic graphs and employs a



TABLE II

EXPERIMENTAL RESULTS OF ALL MODELS ON THE RECCON-DD AND CONVECPE DATASETS.  $\triangle$  AND  $\blacktriangle$  DENOTE THE RESULTS ARE REFERRED FROM [22] AND [25], RESPECTIVELY. NOTE THAT THE RESULTS ARE THE MEANS OF FIVE RUNS AND  $(\cdot)$  DENOTES THE STANDARD DEVIATION OF MODELS

Model		RECCON-DD			ConvECPE		
		Pos. F1	Neg. F1	macro F1	Pos. F1	Neg. F1	macro F1
1	ECPE-2D $\triangle$	55.50	94.96	75.23	-	-	-
	ECPE-MLL $\triangle$	48.48	94.68	71.58	-	-	-
	RankCP $\triangle$	33.00	97.30	65.15	-	-	-
2	TUCORE-BERT	67.08	88.53	76.90	-	-	-
	TUCORE-RoBERTa	68.59	90.12	79.35	-	-	-
	DAG-ERC $\blacktriangle$	63.56	95.33	79.44	-	-	-
3	Joint-EC	43.75 <sub>(0.44)</sub>	83.61 <sub>(0.26)</sub>	63.68 <sub>(0.33)</sub>	41.34 <sub>(0.49)</sub>	94.27 <sub>(0.15)</sub>	67.80 <sub>(0.32)</sub>
4	RoBERTa-base	63.31 <sub>(3.02)</sub>	87.99 <sub>(0.47)</sub>	75.65 <sub>(1.52)</sub>	72.48 <sub>(2.60)</sub>	72.49 <sub>(1.23)</sub>	72.49 <sub>(1.45)</sub>
	RoBERTa-large	66.12 <sub>(5.16)</sub>	88.76 <sub>(0.60)</sub>	77.44 <sub>(1.05)</sub>	69.85 <sub>(3.74)</sub>	71.43 <sub>(3.45)</sub>	70.64 <sub>(3.53)</sub>
	KEC	63.85 <sub>(0.80)</sub>	95.63 <sub>(0.08)</sub>	79.74 <sub>(0.44)</sub>	78.00 <sub>(0.93)</sub>	75.10 <sub>(1.32)</sub>	76.55 <sub>(1.00)</sub>
	MuTEC <sub>CEE</sub>	61.62 <sub>(1.14)</sub>	83.46 <sub>(0.46)</sub>	72.54 <sub>(0.50)</sub>	76.46 <sub>(0.57)</sub>	78.24 <sub>(0.33)</sub>	77.35 <sub>(0.15)</sub>
	KBCIN	69.06 <sub>(0.23)</sub>	88.84 <sub>(0.57)</sub>	79.21 <sub>(0.15)</sub>	89.08 <sub>(0.35)</sub>	90.29 <sub>(0.11)</sub>	89.68 <sub>(0.23)</sub>
	PAGE	65.20 <sub>(0.63)</sub>	89.42 <sub>(0.38)</sub>	77.02 <sub>(0.50)</sub>	90.10 <sub>(0.55)</sub>	89.32 <sub>(0.41)</sub>	89.70 <sub>(0.43)</sub>
5	<b>Ours</b>	<b>71.18</b> <sub>(0.50)</sub>	<b>90.35</b> <sub>(0.12)</sub>	<b>80.76</b> <sub>(0.22)</sub>	<b>90.80</b> <sub>(0.38)</sub>	<b>90.72</b> <sub>(0.35)</sub>	<b>90.76</b> <sub>(0.36)</sub>

directed acyclic neural network to learn the intrinsic structures within conversations.

- *Joint-EC* A two-step framework for the ECPEC task, which consists of two multi-task models to extract the emotion-cause pairs in conversations.
- *RoBERTa-base/large* A baseline for the CEE task, utilizing the classification model of RoBERTa-base/large to process concatenated pairs of utterances and their corresponding contexts.
- *KEC* A SOTA method for the CEE task, which builds a knowledge-enhanced dialogue graph and enhances the background knowledge of utterances using a sentiment-realized knowledge selection strategy.
- *MuTEC<sub>CEE</sub>* An end-to-end multi-task learning framework for extracting emotions, emotion cause, and entailment in conversations.
- *KBCIN* A knowledge-bridged causal interaction network that leverages commonsense knowledge as three bridges.
- *PAGE* A position-aware graph-based model that distinguish utterances of different speakers for better causal reasoning.

## B. Implementation Details

The hyperparameter configurations used by MPEG are as follows. In the utterance encoder, we utilized the pre-trained RoBERTa-large uncased model [27] with default hidden layer size of 1024, and the parameters of RoBERTa-large are not frozen in our experiments. We employed two multi-head self-attention modules, each with 16 heads, and a window size of 1 for the window-limited attention layer. In the conversation encoder, we used two layers of HAN, each with a single head, a dropout rate of 0.2, and the same hidden layer size as the pre-trained model. ELU was adopted as the activation function for HAN. The PFFN layer after each HAN layer was implemented by a two-layer CNN network with a dropout rate of 0.1, kernel size of 1, input channel size of 1 for the first layer, and a hidden layer

size consistent with that of the HAN layer for the second layer. In addition, the window size of the context edges was set to 2 in the graph construction process.

