TriKF: Triple-Perspective Knowledge Fusion Network for Empathetic Question Generation

Tiantian Chen, Ying Shen, Xuri Chen, Lin Zhang, *Senior Member, IEEE*, and Shengjie Zhao, *Senior Member, IEEE*

***Abstract*—Questioning is one of the essential tactics for demonstrating empathy in social dialogues. Effective questioning can guide individuals to express their experiences, feelings, and thoughts, aiming to establish emotional connections and deepen interpersonal understanding. However, how to generate empathetic questions in emotional support conversations remains an unresolved issue. To fill this research gap to some extent, we propose an empathetic question generation (QG) framework called triple-perspective knowledge fusion (TriKF), which incorporates external knowledge from the perspectives of events, cognition, and affection to comprehensively understand the dialogue context. Specifically, this framework acquires commonsense knowledge from these three perspectives and integrates them into the dialogue context to enrich the contextual information. To the best of our knowledge, this is the first method proposed for empathetic QG. Additionally, we construct an empathetic question dataset, namely EQ-EMAC. This dataset comprises 4213 dialogues with single user inputs and multiple empathetic question responses, which can be utilized to assess the effectiveness and generalization capability of empathetic QG models. Experimental results have demonstrated the effectiveness of TriKF on the task of empathetic QG compared with seven baseline models.**

***Index Terms*—Cognitive therapy, empathy, question generation (QG), social dialog.**

I. Introduction

EMPATHY is an essential trait of a good listener in social communications, which shows the listener's understanding of the speaker's thoughts and feelings by providing empathetic feedback [1], [2]. Empathy words can enhance the positive impression of the listener, promote the speaker's willingness to express himself/herself, and facilitate the establishment of interpersonal connections between the listener and the speaker. Existing research indicates that dialogue generation models owning empathy characteristics can generate responses that increase users' satisfaction with the dialogue systems [3], [4], [5], [6], [7]. However, how to adequately express empathy in dialogue systems remains a technical challenge.

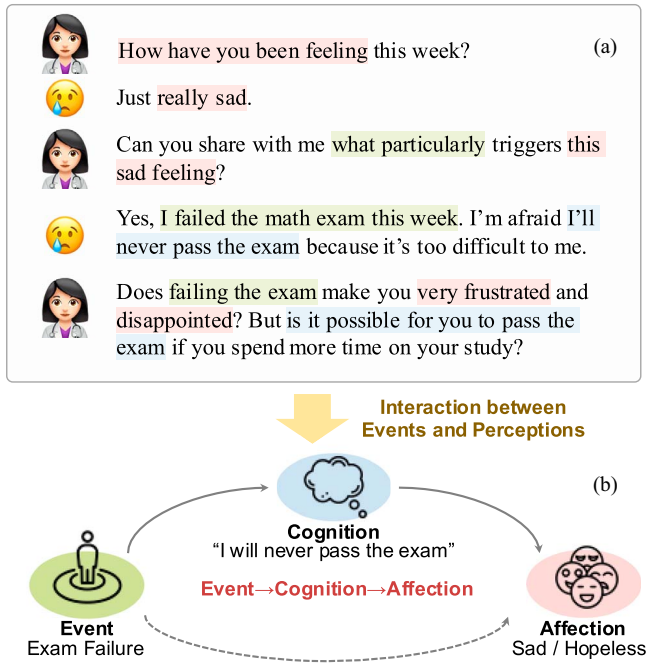

Fig. 1. Empathetic question example in a simulated counseling conversation. (a) Partial conversational content between the speaker and the listener. (b) The interaction between the event, cognition, and affection of the speaker. This example illustrates the interaction between objective events and subjective cognition that influences an individual's emotions. In this example, the texts in green, blue, and pink, respectively, highlight the contents related to the speaker's central events, thoughts, and emotions.

Questioning is one of the tactics to express empathy [8], [9], [10]. In psychology and sociology research, empathetic questions typically guide individuals to express their experiences, feelings, and thoughts, aiming to establish emotional connections and deepen interpersonal understanding [11], [12], [13]. Fig. 1 gives a segment of a simulated counseling conversation. As shown in Fig. 1(a), when the speaker expresses "(I am) Just really sad," an empathetic question, such as "Can you share with me what particularly triggers this sad feeling?" shows the listener's solicitude for the speaker, and reveals the listener's readiness to listen and understand to speaker's feelings, thereby consolidating the link to the speaker. In addition, empathetic questions can guide the speaker to share more information in the conversation and keep the conversation going in the right

Manuscript received 24 April 2024; revised 13 June 2024; accepted 21 June 2024. Date of publication 12 July 2024; date of current version 3 December 2024. This work was supported by the Fundamental Research Funds for the Central Universities. *(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).

Digital Object Identifier 10.1109/TCSS.2024.3418820

direction. For example, in the final utterance in Fig. 1(a), the listener first expresses her sympathy for the speaker's bad luck in his/her exam. Then, the listener encourages the speaker to reflect on his/her thoughts by asking "But is it possible for you to pass the exam if you spend more time on your study?" to make the counseling conversation move on. Viewing the aforementioned benefits, Svikhnushina et al. [14] suggested that dialogue systems can express empathy by introducing empathetic questions. They collected empathetic questions, performed a taxonomy study to investigate their roles in social dialogues, and consequently constructed an empathetic question dataset named EQT. However, their research only focused on the classification of empathetic questions. These collected questions cannot be directly used as empathetic responses by dialogue systems. In real-life situations, empathetic questions should be tailored to the speaker's inputs and problems. Therefore, we introduce the task of generating empathetic questions here for the first time and attempt to address this problem.

Generating empathetic questions is a challenging task. It involves capturing the central events and thoughts that arouse the speaker's emotions and formulating questions around them. As shown in Fig. 1(b), the conversation revolves around the central event "exam failure," which leads the speaker to develop a thought of "I will never pass the exam" and feelings of "sad" and "hopeless." The listener responds with an empathetic question "Does failing the exam make you very frustrated and disappointed?" formulated around the central event and the speaker's emotions. However, due to the limited conversational contents and the absence of background knowledge, dialogue systems often have difficulties extracting sufficient information about the central events, thoughts, and emotions. Related efforts attempt to enrich conversational contexts by incorporating external knowledge, such as commonsense knowledge (CSK) and emotion lexicon knowledge [15], [16], [17], [18]. However, the injection of external knowledge may introduce noise that makes dialogue systems generate incorrect responses. In addition, related research in the empathetic dialogue (ED) generation field primarily generates empathetic responses from affective and cognitive perspectives, i.e., the speaker's affection and cognition expressed in dialogues [19], [20], [21]. Affection refers to the speaker's emotional states, while cognition refers to the speaker's thoughts. However, according to cognitive-behavioral therapy [22], [23], empathetic questions should also be formulated around the central events in the dialogue, which serve as the background or causation for the speaker's emotions and thoughts. The central events interact with the affection and cognition of the speaker, influencing his/her perceptions of situations and subsequently arousing his/her emotions, as illustrated in Fig. 1(b). Therefore, the involvement of central events can help to generate better empathetic questions.

To address the research gaps mentioned above, we propose a novel empathetic question generation (QG) framework, namely "*Tri*ple-Perspective *K*nowledge *F*usion Network for empathetic QG (TriKF)." TriKF enriches conversational contexts comprehensively by incorporating knowledge from three perspectives, i.e., events, cognition, and affection. Additionally, we create the second empathetic question dataset named EQ-EMAC to support further research in this area. Our contributions can be summarized as follows.

1) We propose an empathetic QG model that comprehensively leverages CSK to analyze conversations from the perspectives of events, cognition, and affection. To the best of our knowledge, it is the first empathetic QG framework.
2) We propose a knowledge-context fusion algorithm that effectively integrates and aligns knowledge and contexts, thereby greatly enriching conversational contexts and alleviating the impact of the noise introduced by CSK.
3) We establish an empathetic question dataset named EQ-EMAC, which serves to evaluate the effectiveness of models in this field and facilitate the studies to find generic patterns of effective empathetic questions.
4) Experimental results have shown that TriKF outperforms baseline models across various automatic and human evaluation metrics. These results suggest that TriKF is an effective solution for the task of empathetic QG.