AdamW [64] was used as the optimizer with a learning rate of 3.6384e-6, which is derived by the wandb<sup>1</sup> package using a Bayesian search algorithm for automatic parameter tuning. The model was trained for 10 epochs on the RECCON-DD dataset and 5 epochs on the ConvCEPE dataset with a batch size of 12. All the experiments were conducted on an NVIDIA TITAN RTX GPU with 24 GB memory. Reported results are the average scores of 5 runs with fixed random seeds on the test sets obtained from 14 evaluated models.

## C. Experimental Results

Table II demonstrates the experimental results of MPEG as well as other 13 baseline models. The evaluated models are presented in groups based on their categories illustrated in Section IV-A3. The baseline models of ECPE, ERC, ECPEC, and CEE categories are listed in the first four rows, respectively. The proposed MPEG model is listed in the last row to highlight its superiority over all the other competitors.

1) *RECCON-DD Dataset*: From the first row of Table II, it is evidenced that ECPE models perform greatly worse than the models from the ERC and CEE categories. For example, the best-performing ECPE model, i.e., ECPE-2D, only obtains a macro F1 score of 75.23%. As a comparison, the macro F1 scores of the best-performing models in ERC and CEE categories are 79.44% and 79.74%, respectively. These results suggest that the existing models for the ECPE task may not be suitable for solving the CEE task.

The worse performance of ECPE models can be attributed to the methods of pairing clauses in the document one by one. This often leads to a relatively small proportion of positive samples. Consequently, the models' positive F1 scores become very low.

<sup>1</sup><https://wandb.ai/>



Moreover, ECPE models fail to leverage the available emotion labels of utterances and they don't explicitly model conversation structures, such as the speaker information, interpersonal- and self-dependencies of emotions. The results of ECPE models indicate that utilizing known emotional information and modeling the conversation structures are essential for the CEE task.

In the second row of Table II, TUCORE-RoBERTa and DAG-ERC achieve better performances with their macro F1 scores of 79.35% and 79.44%. TUCORE-BERT performs the worst among the three models in the ERC category. Compared with the CEE baseline model, i.e., RoBERTa-large, which is proposed by the authors of RECCON-DD, TUCORE-RoBERTa, and DAG-ERC still achieve improvements in macro F1 of 1.91% and 2.0%, respectively. The better performances of TUCORE-RoBERTa and DAG-ERC can be attributed to the similarity between the tasks of ERC and RECCON. Both tasks require modeling conversation structures and identifying speakers' emotional dynamics.

In the third row of Table II, Joint-EC only achieves a macro F1 score of 63.68%, although it takes into account the conversational characteristics compared to other ECPE models. This may be due to the fact that Joint-EC needs to first detect the emotional and causal utterances in conversations, and then perform a Cartesian product on them before predicting the emotion-cause pairs.

The last two rows of Table II present the experimental results of SOTA models for the CEE task and the proposed MPEG model. As baselines for RECCON-DD, macro F1 scores of RoBERTa-base and RoBERTa-large are 75.65% and 77.44%, respectively. KEC outperforms RoBERTa-base and RoBERTa-large by 4.09% and 2.3%, respectively, making it the second-best model in the CEE category. Its better performance is due to the introduction of sentiment-realized CSK to DAG-ERC. KBCIN delivers a commendable performance, surpassing RoBERTa-base and RoBERTa-large by 3.56% and 1.77% in terms of macro F1, respectively, and falling short of KEC by only 0.53%. Both KBCIN and KEC incorporate CSK and achieve a satisfactory performance, indicating the significance of incorporating external knowledge in causal reasoning tasks. PAGE achieves a macro F1 score of 77.02%, 1.37% higher than RoBERTa-base but 0.42% lower than RoBERTa-large. MuTEC<sub>CEE</sub> performs suboptimally compared to other CEE models, obtaining a macro F1 score of 72.54%, which is 3.11% and 4.9% lower than RoBERTa-base and RoBERTa-large, respectively. This can be attributed to the fact that MuTEC<sub>CEE</sub> treats the emotions of utterances as unknown information and does not make use of them.

Notably, the proposed MPEG model obtains a macro F1 score of 80.76%, surpassing RoBERTa-base and RoBERTa-large by 5.11% and 3.32%, respectively. Despite not incorporating external knowledge like KEC and KBCIN, MPEG outperforms them by 1.02% and 1.55%, respectively. MPEG also outperforms PAGE and MuTEC<sub>CEE</sub> by 3.74% and 8.22% in macro F1, respectively.

2) *ConvECPE Dataset*: Six CEE models and one ECPEC baseline were implemented and tested on the ConvECPE dataset, the results of which have been listed in Table II. Among these

baseline models, PAGE achieves the best performance with a macro F1 score of 89.70%. KBCIN closely follows with a marginally lower score, trailing by only 0.02%. However, KEC, which is the second-best model on RECCON-DD, only achieves a macro F1 score of 76.55% this time. Its performance is 13.15% lower than that of PAGE. The underperformance of KEC on ConvECPE indicates that it may not be well-suited for long conversation scenarios. This could be attributed to its incorporation of excessive external knowledge that may not be suitable for the conversation environment. MuTEC<sub>CEE</sub> achieves a respectable macro F1 score of 77.35% on ConvECPE, without the help of additional emotional information. While this score is 12.33% lower than that of KBCIN, MuTEC<sub>CEE</sub> outperforms RoBERTa-base, RoBERTa-large, and KEC by 4.86%, 6.95%, and 0.8%, respectively. RoBERTa-base/large exhibits the worst performance among the CEE models, possibly due to its simplistic approach of concatenating contextual information. When the conversations become too long, RoBERTa-base/large may truncate some contextual information, resulting in an information loss which leads to a poor performance. Joint-EC achieves a macro F1 score of 67.80%, with a smaller gap compared to RoBERTa-base/large on RECCON-DD, but is still much inferior to the SOTA models in the CEE task. In general, Joint-EC exhibits poor performance on both datasets of the CEE task.

Our proposed MPEG model still achieves the best performance on the ConvECPE dataset with its pos. F1, neg. F1, and macro F1 scores of 90.80%, 90.72%, and 90.76%, respectively. It overwhelms all the CEE models on all three metrics. In particular, MPEG outperforms RoBERTa-base and RoBERTa-large by 18.27% and 20.12% in macro F1, respectively, and even surpasses the second-best PAGE model by 1.06% in macro F1. Furthermore, MPEG outperforms KEC with a notable margin, which ranks second on the RECCON-DD dataset, exhibiting a 14.21% higher macro F1 score. The experimental results indicate that MPEG exhibits outstanding performance not only on RECCON-DD consisting of short conversations but also on ConvECPE consisting of long conversations. Such results demonstrate the outstanding performance and generalization ability of MPEG, making it a highly effective model for handling different conversation scenarios.

## V. ANALYSIS

### A. Ablation Study

To verify the effectiveness of different modules in MPEG, ablation studies were performed on the RECCON-DD dataset. Different modules which include SA-RoBERTa, window-limited attention layer, emotional fusion layer, PFFN layer, emotional information, and speaker characteristics were removed from MPEG respectively and the performances were evaluated thereafter. The experimental results are presented in Table III.

- $\sim$ SA-RoBERTa indicates that the SA-RoBERTa model in the input embedding layer of MPEG is substituted with the regular RoBERTa model.
- $\sim$ Attention indicates that the window-limited attention layer is removed from MPEG.

TABLE III  
RESULTS OF THE ABLATION STUDY ON THE RECCON-DD DATASET

Model	RECCON-DD		
	Pos. F1	Neg. F1	macro F1
~SA-RoBERTa	70.56 <sub>(1.37)</sub>	90.21 <sub>(0.28)</sub>	80.38 <sub>(0.67)</sub>
~Attention	69.54 <sub>(0.43)</sub>	90.63 <sub>(0.07)</sub>	80.09 <sub>(0.09)</sub>
~Fusion	69.31 <sub>(0.87)</sub>	90.14 <sub>(0.23)</sub>	79.73 <sub>(0.20)</sub>
~PFFN	69.13 <sub>(0.64)</sub>	90.50 <sub>(0.18)</sub>	79.82 <sub>(0.38)</sub>
~Emotion	68.07 <sub>(1.10)</sub>	89.91 <sub>(0.18)</sub>	78.99 <sub>(0.49)</sub>
~Speaker	68.97 <sub>(0.43)</sub>	89.90 <sub>(0.27)</sub>	79.43 <sub>(0.34)</sub>
<b>MPEG</b>	<b>71.18<sub>(0.50)</sub></b>	<b>90.35<sub>(0.12)</sub></b>	<b>80.76<sub>(0.22)</sub></b>

- ~Fusion indicates that the emotional fusion layer is removed from MPEG.
- ~PFFN indicates that the position-wise feed-forward layer is removed from MPEG.
- ~Emotion indicates that emotion tokens originally concatenated in the input embedding layer are removed and the emotional fusion layer is eliminated simultaneously.
- ~Speaker indicates that the speaker embedding layer in SA-RoBERTa and the speaker edges in graph construction are eliminated simultaneously.