The remainder of this article is organized as follows. Section II introduces the existing work on three tasks related to empathetic QG. Section III describes the proposed architecture of TriKF with details. Section IV introduces the construction of the EQ-EMAC dataset. Section V outlines the datasets, baselines, and evaluation metrics used in the experiments as well as the implementation details. Section VI demonstrates the experimental results of the evaluated methods under both automatic and human evaluation metrics. Section VII provides analyses for the ablation studies and the case study. Finally, Section VIII concludes the article.

II. Related Work

A. ED Generation

Empathy plays an important role in human communication. Recently, there has been a surge of interest in empathy-related topics in dialogue systems [24], [25], [26], [27], [28], [29], [30], [31]. Rashkin et al. [32] were the pioneers in empathizing the significance of empathy in social dialogues. They created the task of ED generation, constructed the first large-scale ED dataset known as EDs, and established several baselines for the task. Lin et al. [19] suggested that expressing empathy requires a focus on speakers' emotions. Therefore, they designed a corresponding decoder for each emotion category and softly integrated the decoding outputs. Zheng et al. [28] proposed that empathetic expression comprises three factors: communication mechanism, dialogue act, and emotion. They employed communication mechanisms to guide empathy expression at a high level while utilizing dialogue acts and emotions for fine-grained realization.

Constrained by the limited length of conversations, dialogue systems encounter challenges in acquiring adequate valid information to comprehend speakers' states and offer appropriate feedback. Therefore, some efforts have been made to bring in external knowledge to enhance the understanding of speakers' utterances [15], [16], [17], [18]. KEMP [15] is the first model

to incorporate external knowledge to generate empathetic responses by utilizing CSK and emotional lexicon knowledge. Sabour et al. [16] obtained affective and cognitive CSK from COMET [33] to enhance the comprehension of speakers, as they pointed out that empathy encompasses both affective and cognitive aspects. Zhou et al. [18] designed a two-level strategy to model the interaction between affection and cognition and generate integrated empathetic responses.

Despite the efforts made in ED generation, the models in this field cannot be directly applied to solve the task of empathetic QG. A typical ED, such as "This matter is so frustrating!," may hinder the speaker's willingness to further express himself/herself and implies the termination of the talk. As a comparison, an empathetic question, such as "This matter is so frustrating. How did you get through it?" can make the speaker feel that he/she is understood and encouraged to continue the talk. In addition, to convey agreement with the speaker, the ED can be easily generated using a simple structure with some emotional words like "That is great." As a comparison, generating empathetic questions requires an overall consideration of central events, affection, and cognition as well as a logical arrangement of the three perspectives. Therefore, it is necessary to develop specific models for the task of empathetic QG.

Many of the aforementioned studies attempt to better express empathy in dialogues by utilizing external knowledge. However, they only focus on affective and/or cognitive knowledge, neglecting the central events that facilitate the analysis of speakers' emotions and thoughts. Therefore, our work aims to generate empathetic questions from the perspectives of events, affection, and cognition, with the goal of understanding speakers' situations in-depth and proposing targeted empathetic questions.

B. QG

QG aims to generate natural and relevant questions from diverse input formats [34], [35], [36]. It has attracted increasing attention across various domains due to its potential applications. For example, QG can serve as a data augmentation technique in the reading comprehension and question–answering (QA) fields by generating abundant testing questions or QA pairs, reducing the manpower of manually making up these questions [37], [38], [39], [40], [41]. In the intelligent education field, QG can serve as an important component of intelligent tutoring systems, facilitating the automatic assessment of students' knowledge levels and self-directed learning [42], [43], [44].

QG, as an essential communication tactic, also has extensive applications in dialogue systems [45], [46], [47], [48], [49], [50]. It is utilized to initiate and sustain conversations to achieve good interactivity with users. Wang et al. [45] proposed the first QG model for single-turn dialogue scenarios. They designed both soft and hard decoders that extract different types of keywords to formulate targeted questions. Pan et al. [46] employed a reinforcement learning mechanism to generate questions based on dialogue histories. Ling et al. [47] generated questions based on conversational contexts, and predicted question types and dialogue topics to promote talk persistence in multiturn dialogues.

The aforementioned work can generate questions based on conversational contexts. However, they mainly cope with daily conversations or conversation comprehension tasks, which only need to capture conversational contexts and central events without considering users' feelings and thoughts. In comparison, empathetic questions should be generated by comprehensively considering dialogue contexts, central events, user emotions, and thoughts. Therefore, it is necessary to explore the expression of empathy in QG within dialogue scenarios to enhance users' satisfaction.

C. Empathetic Question Classification

The first attempt related to empathetic questions is proposed by Svikhnushina et al. [14], which is a classification task on the collected empathetic questions. They developed a taxonomy of empathetic questions collected from social dialogues and constructed an empathetic question dataset named EQT. This dataset provides the actions and intents of questions, where actions represent the communicative strategies of questions such as "request information" (e.g., "What are you studying?"), and intents represent the impact of questions on recipients like "express concern" (e.g., "Why? What happened to her?").

Although their work reveals the significance of empathetic questions in social dialogues, it solely focuses on the classification of empathetic questions. In real-world applications, it is more vital to generate context-related empathetic questions rather than merely predict questions' actions or intents. Therefore, it is necessary to promote the research to generate targeted empathetic questions. In addition, although the EQT dataset has been proposed, no other empathetic question dataset is available. More efforts should be made in the construction of empathetic question datasets.

III. Method

A. Task Definition

Given a dialogue context $C = [(u_1, s_1), \ldots, (u_i, s_i), \ldots, (u_n, s_n)]$, the task of empathetic QG is to generate proper empathetic questions Y based on the dialogue context C. Here, u_i and s_i denote the content and the speaker of the ith utterance, respectively, n is the number of utterances in C. Each dialogue is constructed between a user (denoted as U) and a listener (i.e., the dialogue system, denoted as L), $s_i \in \{U, L\}$. Additionally, the user expresses a specific emotion e during the dialogue, where e is one of 32 emotion types[1].

B. Model Overview

The framework of TriKF, as illustrated in Fig. 2, consists of four main components which include *context feature extraction*,

[1] $e \in$ {surprised, excited, proud, grateful, impressed, confident, hopeful, joyful, prepared, anticipating, content, caring, trusting, faithful, annoyed, angry, sad, lonely, afraid, disgusted, terrified, anxious, disappointed, guilty, furious, nostalgic, jealous, embarrassed, devastated, sentimental, ashamed, apprehensive}

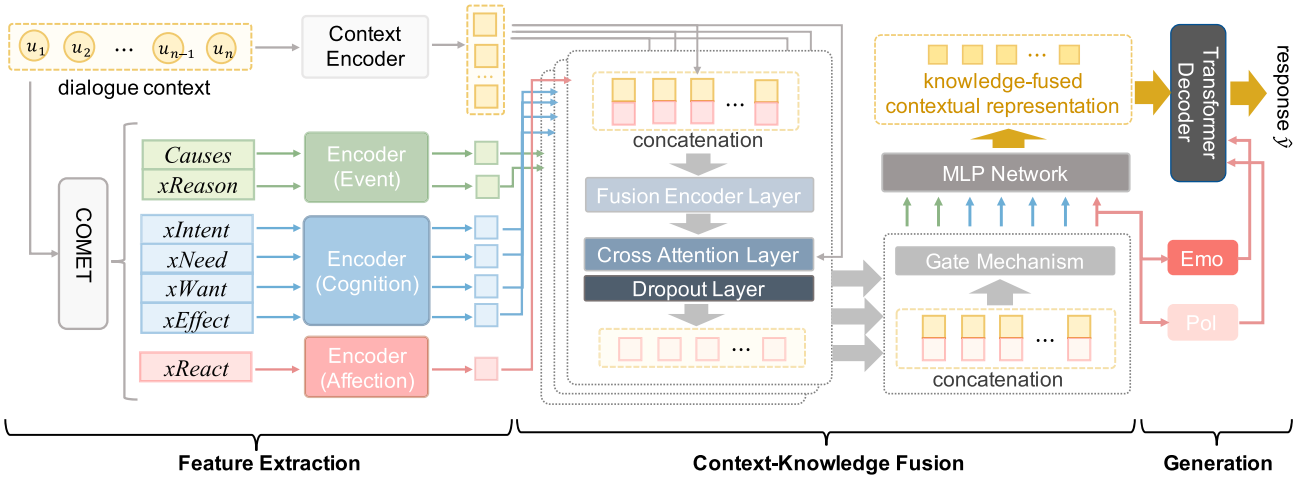

Fig. 2. Overall architecture of TriKF, which consists of four main components: context feature extraction, knowledge feature extraction (referred to as feature extraction stage with context feature extraction), context-knowledge fusion, and response generation.