From Table III, the contributions of different modules in MPEG can be summarized as follows:

- *Efficacy of SA-RoBERTa*: In comparison to MPEG, ~SA-RoBERTa drops by 0.38% on macro F1, indicating that the addition of the speaker embedding layer helps reveal speaker characteristics but to a lesser extent. This may be due to the subsequent modules being more complex, causing the already small contribution of SA-RoBERTa to be further diluted over a prolonged training period.
- *Efficacy of window-limited attention layer*: In comparison to MPEG, ~Attention drops by 0.67% on macro F1, indicating that the window-limited attention layer can incorporate local contextual semantics effectively. However, the impact of attention is not as pronounced as that of the fusion layer, despite both utilizing MHSA for information fusion. We believe that contextual information is more complex and changeable than emotional information, and thus, the effect of attention may not be as significant as that of the fusion layer.
- *Efficacy of emotional fusion layer*: In comparison to MPEG, ~Fusion drops by 1.03% on macro F1, indicating that our emotional fusion layer can effectively integrate emotional information into the utterance representation, thereby enhancing the accuracy of emotion cause prediction.
- *Efficacy of PFFN*: Compared to MPEG, ~PFFN drops by 0.94% on macro F1, indicating that PFFN contributes to information fusion to some extent and can help discover deep semantic information.
- *Efficacy of emotion*: After removing the emotional information, MPEG's model performance drops by 1.77% on

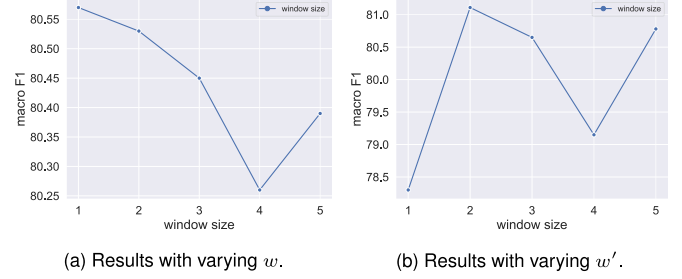


Fig. 5. Experimental results with varying hyperparameters on the RECCON-DD dataset.

macro F1, indicating that the inclusion of emotion labels plays a significant role in the CEE task. However, it is worth noting that MPEG still outperforms the CEE baselines without the use of emotion labels and only slightly underperforms KBCIN and KEC. Moreover, its performance is 6.45% higher than MuTEC<sub>CEE</sub>, which also doesn't utilize emotion labels. These experimental results demonstrate that MPEG can still perform well in situations where the emotions of utterances are unknown, which is in line with real-life conversational scenarios.

- *Efficacy of speaker*: Compared to MPEG, ~Speaker exhibits a decrease of 1.33% in terms of macro F1, indicating that our SA-RoBERTa and speaker edges can effectively model the intra-speaker emotional dynamics and further enhance the model performance.

### B. Impacts of Hyperparameters

Our model utilizes two window sizes: the window-limited size  $w$  in Section III-C2 and the window size  $w'$  in Section III-D1. To gain a better understanding of our model, we conducted the following experiments to investigate the impacts of these two hyperparameters on the model performance: (a) varying sizes of  $w$ , and (b) varying sizes of  $w'$ . The impacts of the two window sizes on the performance of MPEG are shown in Fig. 5.

- *Impact of the value of  $w$* : The window-limited size  $w$  in the MHSA layer determines the range of local contexts with which the current utterance should be fused, with the aim of modeling local utterance-level features. We set the value of  $w$  from 1 to 5 for analyzing the results using different values of  $w$ , as shown in Fig. 5(a). It can be observed that when  $w$  is set to 1, the model achieves the highest performance. As  $w$  increases, the macro F1 value decreases, indicating that incorporating too many local contexts makes the model fail to capture the key information effectively. When  $w$  is set to 5, the model's performance shows a slight improvement compared to  $w = 4$ , but it is still not optimal. Therefore, we set  $w$  as 1 in the final version of our model, as it corresponds to the best performance.
- *Impact of the value of  $w'$* : The window size  $w'$  in the graph construction process determines which preceding nodes should be connected to the current node. We set

TABLE IV  
RESULTS OF CASE STUDY ON THE RECCON-DD TEST SET

**Case 1:** A conversation about mischarging. The speakers' emotions are rather homogeneous. Speaker A is mostly in an angry mood and Speaker B remains in a neutral mood.

Turn	Speaker	Utterance	Emotion	Truth	Predicted
1	A	You guys are charging me \$10 for a movie that I never ordered or saw.	anger	[1]	[1]
2	B	Let's see, sir. According to your file, you watched 'Titanic' Monday evening.	neutral	-	-
3	A	Well, the wrong information is in my file. I was at a concert Monday night.	anger	[1, 3]	[1, 3]
4	B	Well, your word overrules the file, sir. One moment, please.	neutral	-	-
5	A	I knew you'd see it my way.	neutral	-	-
6	B	Sir, I deleted the \$10, but I had to add a \$ 2 service charge to your bill.	neutral	-	-
7	A	Am I in the Twilight Zone? You're charging me for a movie I never saw?	anger	[6, 7]	[1, 6, 7]
8	B	Please don't blame me, sir. Blame the computer programmer.	neutral	-	-
9	A	This is highway robbery. I've got a good mind to call the police!	anger	[6, 7]	[1, 6, 7]
10	B	If it makes you feel any better, other guests feel the same way.	neutral	-	-

**Case 2:** A conversation about affairs. The speakers' emotions are complex and variable, involving surprise, anger, and disgust.