knowledge feature extraction, *context-knowledge fusion*, and *response generation*. The context feature extraction component captures representations of dialogue contexts, while the knowledge feature extraction component extracts CSK features related to events, cognition, and affection. The context-knowledge fusion component is designed to enrich dialogue contexts with external knowledge. It first integrates various types of knowledge with contexts effectively to obtain knowledge-aware contextual representations. Then, it aligns these representations with the original contexts to ensure consistency between knowledge and contexts and reduce the external noise. Finally, the response generation component generates empathetic questions based on the contextual representations enriched by all knowledge.

C. Context Feature Extraction

To extract context features, the utterances in C are first concatenated and capped with a [CLS] token to construct the context sequence I, i.e., $I = \{[\text{CLS}], u_1, u_2, \ldots, u_n\}$, where [CLS] is a special token marking the start of a sequence. Following prior work [19], I is fed into a modified transformer encoder Enc_{ctx} to learn its feature representation. The "modified" transformer encoder means that there is an additional speaker embedding e_s added to the embedding layers of Enc_{ctx} with the aim of distinguishing utterances from different speakers. Therefore, the input embedding $\boldsymbol{E}$ of sequence I is the sum of three embeddings: token embedding e_t, position embedding e_p, and speaker embedding e_s, as shown in the following equation:

$$\boldsymbol{E} = \text{e}_t(I) + \text{e}_p(I) + \text{e}_s(I). \tag{1}$$

Then, the input embedding $\boldsymbol{E}$ is fed into the context encoder Enc_{ctx} to obtain the contextual feature representation $\boldsymbol{H}_{\text{ctx}}$

$$\boldsymbol{H}_{\text{ctx}} = \text{Enc}_{ctx}(\boldsymbol{E}) \tag{2}$$

where $\boldsymbol{H}_{\text{ctx}} \in \mathbb{R}^{L \times d}$, L is the length of I, and d is the feature dimension.

D. Knowledge Feature Extraction

1) Knowledge Acquisition: To enrich dialogue contexts and understand users' situations and reactions, we comprehensively leverage three perspectives of CSK: event-centered CSK, affective CSK, and cognitive CSK, as introduced in Section I. These three types of knowledge are generated by COMET [33], a generative model trained on ATOMIC-2020 [51], which encompasses social, physical, and event-related aspects of everyday inferential knowledge.

Specifically, we utilize the BART-based [52] variation of COMET. Given the last utterance u_n, COMET generates inferential CSK for u_n under a certain CSK relation type r. This process follows a structured input format of $(u_n, r, [\text{GEN}])$, where [GEN] is a special token marking the beginning of the generated content. There are 23 distinct CSK relation types in ATOMIC-2020. Among these, we selectively utilize seven relations closely associated with events, cognition, and affection, i.e., relation set $R = \{$xReason, Causes, xReact, xIntent, xNeed, xWant, xEffect$\}$, as illustrated in Fig. 3. xReason and Causes elucidate the objective antecedents and consequences of the events talked about in conversations, reviewing conversations from the perspective of events. xReact unveils users' emotions by typically generating words that describe their emotions, such as "happy" or "excited," providing insights on users from an affective perspective. xIntent, xNeed, xWant, and xEffect reveal users' thoughts and interpretations about their situations, providing observations of users from a cognitive perspective. By utilizing these seven relations of knowledge, TriKF can thoroughly comprehend the antecedents and consequences of central events, as well as users' emotions and thoughts.

2) Feature Extraction: Following prior work [16], COMET is employed to generate five CSK statements $[k_1^r, k_2^r, \ldots, k_5^r]$ for each relation r based on the user's final utterance u_n, where $r \in R$. The five generated CSK statements for each relation r are concatenated to construct the relation-aware CSK sequence K_r. Note that CSK sequences under xReact consist of single words, while those under other relations consist of phrases or

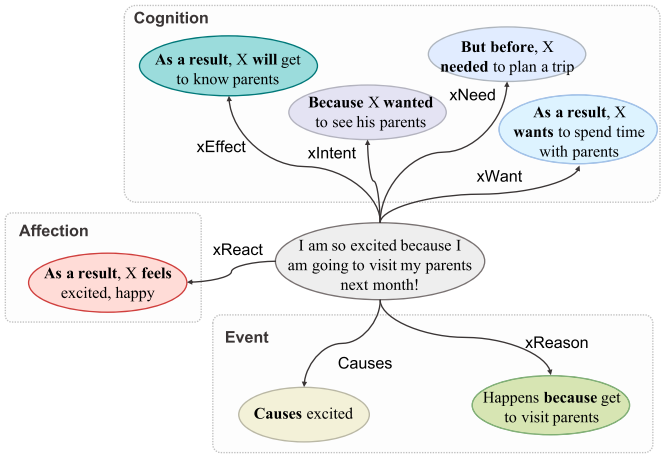

Fig. 3. Examples of the CSK from the perspectives of events, cognition, and affection.

short sentences. Therefore, specifically for CSK sequences generated under relation r', $r' \in R \setminus \{\text{xReact}\}$, they are prefixed with a [CLS] token to learn the phrase or sentence structure, i.e., $K_{r'} = [\text{CLS}] \oplus k_1^{r'} \oplus k_2^{r'} \oplus \ldots \oplus k_5^{r'}$, where $\oplus$ denotes the concatenation operation. Then, the constructed knowledge sequences K_r are fed into a transformer's embedding layer to get the input embeddings of knowledge sequences from three perspectives, i.e., $\boldsymbol{E}_{\text{EVT}}$, $\boldsymbol{E}_{\text{COG}}$, and $\boldsymbol{E}_{\text{AFF}}$. These three types of input embeddings are fed into three independent encoders, each corresponding to a CSK perspective, aiming to uncover the commonalities of knowledge within the same perspectives

$$\boldsymbol{H}_{\text{EVT}} = \text{Enc}_{\text{evt}}(\boldsymbol{E}_{\text{EVT}}) \tag{3}$$

$$\boldsymbol{H}_{\text{COG}} = \text{Enc}_{\text{cog}}(\boldsymbol{E}_{\text{COG}}) \tag{4}$$

$$\boldsymbol{H}_{\text{AFF}} = \text{Enc}_{\text{aff}}(\boldsymbol{E}_{\text{AFF}}) \tag{5}$$

where EVT $= \{\text{Causes}, \text{xReason}\}$, COG $= \{\text{xIntent}, \text{xNeed}, \text{xWant}, \text{xEffect}\}$, AFF $= \{\text{xReact}\}$, $\boldsymbol{H}_{\text{EVT}}$, $\boldsymbol{H}_{\text{COG}}$, and $\boldsymbol{H}_{\text{AFF}}$ are the encoded outputs of knowledge sequences. Note that for the word-formatted xReact knowledge, its feature representation is obtained by averaging the encoded outputs. In comparison, for other knowledge in phrase format, its feature representation is designated as the encoded output of [CLS] to capture the overall sequence structure information. Finally, seven relation-aware knowledge representations $\boldsymbol{H}_r$ are obtained using as follows:

$$\boldsymbol{h}_{\text{xReact}} = \text{AvgPool}(\boldsymbol{H}_{\text{xReact}}) \tag{6}$$

$$\boldsymbol{h}_{r'} = \boldsymbol{H}_{r'}[0] \tag{7}$$

$$\boldsymbol{H}_r = \{\boldsymbol{h}_{\text{xReact}}, \boldsymbol{h}_{r'}\} \tag{8}$$

where $\boldsymbol{h}_{\text{xReact}}, \boldsymbol{h}_{r'} \in \mathbb{R}^d$, $r' \in \{\text{xReason}, \text{Causes}, \text{xIntent}, \text{xNeed}, \text{xWant}, \text{xEffect}\}$, $\boldsymbol{H}_r \in \mathbb{R}^{7\times d}$, $r \in R$.

E. Context-Knowledge Fusion

We enrich dialogue contexts by introducing three perspectives of CSK. However, the introduced knowledge may deviate from contexts and introduce external noise. Therefore, we propose a knowledge fusion algorithm consisting of two stages: the integration stage and the alignment stage. In the first stage, CSK is thoroughly integrated with contexts to obtain knowledge-aware contextual representations. In the second stage, these representations are further aligned with the original contexts, ensuring consistency between knowledge and contexts and reducing external noise. Through these two stages, the contextual representations are correctly enriched by CSK from the perspectives of events, affection, and cognition, which guarantees the accuracy (Acc) and consistency of generated empathetic questions.

1) Integrating Knowledge and Context: The generated relation-aware knowledge is integrated with the dialogue context to obtain relation-aware contextual representations. Specifically, each knowledge representation $\boldsymbol{h}_r \in \boldsymbol{H}_r$ is concatenated with the contextual representation $\boldsymbol{H}_{\text{ctx}}$ at the token level for better fuzing the knowledge features into each token in the contextual sequence. The concatenated representation $\boldsymbol{H}_r^{'}$ can be expressed as

$$\boldsymbol{H}_r^{'} = [\boldsymbol{H}_{\text{ctx}}; \boldsymbol{h}_r] \tag{9}$$

where $\boldsymbol{H}_r^{'} \in \mathbb{R}^{L\times 2d}$. Similarly, three separate fusion encoders $\text{Enc}_{\text{fu,evt}}$, $\text{Enc}_{\text{fu,cog}}$, and $\text{Enc}_{\text{fu,aff}}$ are designed for the three perspectives of knowledge. The concatenated representations $\boldsymbol{H}_r^{'}$ are categorized into three groups, i.e., $\boldsymbol{H}_{\text{EVT}}^{'}$, $\boldsymbol{H}_{\text{COG}}^{'}$, and $\boldsymbol{H}_{\text{AFF}}^{'}$, and fed into the corresponding fusion encoders to derive preliminary relation-aware contextual feature representations $\boldsymbol{H}_{\text{EVT}}^{''}$, $\boldsymbol{H}_{\text{COG}}^{''}$, and $\boldsymbol{H}_{\text{AFF}}^{''}$

$$\boldsymbol{H}_{\text{EVT}}^{''} = \text{Enc}_{\text{fu,evt}}(\boldsymbol{H}_{\text{EVT}}^{'}) \tag{10}$$

$$\boldsymbol{H}_{\text{COG}}^{''} = \text{Enc}_{\text{fu,cog}}(\boldsymbol{H}_{\text{COG}}^{'}) \tag{11}$$

$$\boldsymbol{H}_{\text{AFF}}^{''} = \text{Enc}_{\text{fu,aff}}(\boldsymbol{H}_{\text{AFF}}^{'}) \tag{12}$$

where $\boldsymbol{H}_{\text{EVT}}^{'}, \boldsymbol{H}_{\text{EVT}}^{''} \in \mathbb{R}^{2\times L\times d}$, $\boldsymbol{H}_{\text{COG}}^{'}, \boldsymbol{H}_{\text{COG}}^{''} \in \mathbb{R}^{3\times L\times d}$, and $\boldsymbol{H}_{\text{AFF}}^{'}, \boldsymbol{H}_{\text{AFF}}^{''} \in \mathbb{R}^{L\times d}$.

Subsequently, with $\boldsymbol{H}_{\text{ctx}}$ as the query and $\boldsymbol{H}_r^{''} \in \{\boldsymbol{H}_{\text{EVT}}^{''}, \boldsymbol{H}_{\text{COG}}^{''}, \boldsymbol{H}_{\text{AFF}}^{''}\}$ as the key and the value, $\boldsymbol{H}_{\text{ctx}}$ and $\boldsymbol{H}_r^{''}$ are fed into a cross-attention layer to integrate CSK with the dialogue context deeply

$$\boldsymbol{H}_{fg,r} = \text{Cross-Attention}(\boldsymbol{H}_{\text{ctx}}, \boldsymbol{H}_r^{''}, \boldsymbol{H}_r^{''}). \tag{13}$$

The fine-grained fusion result $\boldsymbol{H}_{fg,r}$ then passes through a dropout layer to prevent overfitting. Consequently, the fused representation of the knowledge and the context $\boldsymbol{H}_{\text{fuse},r}$ is obtained for each type of relation r

$$\boldsymbol{H}_{\text{fuse},r} = \text{Dropout}(\boldsymbol{H}_{fg,r}) \tag{14}$$

where $\boldsymbol{H}_{fg,r}, \boldsymbol{H}_{\text{fuse},r} \in \mathbb{R}^{L\times d}$, $r \in \{\text{xReason}, \text{Causes}, \text{xReact}, \text{xIntent}, \text{xNeed}, \text{xWant}, \text{xEffect}\}$.

2) Aligning Knowledge and Context: Although CSK has been tightly integrated with the dialogue context, not every piece of knowledge contributes equally to the understanding of the context. Therefore, a gating mechanism is employed to align the relation-aware contextual representations $\boldsymbol{H}_{\text{fuse},r}$ with the original dialogue context $\boldsymbol{H}_{\text{ctx}}$, with the goal of reducing

noise. Specifically, the relation-aware contextual representations $\boldsymbol{H}_{\text{fuse},r}$ are first concatenated with the original context $\boldsymbol{H}_{\text{ctx}}$ at the token level. Then, they are fed into a fully connected layer with sigmoid activation to compute the contribution of each relation-aware contextual representation $\boldsymbol{g}_{\text{fuse},r}$

$$\boldsymbol{g}_{\text{fuse},r} = \sigma(\text{FC}_{\text{contr}}([\boldsymbol{H}_{\text{fuse},r}; \boldsymbol{H}_{\text{ctx}}])) \tag{15}$$

where $\boldsymbol{g}_{\text{fuse},r} \in \mathbb{R}^{L\times d}$. Subsequently, $\boldsymbol{H}_{\text{fuse},r}$ are integrated with $\boldsymbol{H}_{\text{ctx}}$ based on the contribution weights $\boldsymbol{g}_{\text{fuse},r}$ to further align knowledge with the original context, resulting in updated relation-aware contextual representations $\boldsymbol{H}_{\text{fuse},r}$

$$\boldsymbol{H}_{\text{fuse},r} = \boldsymbol{g}_{\text{fuse},r} \odot \boldsymbol{H}_{\text{fuse},r} + (1 - \boldsymbol{g}_{\text{fuse},r}) \odot \boldsymbol{H}_{\text{ctx}} \tag{16}$$

where $\boldsymbol{H}_{\text{fuse},r} \in \mathbb{R}^{L\times d}$, $\odot$ denotes elementwise multiplication.