Turn	Speaker	Utterance	Emotion	Truth	Predicted
1	A	The Blake's got divorced.	neutral	-	-
2	B	Really? Why?	neutral	-	-
3	A	Mr. Blake has been getting a little around aside.	neutral	-	-
4	B	I'm surprised. He doesn't look like a guy who'd ever cheat on his wife, does he?	surprise	[3, 4]	[3, 4]
5	A	No, he doesn't. But his wife found out he has been too charming for a long time. Incredibly, he has many different girlfriends. Starting almost right after they married 20 years ago.	neutral	-	-
6	B	Well, I'm really surprised. You are not doing anything behind my back are you?	anger	[6]	[4, 5, 6]
7	A	No, the only thing I've ever done behind your back is zip you up, besides I told all my other girlfriends, and my wife who's getting suspicious. And we had a cold for a while, tell the	neutral	-	-
8	B	Haha... You are not very funny. I guess that means - - except me to tell my lover we have stopped seeing each other too.	disgust	[8]	[7, 8]
9	A	You are not funny either. I can't believe I married a woman like you.	disgust	[9]	[7, 8, 9]

the value of  $w'$  from 1 to 5, as shown in Fig. 5(b). It can be observed that when  $w'$  is set to 1, the model's macro F1 is below 78.5%, indicating insufficient learning of the conversational graph structures. When  $w'$  is set to 2, the model achieves the best performance with a macro F1 value exceeding 81.0%. As  $w'$  is increased from 2, the model's performance decreases until it rises again at the value of 5. The experimental results demonstrate that when  $w'$  is set to 2, it is sufficient to propagate the contextual passages in graphs effectively. Therefore, we choose 2 as the final value of  $w'$ , as it corresponds to the highest macro F1 value.

### C. Case Study

To intuitively demonstrate the performance of MPEG, two representative examples are selected from the RECCON-DD test set for the case study. The conversation contents, the emotion label of each utterance, and the truth and predicted indices of causal utterances are presented in Table IV.

In the first example (i.e., Case 1 in Table IV), a conversation about a mischarge is presented, in which Speaker A speaks with an anger emotion most of the time while Speaker B remains in a neutral mood. In utterance 1, Speaker A emphasizes that he had

been mischarged \$10 for a movie that he did not order. The anger emotion is obviously triggered by the mischarge event illustrated in utterance 1. In utterance 3, Speaker A has to correct another mistake made by Speaker B with mischarge unsolved, which intensifies his anger emotion in the current utterance. Therefore, Speaker A's anger emotion in utterance 3 is caused by both events described in utterances 1 and 3. Similarly, in utterances 7 and 9, Speaker A is annoyed by the extra fee to solve the mischarge problem described in utterance 6. Therefore, the cause of the anger emotion in utterances 7 and 9 should be utterances 6 and 7.

In this example, MPEG successfully identified the correct causal utterance indices for utterances 1 and 3. However, for utterances 7 and 9, MPEG erroneously treated utterance 1 as an emotion cause, although it also identified the correct causal utterances. The extra incorrect prediction can be attributed to the phrase "you're charging me for a movie" in utterance 6, which confuses MPEG and leads to the incorrect assumption that the event in utterance 6 is highly correlated with the mischarge mentioned in utterance 1. Similarly, MPEG mistakenly treated the mischarge mentioned in utterance 1 as a cause of utterance 9.

As demonstrated by Case 1, MPEG can identify the central event running through the conversation and find all the correct causal utterances. However, it cannot always capture the subtle



shifts of the main contradictions and may mistakenly treat the original event as the emotion cause.

In Case 2, a couple's discussion about an extramarital affair is presented. In this example, speakers' emotions are complex and varied, involving surprise, anger, and disgust. In utterance 4, Speaker B is in surprise when learning that Mr. Blake, who looks decent, is cheating on his wife. The surprise emotion is triggered by the cheating event mentioned in utterance 3 and Speaker B's own perceptions revealed in utterance 4. However, in utterance 6, Speaker B abruptly questions Speaker A about whether he has ever cheated on her. This marks a significant contextual shift in the discussion, and only utterance 6 should be regarded as the emotion cause of utterance 6. In utterance 8, Speaker B is disgusted by Speaker A's humor, and this emotion is triggered by the utterance itself. Similarly, in utterance 9, Speaker A feels disgusted that Speaker B is not funny. Therefore, only utterance 9 should be considered as the emotion cause of itself.

In this example, MPEG again identifies all the correct emotion causes but tends to include more unrelated utterances as causes. For utterance 4, MPEG successfully identified the correct causal utterances. But for utterance 6, MPEG did not recognize the contextual shift, and mistakenly attributed Speaker A's anger to the affair mentioned in utterances 4 and 5. For utterance 8, MPEG also misidentified the joke made by Speaker A in utterance 7 as the cause of Speaker B's disgust. However, we think that this inference is reasonable since Speaker B does scold Speaker A due to the ill-timed joke. For utterance 9, MPEG incorrectly identified utterances 7 and 8 as causes. Although we believe that utterance 8 can be seen as the emotional cause of utterance 9, there is no direct relationship between utterances 7 and 9. These results suggest that MPEG has room for improvement in accurately identifying emotion causes.

Overall, the performance of MPEG is satisfactory as it can identify all the correct causal utterances while eliminating most unrelated utterances, as observed in the two examples. However, MPEG tends to consider certain semantic-relevant utterances as emotion causes and, as a result, tends to associate emotion trigger events with more conversation contexts. Especially for the target utterance with only one causal utterance, this approach tends to perform worse. There is still room for improvement, particularly in excluding relevant but non-critical utterances from prediction.

## VI. CONCLUSION

In this paper, we proposed a novel multi-perspective enhanced graph attention network, namely MPEG, which has been demonstrated to be highly effective in the task of causal emotion entailment in conversations. Our model effectively integrates local and global causal associations through utterance-level and conversation-level encoders, and leverages a fully-connected network to learn the relevance between utterances. The incorporation of speaker information with an improved pre-trained model, as well as the employment of two attention mechanisms to aggregate local contexts and emotion dynamics, have further enhanced the model's performance. Moreover, the heterogeneous graph attention network and the position-wise feed-forward network facilitate the fusion of multi-perspective

conversational information, enabling MPEG to achieve state-of-the-art performance in the field. Extensive experiments and studies have corroborated the superior performance of MPEG, which outperforms existing models by a significant margin.

## REFERENCES

- [1] H. A. Simon, *Reason in Human Affairs*. Stanford, CA, USA: Stanford Univ. Press, 1983.
- [2] D. Keltner, D. Sauter, J. Tracy, and A. Cowen, "Emotional expression: Advances in basic emotion theory," *J. Nonverbal Behav.*, vol. 43, pp. 133–160, 2019.
- [3] J. S. Lerner, Y. Li, P. Valdesolo, and K. S. Kassam, "Emotion and decision making," *Annu. Rev. Psychol.*, vol. 66, pp. 799–823, 2015.
- [4] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE Access*, vol. 7, pp. 100943–100953, 2019.
- [5] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea, "Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 108–132, First Quarter 2023.
- [6] D. Zhang, L. Wu, C. Sun, S. Li, Q. Zhu, and G. Zhou, "Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations," in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 5415–5421.
- [7] D. Hu, L. Wei, and X. Huai, "DialogueCRN: Contextual reasoning networks for emotion recognition in conversations," in *Proc. Annu. Meeting Assoc. Comput. Linguistics Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 7042–7052.
- [8] H. Zhang and D. Song, "Towards contrastive context-aware conversational emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 1879–1891, Fourth Quarter 2022.
- [9] S. Xing, S. Mai, and H. Hu, "Adapted dynamic memory network for emotion recognition in conversation," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1426–1439, Third Quarter 2022.
- [10] Y. Ma, K. L. Nguyen, F. Z. Xing, and E. Cambria, "A survey on empathetic dialogue systems," *Inf. Fusion*, vol. 64, pp. 50–70, 2020.
- [11] A. SesagiriRamkumar and Y. Yang, "Empathetic conversational systems: A review of current advances, gaps, and opportunities," 2022, *arXiv:2206.05017*.
- [12] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, "Emotional chatting machine: Emotional conversation generation with internal and external memory," in *Proc. AAAI Conf. Artif. Intell. Innov. Appl. Artif. Intell. Conf. AAAI Symp. Educ. Adv. Artif. Intell.*, 2018, pp. 730–738.
- [13] N. Lubis, S. Sakti, K. Yoshino, and S. Nakamura, "Positive emotion elicitation in chat-based dialogue systems," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 4, pp. 866–877, Apr. 2019.
- [14] N. Majumder et al., "MIME: MIMicking emotions for empathetic response generation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 8968–8979.
- [15] S. Katayama, S. Aoki, T. Yonezawa, T. Okoshi, J. Nakazawa, and N. Kawaguchi, "ER-Chat: A text-to-text open-domain dialogue framework for emotion regulation," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 2229–2237, Fourth Quarter 2022.
- [16] T. Ishiwatari, Y. Yasuda, T. Miyazaki, and J. Goto, "Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 7360–7370.
- [17] W. Zhao, Y. Zhao, and X. Lu, "CauAIN: Causal aware interaction network for emotion recognition in conversations," in *Proc. Int. Joint Conf. Artif. Intell.*, 2022, pp. 4524–4530.
- [18] Y. Li et al., "Towards an online empathetic chatbot with emotion causes," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2021, pp. 2041–2045.
- [19] J. Gao et al., "Improving empathetic response generation by recognizing emotion cause in conversations," in *Proc. Findings Assoc. Comput. Linguistics: EMNLP*, 2021, pp. 807–819.
- [20] J. Wang, W. Li, P. Lin, and F. Mu, "Empathetic response generation through graph-based multi-hop reasoning on emotional causality," *Knowl.-Based Syst.*, vol. 233, 2021, Art. no. 107547.
- [21] H. Kim, B. Kim, and G. Kim, "Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 2227–2240.