Finally, all the relation-aware contextual representations $\boldsymbol{H}_{\text{aln},r}$ are concatenated at the token level to derive the knowledge-perceived representation $\boldsymbol{H}_{\text{klg}}$. Then the contribution weight of each $\boldsymbol{H}_{\text{aln},r}$ is computed and multiplied by the corresponding representation $\boldsymbol{H}_{\text{aln},r}$, yielding an adjusted knowledge-perceived representation. The adjusted knowledge-perceived representation is then passed through a two-layer MLP network with ReLU activation to obtain the final contextual representation $\widetilde{\boldsymbol{H}}$ which has fully fused with multiple types of knowledge

$$\boldsymbol{H}_{\text{klg}} = [\boldsymbol{H}_{\text{fuse,xReason}}; \ldots; \boldsymbol{H}_{\text{aln,xEffect}}] \tag{17}$$

$$\widetilde{\boldsymbol{H}} = \text{MLP}(\sigma(\boldsymbol{H}_{\text{klg}}) \odot \boldsymbol{H}_{\text{klg}}) \tag{18}$$

where $\boldsymbol{H}_{\text{klg}} \in \mathbb{R}^{7\times L\times d}$, $\widetilde{\boldsymbol{H}} \in \mathbb{R}^{L\times d}$.

F. Response Generation

1) Emotion Classification: Accurately perceiving a user's emotions is the foundation for generating empathetic questions tailored to the user's inputs. To achieve this, we leverage knowledge-aware contextual information from an affective perspective to predict the user's emotional states during conversations. Specifically, we first extract the [CLS] feature of the relation-aware contextual representation under the xReact relation, as discussed in Section III-E2

$$\boldsymbol{h}_{\text{aff}} = \boldsymbol{H}_{\text{aln,xReact}}[0] \tag{19}$$

where $\boldsymbol{h}_{\text{aff}} \in \mathbb{R}^d$. Then, $\boldsymbol{h}_{\text{aff}}$ is passed through a linear layer with softmax activation to derive the emotion category distribution P_{emo}

$$P_{\text{emo}} = \text{softmax}(\boldsymbol{W}_e^\top \boldsymbol{h}_{\text{aff}}) \tag{20}$$

where $\boldsymbol{W}_e \in \mathbb{R}^{d\times q}$ is a parameter matrix, $P_{\text{emo}} \in \mathbb{R}^q$, q denotes the number of emotion categories. The parameters for emotion prediction are optimized by minimizing the cross-entropy loss between the predicted emotion label and the true emotion label e

$$\mathcal{L}_{\text{emo}} = -\log P_{\text{emo}}(e). \tag{21}$$

An excessive focus on individual labels could introduce biases when predicting emotions in overly detailed classifications. Therefore, we refine the initial set of 32 annotated emotions into two primary polarities: positive and negative, as detailed in Table I. Similarly, $\boldsymbol{h}_{\text{aff}}$ is utilized to predict the user's emotional polarity tendency P_{pol}

TABLE I
32 EMOTIONS ARE DIVIDED INTO TWO PRIMARY POLARITIES: POSITIVE AND NEGATIVE

Positive	Negative
Surprised, excited, proud, grateful, impressed, confident, hopeful, joyful, prepared, anticipating, content, caring, trusting, faithful	Annoyed, angry, sad, lonely, afraid, disgusted, terrified, anxious, disappointed, guilty, furious, nostalgic, jealous, embarrassed, devastated, sentimental, ashamed, apprehensive

$$P_{\text{pol}} = \text{softmax}(\boldsymbol{W}_p^\top \boldsymbol{h}_{\text{aff}}) \tag{22}$$

where $\boldsymbol{W}_p \in \mathbb{R}^{d\times 2}$ is a parameter matrix, $P_{\text{pol}} \in \mathbb{R}^2$. The parameters for polarity prediction are also optimized using cross-entropy loss

$$\mathcal{L}_{\text{pol}} = -\log P_{\text{pol}}(p) \tag{23}$$

where p denotes the truth polarity label.

2) Response Decoding: Finally, a transformer decoder is employed to generate empathetic questions. For a given target question response $Y = [y_1, \ldots, y_T]$, the decoder generates the hidden representation of the tth token y_t one by one, which can be computed as follows:

$$P(y_t|y_{<t}, C) = \text{Decoder}(\hat{\boldsymbol{H}}_{y_{<t}}, \widetilde{\boldsymbol{H}}) \tag{24}$$

where $\hat{\boldsymbol{H}}_{y_{<t}}$ denotes the hidden representations of the tokens that have been generated. The standard negative log-likelihood loss (NLL) is utilized as the response generation loss function:

$$\mathcal{L}_{\text{gen}} = -\sum_{t=1}^{T} \log P(y_t|y_{<t}, C). \tag{25}$$

3) Training Objectives: A multitask learning framework is employed to jointly minimize the weighted sum of three losses

$$\mathcal{L} = \gamma_1 \mathcal{L}_{\text{gen}} + \gamma_2 \mathcal{L}_{\text{emo}} + \gamma_3 \mathcal{L}_{\text{pol}} \tag{26}$$

where γ_1, γ_2, and γ_3 are hyperparameters. In our experiments, we set $\gamma_1 = 1$, $\gamma_2 = 1$, and $\gamma_3 = 1$.

IV. EQ-EMAC DATASET

To facilitate the research on empathetic QG, we construct a new empathetic question dataset called EQ-EMAC based on the ED dataset X-EMAC [53]. X-EMAC is a single-turn Chinese ED dataset, where users input queries and the system generates corresponding empathetic responses. This dataset annotates user emotions, response strategies, and keywords related to emotion causes.

X-EMAC annotates the response strategies for each conversation, with "questioning" identified as one of these strategies. "Questioning" is a strategy that indicates listeners express empathy to users through questions. Therefore, we initially screened the original dataset to retain only dialogues annotated with the "questioning" strategy. The retained dialogues

TABLE II
DATA EXAMPLE FROM THE EQ-EMAC DATASET

Query	I am in a bad mood.
Emotion	Sad
Keyword	bad mood
Reply	What can I do to make you happy? Can you tell me why you are unhappy? Tell me what happen? I'm with you, would you like to talk to me? What's wrong? What happened recently? Did something happen to you? Is there anything I can do to help you? Tell me why? What's on your mind?

TABLE III
STATISTICS OF THE EQT AND EQ-EMAC DATASETS

Statistics	EQT			EQ-EMAC
	Train	Valid	Test	
Conv. Num.	14 943	2058	1826	4213
Avg. Conv. Len.	2.57	2.53	2.53	2.00
Avg. Utt. Len.	15.31	16.81	18.21	6.19
Avg. Res. Len.	12.00	12.38	12.89	7.16

Note: Conv. Num. refers to the number of conversations, Avg. Conv. Len., Avg. Utt. Len., and Avg. Res. Len. refer to the average length of conversations, utterances, and target responses, respectively.

comprised a preliminary empathetic question dataset. Then, we utilized the Baidu Translate API[2] to translate the original Chinese sentences in the dataset into English. Subsequently, we performed data cleaning using manually crafted rules to ensure dataset quality. For example, in Chinese dialogues, there are some simple expressions not suitable for English situations, such as "Mo-mo" (which means pat in English). We removed or replaced these expressions to ensure conversational fluency within English contexts. In addition, some keywords in the user inputs were replaced by the keyword slot "*${keyword}*." Therefore, we inserted the keyword tags provided in the original dataset into the keyword slots to obtain the complete user inputs. After that, we randomly selected 250 dialogue samples to assess their topic complexity and grammatical Acc, identifying potential error types. We scrutinized the entire dataset to remove those too simple dialogues and rectify the grammatical errors. In the end, we obtained nearly 18 000 dialogue samples. Considering that some dialogues have the same user input, we merged these dialogues and finally derived 4213 dialogue samples. Each sample preserves a user query, an emotion type, emotion cause keywords, and replies, as shown in Table II. These dialogue samples comprise a new empathetic question dataset, namely EQ-EMAC.