- [22] S. Poria et al., "Recognizing emotion cause in conversations," *Cogn. Computation*, vol. 13, no. 5, pp. 1317–1332, 2021.
- [23] M. W. Morris and D. Keltner, "How emotions work: The social functions of emotional expression in negotiations," *Res. Organizational Behav.*, vol. 22, pp. 1–50, 2000.
- [24] D. Zhang, Z. Yang, F. Meng, X. Chen, and J. Zhou, "TSAM: A two-stream attention model for causal emotion entailment," in *Proc. Int. Conf. Comput. Linguistics*, 2022, pp. 6762–6772.
- [25] J. Li et al., "Neutral utterances are also causes: Enhancing conversational causal emotion entailment with social commonsense knowledge," in *Proc. Int. Joint Conf. Artif. Intell.*, 2022, pp. 4209–4215.
- [26] W. Zhao, Y. Zhao, Z. Li, and B. Qin, "Knowledge-bridged causal interaction network for causal emotion entailment," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 14 020–14 028.
- [27] Y. Liu et al., "RoBERTa: A robustly optimized bert pretraining approach," 2019, *arXiv: 1907.11692*.
- [28] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "DailyDialog: A manually labelled multi-turn dialogue dataset," in *Proc. Int. Joint Conf. Natural Lang. Process.*, 2017, pp. 986–995.
- [29] A. Bhat and A. Modi, "Multi-task learning framework for extracting emotion cause span and entailment in conversations," in *Proc. Transfer Learn. Natural Lang. Process. Workshop*, 2023, pp. 33–51.
- [30] W. Li, Y. Li, V. Pandelea, M. Ge, L. Zhu, and E. Cambria, "ECPEC: Emotion-cause pair extraction in conversations," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 1754–1765, Jul.–Sep. 2023.
- [31] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, pp. 335–359, 2008.
- [32] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2018, pp. 2122–2132.
- [33] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, "ICON: Interactive conversational memory network for multimodal emotion detection," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 2594–2604.
- [34] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "DialogueRNN: An attentive RNN for emotion detection in conversations," in *Proc. AAAI Conf. Artif. Intell. Innov. Appl. Artif. Intell. Conf. AAAI Symp. Educ. Adv. Artif. Intell.*, 2019, Art. no. 8.
- [35] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, and S. Poria, "COSMIC: Commonsense knowledge for eMotion identification in conversations," in *Proc. Findings Assoc. Comput. Linguistics: EMNLP*, 2020, pp. 2470–2481.
- [36] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "DialogueGCN: A graph convolutional neural network for emotion recognition in conversation," in *Proc. Conf. Empirical Methods Natural Lang. Process. Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 154–164.
- [37] B. Lee and Y. S. Choi, "Graph based network with contextualized representations of turns in dialogue," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 443–455.
- [38] J. Li, Z. Lin, P. Fu, and W. Wang, "Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge," in *Proc. Findings Assoc. Comput. Linguistics: EMNLP*, 2021, pp. 1204–1214.
- [39] S. Y. M. Lee, Y. Chen, and C. Huang, "A text-driven rule-based system for emotion cause detection," in *Proc. NAACL HLT Workshop Comput. Approaches Anal. Gener. Emotion Text*, 2010, pp. 45–53.
- [40] Y. Chen, S. Y. M. Lee, S. Li, and C. Huang, "Emotion cause detection with linguistic constructions," in *Proc. Int. Conf. Comput. Linguistics*, 2010, pp. 179–187.
- [41] I. Russo, T. Caselli, F. Rubino, E. Boldrini, and P. Mart ínez-Barco, "EMO-Cause: An easy-adaptable approach to extract emotion cause contexts," in *Proc. Workshop Comput. Approaches Subjectivity Sentiment Anal.*, 2011, pp. 153–160.
- [42] L. Gui, D. Wu, R. Xu, Q. Lu, and Y. Zhou, "Event-driven emotion cause extraction with corpus construction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 1639–1649.
- [43] A. Neviarouskaya and M. Aono, "Extracting causes of emotions from text," in *Proc. Int. Joint Conf. Natural Lang. Process.*, 2013, pp. 932–936.
- [44] C. Fan, C. Yuan, J. Du, L. Gui, M. Yang, and R. Xu, "Transition-based directed graph construction for emotion-cause pair extraction," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 3707–3717.
- [45] X. Chen, Q. Li, and J. Wang, "Conditional causal relationships between emotions and causes in texts," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 3111–3121.
- [46] E. Turcan, S. Wang, R. Anubhai, K. Bhattacharjee, Y. Al-Onaizan, and S. Muresan, "Multi-task learning and adapted knowledge models for emotion-cause extraction," in *Proc. Findings Assoc. Comput. Linguistics: ACL-IJCNLP*, 2021, pp. 3975–3989.
- [47] H. Yan, L. Gui, G. Pergola, and Y. He, "Position bias mitigation: A knowledge-aware graph model for emotion cause extraction," in *Proc. Annu. Meeting Assoc. Comput. Linguistics Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 3364–3375.
- [48] L. Gui, J. Hu, Y. He, R. Xu, Q. Lu, and J. Du, "A question answering approach for emotion cause extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1593–1602.
- [49] B. Xu, H. Lin, Y. Lin, Y. Diao, L. Yang, and K. Xu, "Extracting emotion causes using learning to rank methods from an information retrieval perspective," *IEEE Access*, vol. 7, pp. 15 573–15 583, 2019.
- [50] C. Fan et al., "A knowledge regularized hierarchical approach for emotion cause analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process. Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 5614–5624.
- [51] R. Xia and Z. Ding, "Emotion-cause pair extraction: A new task to emotion analysis in texts," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 1003–1012.
- [52] Z. Ding, R. Xia, and J. Yu, "ECPE-2D: Emotion-cause pair extraction based on joint two-dimensional representation, interaction and prediction," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 3161–3170.
- [53] P. Wei, J. Zhao, and W. Mao, "Effective inter-clause modeling for end-to-end emotion-cause pair extraction," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 3171–3181.
- [54] J. Gu et al., "Speaker-aware bert for multi-turn response selection in retrieval-based chatbots," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2020, pp. 2041–2044.
- [55] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.
- [56] X. Wang et al., "Heterogeneous graph attention network," in *Proc. Conf. World Wide Web*, 2019, pp. 2022–2032.
- [57] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, "PathSim: Meta path-based top-K similarity search in heterogeneous information networks," in *Proc. VLDB Endowment*, vol. 4, no. 11, pp. 992–1003, 2020.
- [58] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–12.
- [59] D. Wang, P. Liu, Y. Zheng, X. Qiu, and X. Huang, "Heterogeneous graph neural networks for extractive document summarization," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 6209–6219.
- [60] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 2383–2392.
- [61] W. Shen, S. Wu, Y. Yang, and X. Quan, "Directed acyclic graph network for conversational emotion recognition," in *Proc. Annu. Meeting Assoc. Comput. Linguistics Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 1551–1560.
- [62] X. Gu, R. Lou, L. Sun, and S. Li, "PAGE: A position-aware graph-based model for emotion cause entailment in conversation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023, pp. 1–5.
- [63] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–14.
- [64] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–18.



**Tiantian Chen** received the BS degree in 2021, from the School of Software Engineering, Tongji University, Shanghai, China, where she is currently working toward the PhD degree. Her research interests include dialogue systems, sentiment analysis, and natural language processing.



**Ying Shen** received the BS and MS degrees from Software School, Shanghai Jiao Tong University, Shanghai, China, in 2006 and 2009, respectively, and the PhD degree from the Department of Computer Science, City University of Hong Kong, Hong Kong, in 2012. In 2013, she joined the School of Software Engineering, Tongji University, Shanghai, where she is currently an associate professor. Her research interests include speech and language processing, and sentiment analysis.



**Lin Zhang** (Senior Member, IEEE) received the BSc and MSc degrees from the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2003 and 2006, respectively, and the PhD degree from the Department of Computing, Hong Kong Polytechnic University, Hong Kong, in 2011. From March 2011 to August 2011, he was a research associate with the Department of Computing, Hong Kong Polytechnic University. In August 2011, he joined the School of Software Engineering, Tongji University, Shanghai, China, where he is currently a full professor. His current research interests include environment perception of intelligent vehicle, pattern recognition, computer vision, and perceptual image/video quality assessment. He serves as an associate editor of the *IEEE Robotics and Automation Letters*, and *Journal of Visual Communication and Image Representation*. He was awarded as a Young Scholar of Changjiang Scholars Program, Ministry of Education, China.



**Xuri Chen** received the MS degree from the School of Psychology, Beijing Normal University, Beijing, China, in 2004. She is currently working toward the PhD degree in psychology with the School of Humanities, Tongji University. She joined Tongji University in 2004, where she is engaged in student management and mental health education. She is a national level two psychological counselor, China. She has been trained in the cognitive-behavioral therapy and is now engaged in the research on anxiety and depression counseling strategies for college students.



**Shengjie Zhao** (Senior Member, IEEE) received the BS degree in electrical engineering from the University of Science and Technology of China, Hefei, China, in 1988, the MS degree in electrical and computer engineering from the China Aerospace Institute, Beijing, China, in 1991, and the PhD degree in electrical and computer engineering from Texas A&M University, College Station, TX, USA, in 2004. He is currently the dean with the College of Software Engineering and a professor with the College of Software Engineering and the College of Electronics and Information Engineering, Tongji University, Shanghai, China. In previous postings, he conducted research with Lucent Technologies, Whippany, NJ, USA, and the China Aerospace Science and Industry Corporation, Beijing. He is a fellow of the Thousand Talents Program of China and an Academician of the International Eurasian Academy of Sciences. His research interests include artificial intelligence, Big Data, wireless communications, image processing, and signal processing.