The EQ-EMAC dataset can be utilized to assess the effectiveness and generalization capability of empathetic QG models. It offers a set of universal empathetic questioning examples that can be employed to investigate effective response patterns and strategies. In our experiments, it is used as a test set to verify the generalization capabilities of TriKF and other baseline models.

V. EXPERIMENTAL SETTINGS

A. Datasets

The experiments were conducted on the EQT [14] and EQ-EMAC datasets, respectively. The statistics of the two datasets are shown in Table III.

The EQT dataset consists of around 15 000 dialogues. Each dialogue is initiated by a speaker describing his/her situations, feelings, or thoughts. Then a listener responds with an empathetic question based on the historical context. The dataset provides speakers' emotions at the conversation level, with emotions categorized into 32 classes covering a wide range of positive and negative sentiments. Following Sharma et al. [3], we partitioned the dataset into training, validation, and test sets with a ratio of 8:1:1.

The EQ-EMAC dataset has been introduced in Section IV. Considering its relatively small volume, in our experiments, it is utilized as the test set to further evaluate different models' performances. The evaluated models were first trained and tuned using the training and validation sets of EQT and then evaluated using the EQ-EMAC dataset.

B. Baselines

To demonstrate the effectiveness of our proposed model, TriKF was compared with the following competitive baseline models in our experiments. These baseline models were originally designed for ED generation. However, considering that there are no efforts specially designed for the task of empathetic QG, they were retrained on the EQT and EQ-EMAC datasets to generate empathetic questions. In addition, their word embeddings were all initialized with 300-dimensional pretrained GloVE vectors [54].

1) *Transformer (Trs)* [55]: A dialogue generation model based on the original transformer architecture.
2) *Multitask Transformer (Multi-Trs)* [32]: A transformer-based model that jointly predicts users' emotions and generates EDs.
3) *MoEL* [19]: A method that designs a specialized decoder for each emotion and softly combines the outputs of these decoders to generate EDs.
4) *MIME* [20]: A method that utilizes polarity-based emotion clusters and emotion simulation matrices to improve empathy and contextual relevance of the generated dialogues.
5) *EmpDG* [27]: A multiresolution adversarial framework utilizing multiresolution emotions and users' feedback.
6) *KEMP* [15]: A model that leverages external knowledge to construct an emotional context graph and generates EDs based on emotional cross-attention mechanisms.

[2]https://fanyi-api.baidu.com/

TABLE IV
EXPERIMENTAL RESULTS OF EIGHT MODELS EVALUATED ON THE EQT AND EQ-EMAC DATASETS

Model	EQT					EQ-EMAC				
	PPL ↓	B-2 ↑	B-3 ↑	METEOR ↑	Acc(%) ↑	PPL ↓	B-2 ↑	B-3 ↑	METEOR ↑	Acc(%) ↑
Trs	28.02	4.11	1.42	19.04	-	45.88	3.12	1.34	20.23	-
Multi-Trs	27.77	4.07	1.47	19.55	29.03	47.80	3.38	1.41	22.74	9.73
MoEL	36.97	3.64	1.21	18.23	28.92	63.93	3.68	1.78	22.51	19.80
MIME	28.07	4.59	1.76	19.66	27.71	45.37	3.94	1.15	19.92	**25.35**
EmpDG	29.12	3.80	1.42	18.11	23.44	44.00	4.97	2.69	23.24	20.22
KEMP	25.93	4.56	1.57	20.20	32.15	42.96	1.93	0.71	**24.19**	10.94
CEM	30.37	4.23	1.55	19.77	32.37	46.52	5.92	3.41	23.20	12.32
TriKF	**25.13**	**5.10**	**1.96**	**21.41**	**34.28**	**39.17**	**6.73**	**4.21**	23.09	23.64

Note: The bold values are highlight the best performing model on a specific evaluation metric.

7) *CEM* [16]: A model that uses CSK to acquire cognitive and affective information about users' inputs, and leverages this additional information to enhance the generation of EDs.

These baselines and TriKF are all built on the transformer architecture. Trs refers to the original transformer model without any additional modules. Multi-Trs, MoEL, MIME, EmpDG, and KEMP generate empathetic responses by predicting and utilizing users' emotional information only from an affective perspective. CEM incorporates both affective and cognitive CSK to understand users' feelings and thoughts better. By contrast, our proposed TriKF further introduces objective central events to comprehensively enrich the dialogue contexts from the perspectives of affection, cognition, and events, thereby enhancing the empathy and contextual relevance of the generated questions.

C. Evaluation Metrics

1) Automatic Evaluation: We employ five automatic evaluation metrics to assess the performance of all the baseline models and the proposed TriKF model. Perplexity (PPL) [56] is a metric utilized to evaluate the performance of language models, where a smaller PPL indicates greater Acc and confidence in the model's judgment. BLEU-n [57] is a metric originally designed for the machine translation task but also widely adopted for the dialogue generation task. This metric calculates the similarity between generated responses and reference responses based on the overlap of n-gram phrase sequences. In this study, BLEU-2 (B-2) and BLEU-3 (B-3) are chosen to measure the quality of the generated responses. METEOR [58] builds upon the BLEU metric by incorporating various linguistic factors such as vocabulary matching and synonyms to provide a more comprehensive evaluation. Additionally, we employ Acc to compute the precision of emotion prediction. A higher Acc implies more accurate emotion predictions, indicating the model's proficiency in capturing user sentiments.

2) Human Evaluation: We conducted human evaluations on the generated empathetic questions for the EQT test set and the EQ-EMAC dataset. First, we randomly selected 50 dialogues from the EQT test set and the EQ-EMAC test set, respectively, to construct the sample set for human evaluation. Next, three evaluators were recruited to assess the quality of 50 dialogues generated by each model. The human evaluation was performed in two ways. In the first way, generated empathetic questions from each model were scored according to three human rating criteria (known as human ratings). Specifically, each evaluator was required to assess the generated questions based on three different criteria using a scale of 1–5 (with 5 being the best score): empathy (denoted as *Emp.*, degree of empathetic understanding of the question), relevance (denoted as *Rel.*, relevance of the question to the dialogue's topic), and fluency (denoted as *Flu.*, expression fluency of the question). A model's scores corresponding to the three metrics were computed by averaging all evaluators' ratings over 50 generated questions. In the second way, the overall performance of TriKF was compared with those of other baselines (known as human A/B test). In this test, evaluators were instructed to compare the performance of TriKF with other baselines and select the best response for each sample. Evaluators could choose the "*Tie*" option if they believed that the quality of questions generated by the two models was equal.

D. Implementation

TriKF was implemented using PyTorch. The word embeddings were initialized with 300-D pretrained GloVe vectors. Therefore, the hidden layer sizes of all relevant components were also set to 300. The coefficient hyperparameters of the three training losses in (26) were set to 1.0, i.e., $\gamma_1 = 1.0$, $\gamma_2 = 1.0$, and $\gamma_3 = 1.0$. Adam was employed as the optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. The learning rate was initialized to 0.0001 and adjusted during training according to Vaswani et al. [55]. During the training process, the model was equipped with an early stopping mechanism to prevent overfitting, and its batch size was set to 16 and its minimum number of iterations was set to 9000. All the experiments were conducted on an NVIDIA TITAN RTX GPU with 24GB memory.

VI. EXPERIMENTAL RESULTS

A. Automatic Evaluation

The automatic evaluation results of TriKF and the other seven baseline models are shown in Table IV. The "EQT" column shows the experimental results of eight models evaluated on the EQT dataset, while the "EQ-EMAC" column shows the results

of models on the EQ-EMAC dataset. When evaluated using the EQT dataset, models were trained using the training set of EQT and tested on its test set. Considering the relatively small scale of the EQ-EMAC dataset, models were trained using the EQT dataset and then tested on EQ-EMAC.

1) EQT Dataset: Trs and Multi-Trs are two models based on the transformer architecture. The difference between Trs and Multi-Trs is that the latter employs multitask learning to predict emotions in conversations, thereby incorporating emotional features to express empathy. According to Table IV, Multi-Trs has a decreased PPL score of 27.77 compared with Trs. In addition, its B-3 and METEOR scores are slightly increased, which are 1.47 and 19.55, respectively. These observations suggest that joint emotion prediction based on multitask learning can enhance the quality of generated questions.

MoEL exhibits the worst performance among all the compared models with a maximum PPL score of 36.97, whereas the other baselines' scores are all below 31. Its B-2 and B-3 scores are also the lowest, only 3.64 and 1.21, respectively. In terms of the METEOR metric, MoEL achieves a score of 18.23, which is the second lowest and only surpasses EmpDG's 18.11. MoEL performs poorly across all metrics, possibly due to its utilization of separate decoders for each emotion, resulting in overly complex parameters not being adequately trained.

MIME performs poorly in terms of PPL and Acc, with scores of 28.07% and 27.71%, respectively. However, it achieves the highest scores in terms of B-2 and B-3 among all the baseline models. The good performance of MIME on B-2/3 illustrates that simulating users' emotions is an effective approach for generating empathetic questions. EmpDG exhibits poor performance on several metrics, including PPL, Acc, and METEOR. This may be owing to its complex interactive discriminator, which cannot be sufficiently trained to discriminate the quality of generated empathetic questions.

KEMP and CEM are two state-of-the-art models that introduce external knowledge in the ED generation task. KEMP demonstrates superior performance among all the baseline models, with the best PPL and METEOR scores of 25.93 and 20.20, respectively. As for CEM, it achieves a high Acc score of 32.37%, which indicates a good performance in emotion prediction. However, its PPL score is only 30.37. The poor PPL score of CEM may be due to the inclusion of a diversity loss, which enhances text diversity but decreases the model's confidence in its judgment. Overall, the strong performance of KEMP and CEM across several metrics indicates that external knowledge can enhance the quality of generated responses. In particular, affective knowledge contributes significantly to performance improvement.

As a comparison, the proposed TriKF model surpasses all the baseline models on various metrics, indicating its comprehensive superiority in the task of empathetic QG. Specifically, it has the smallest PPL score of 25.13, which is 0.80 lower than the second-ranked KEMP and 11.84 lower than the worst MoEL. Its emotion prediction Acc achieves the highest score of 34.28%, surpassing the second-ranked CEM by 1.91% and the worst EmpDG by 10.84%. Additionally, TriKF exhibits the highest performance on several metrics associated with the quality of generated questions, e.g., B-2, B-3, and METEOR, the scores of which are 5.10, 1.96, and 21.41, respectively. These scores exceed those of the second-ranked models by 0.51, 0.20, and 1.21, respectively. These results indicate that our proposed method can generate coherent and topic-relevant questions and also accurately recognize users' emotions and provide empathetic support.

TABLE V
RESULTS OF HUMAN RATINGS ON THE EQT DATASET AND THE EQ-EMAC DATASET

Models	EQT			EQ-EMAC		
	Emp.	Rel.	Flu.	Emp.	Rel.	Flu.
Trs	3.67	3.75	4.29	3.59	3.49	3.94
Multi-Trs	3.74	3.74	4.23	3.62	3.51	3.90
MoEL	3.80	3.77	4.31	3.59	3.47	3.89
MIME	3.85	3.79	4.30	3.63	3.54	3.97
EmpDG	3.77	3.76	4.29	3.61	3.50	3.95
KEMP	3.89	3.83	**4.32**	3.64	3.51	3.89
CEM	3.91	3.80	4.26	3.64	3.53	3.96
TriKF	**3.95**	**3.86**	**4.32**	**3.66**	**3.58**	**3.99**

Note: The bold values are highlight the best performing model on a specific evaluation metric.

2) EQ-EMAC Dataset: To evaluate the generalization capability of TriKF, we constructed the EQ-EMAC dataset (introduced in Section IV) and conducted experiments based on it. These experimental results provide substantial evidence to demonstrate the effectiveness of the proposed TriKF.

First of all, TriKF achieves the smallest PPL score of 39.17 among all eight models, which is 3.79 lower than the second-ranked KEMP and 24.76 lower than the worst MoEL. TriKF also outperforms the other models on the B-2 and B-3 metrics, surpassing the second-ranked models by 0.81 and 0.80, respectively. These gaps are even more significant than those observed on the EQT dataset. In terms of Acc, TriKF achieves a score of 23.64%, ranking second among all eight models, just behind MIME, whose score is 25.35%. Its performance in METEOR is also competitive, lower than KEMP's 24.19, and close to the scores of EmpDG and CEM.

Although TriKF is not the best performer on METEOR and Acc metrics, it significantly outperforms the baselines on all the other metrics. Specifically, TriKF achieves the best performance on the metrics of PPL, B-2, and B-3. Especially, its PPL score is much smaller than those of its counterparts. For MIME, which defeats TriKF on the Acc metric, its performance significantly falls behind on the PPL, B-2, B-3, and METEOR metrics. For KEMP, which achieves the highest METEOR score, its PPL and Acc scores are lower than those of TriKF by 3.79% and 12.70%, respectively. Furthermore, it performs the worst in terms of B-2 and B-3 among all models, lower than TriKF by 4.80 and 3.50, respectively. From an overall perspective, TriKF shows the best performance in the experiment.

B. Human Evaluation

1) EQT Dataset: As illustrated in Table V, TriKF surpasses the other baseline models across three human evaluation criteria. In terms of *fluency*, the fluency of responses generated

TABLE VI
RESULTS OF HUMAN A/B TEST ON THE EQT DATASET AND THE EQ-EMAC DATASET

Models	EQT			EQ-EMAC		
	Win	Lose	Tie	Win	Lose	Tie
TriKF versus Trs	44.7%	18.7%	36.7%	38.7%	21.3%	40.0%
TriKF versus Multi-Trs	40.0%	20.0%	40.0%	48.7%	26.0%	25.3%
TriKF versus MoEL	40.7%	28.7%	30.7%	44.0%	11.3%	44.7%
TriKF versus MIME	38.7%	24.0%	37.3%	21.3%	16.0%	62.7%
TriKF versus EmpDG	40.7%	28.0%	31.3%	38.7%	16.0%	45.3%
TriKF versus KEMP	33.3%	29.3%	37.3%	33.3%	17.7%	50.0%
TriKF versus CEM	34.0%	27.3%	38.7%	28.7%	21.3%	50.0%

by each model is quite satisfactory, with all scores exceeding 4. TriKF and KEMP receive the highest score of 4.32. CEM, despite being a SOTA model, exhibits lower fluency and generates repetitive sentences sometimes. Multi-Trs exhibits the lowest fluency among all models, with a score as low as 4.23, possibly due to its simplistic introduction of emotion prediction which affects the decoding process of transformer. In terms of *relevance*, TriKF achieves the highest score of 3.86, with most of the generated questions able to discuss the central events in conversations. CEM and KEMP also perform well, scoring 3.80 and 3.83, respectively. Multi-Trs and Trs exhibit the poorest performance, often generating contextually irrelevant responses. In terms of *empathy*, TriKF achieves the highest score of 3.95, indicating its ability to effectively capture users' emotions and generate empathetic questions. CEM and KEMP also demonstrate strong performance, scoring 3.91 and 3.89, respectively, indicating that the incorporation of emotional knowledge can indeed enhance the empathetic capabilities of models. Trs displays the weakest empathetic ability among all models, with a score as low as 3.67, possibly due to the absence of prediction or utilization of users' emotions.

Table VI shows the comparative evaluation results between TriKF and the seven baseline models. As shown in Table VI, TriKF outperforms all baseline models in the overall evaluation. Specifically, TriKF surpasses KEMP with 33.3% of its responses being superior and 29.3% inferior, yielding an overall superiority of 4.0%. Compared with CEM, 34.0% of TriKF's responses are superior and 27.3% inferior, resulting in an overall superiority of 6.7%. Moreover, when compared to Trs, TriKF exhibits a considerable advantage, with 44.7% of its responses superior and only 18.7% inferior. These results greatly demonstrate the TriKF's notable superiority over all baseline models.

2) EQ-EMAC Dataset: As shown in Table V, TriKF exhibits good generalization capability on the EQ-EMAC dataset. It achieves the highest scores in terms of empathy, relevance, and fluency, with values of 3.66, 3.58, and 3.99, respectively. According to Table IV, KEMP and MIME are the only two baselines that surpass TriKF and achieve the highest scores in METEOR and Acc metrics of the EQ-EMAC test set, respectively. However, KEMP exhibits poor performance in human evaluation metrics, lower than TriKF by 0.02, 0.07, and 0.10 in the empathy, relevance, and fluency metrics, respectively. Its relevance and fluency scores are only 3.51 and 3.89, the second-lowest and third-lowest among all models. As to MIME, it achieves a competitive performance in the human evaluation metrics of EQ-EMAC. It scores 3.63, 3.54, and 3.97 in terms of empathy, relevance, and fluency, respectively. However, it still performs worse than TriKF on the three metrics, whose scores are lower than those of TriKF by 0.03, 0.04, and 0.02, respectively.

In addition, the human A/B test results of TriKF compared to other baseline models are illustrated in Table VI. The comparative results indicate that the question quality of TriKF significantly surpasses that of the other models. Specifically, compared to KEMP, TriKF produces 33.0% better responses and only 17.0% inferior responses. Compared to MIME, TriKF produces 21.3% better responses and only 16.0% inferior responses. These results indicate TriKF's outstanding performance in the human evaluation on the EQ-EMAC test set.

TriKF outperforms all the baselines in both automatic and manual evaluations for the EQT and EQ-EMAC datasets, demonstrating that TriKF has excellent performance and good generalization capability.

VII. EXPERIMENTAL ANALYSIS

A. Ablation Study for Knowledge Analysis

To validate the effectiveness of various types of knowledge, an ablation study was conducted on the EQT dataset. The models involved in this study are divided into three groups: the first group comprises models with one type of knowledge removed, the second group comprises models with two types of knowledge removed, and the third group is TriKF itself which retains all three perspectives of knowledge. Specifically, "~Aff.," "~Cog.," and "~Evt." represent the models which remove the knowledge of affection, cognition, and events in the training process, respectively. "+Aff.," "+Cog.," and "+Evt." represent the models which only keep the knowledge of affection, cognition, and events in the training process, respectively. The experimental results are shown in Table VII.

According to the first group of Table VII, after removing affective knowledge, the scores in all metrics decrease to some extent. Especially the Acc score, decreases greatly from 34.28% to 29.30%, which indicates that the introduction of affective knowledge can significantly improve the Acc of the model in

TABLE VII
RESULTS OF THE ABLATION STUDY FOR KNOWLEDGE ANALYSIS ON THE EQT DATASET

Model	EQT Dataset				
	PPL ↓	METEOR ↑	B-2 ↑	B-3 ↑	Acc(%) ↑
˜Aff.	25.83	19.24	4.34	1.61	29.30
˜Cog.	26.16	**21.45**	**5.13**	1.69	34.01
˜Evt.	25.97	18.55	4.20	1.66	33.02
+Aff.	25.83	19.06	4.38	1.69	32.48
+Cog.	25.91	19.40	4.29	1.34	31.33
+Evt.	25.93	18.20	4.18	1.49	30.45
TriKF	**25.13**	21.41	5.10	**1.96**	**34.28**

Note: The bold values are highlight the best performing model on a specific evaluation metric.

TABLE VIII
RESULTS OF THE ABLATION STUDY FOR MODULE ANALYSIS ON THE EQT DATASET

Model	EQT Dataset				
	PPL ↓	METEOR ↑	B-2 ↑	B-3 ↑	Acc(%) ↑
˜Enc.	26.05	21.00	5.03	1.84	32.53
˜Cross	26.30	19.70	4.27	1.42	**35.65**
˜Integ.	27.02	18.96	4.45	1.93	33.41
˜Gate	25.53	**21.89**	**5.26**	2.05	**35.65**
˜MLP	25.78	20.61	4.89	**2.15**	34.83
˜Aln.	25.35	19.81	4.57	1.65	33.19
˜Pol.	25.24	21.17	4.98	1.93	32.53
TriKF	**25.13**	21.41	5.10	1.96	34.28

Note: The bold values are highlight the best performing model on a specific evaluation metric.

capturing users' emotions. Additionally, the declines in other metrics such as PPL and METEOR also indicate that affective knowledge can enhance the quality of generated questions. When cognitive knowledge is removed, although there is a slight increase in METEOR and B-2 scores, the scores in all the other metrics decrease. Especially after removing cognitive knowledge, the PPL value rises by 1.03, indicating that cognitive knowledge can effectively improve the model's Acc and confidence in its judgment. When event knowledge is removed, there is a decrease in scores across all metrics. Especially for METEOR and B-2, the scores of which decrease by 2.86 and 0.90, respectively. These results suggest that event knowledge contributes to an overall improvement in the quality of generated questions.

The second group of data in Table VII emphasizes the importance of various types of knowledge from another angle. +Aff., the model that only retains affective knowledge, achieves the best overall performance in this group, indicating the significance of affective knowledge in empathetic QG. +Cog. is inferior to +Aff. in several metrics, except for METEOR, where +Cog. achieves a score of 19.40, surpassing +Aff.'s 19.06 and +Evt.'s 18.20. This result indicates that cognitive knowledge mainly contributes to the quality of generated questions. +Evt. performs worse than +Aff. and +Cog. across various metrics, suggesting that event-centered knowledge is less effective when used alone. However, when it is used with other knowledge, as illustrated by the comparative results between ˜Evt. and TriKF, event-centered knowledge can comprehensively improve the scores in all metrics.

Based on the above analysis, each type of knowledge makes a unique contribution to improving the model's performance. The model that combines knowledge from all three perspectives comprehensively demonstrates superior performance compared to those leveraging only one or two types of knowledge across all metrics.

B. Ablation Study for Module Analysis

To validate the effectiveness of each module in TriKF, another ablation study was performed on the EQT dataset, and the experimental results are shown in Table VIII. The experiments in this study are categorized into four groups: the first group comprises ablation experiments on the integration module (described in Section III-E1), the second group comprises ablation experiments on the alignment module (described in Section III-E2), the third group comprises the ablation experiment on the polarity prediction, and the fourth group comprises the experiment on the original TriKF model. Specifically, the design of each ablation experiment is as follows.

1) ˜Enc. denotes the removal of fusion encoders from the integration module of TriKF, corresponding to (9) and (12).
2) ˜Cross denotes the removal of cross-attention mechanisms from the integration module of TriKF, corresponding to (13) and (14).
3) ˜Integ. denotes the complete removal of the integration module from TriKF.
4) ˜Gate denotes the removal of alignment operations between each knowledge and context from the alignment module of TriKF, corresponding to (15) and (16).
5) ˜MLP denotes the substitution of the MLP-based fusion operations with a simpler linear layer in the alignment module of TriKF, corresponding to (17) and (18).
6) ˜Aln. denotes the complete removal of the alignment module from TriKF. Specifically, this involves removing the alignment operations between knowledge and contexts and replacing the MLP-based fusion operations with a simple linear layer.
7) ˜Pol. denotes the removal of the prediction for users' emotional polarity from TriKF.

According to the first group of data in Table VIII, ˜Enc. exhibits a detrimental effect on the scores across all five metrics, particularly in Acc, where it leads to a reduction of 1.75%. ˜Cross results in reduced performance in various metrics, especially in PPL and METEOR, the values of which increase by 1.17 and decrease by 1.71, respectively. These results suggest that the independent utilization of fusion encoders or cross-attention mechanisms can greatly contribute to the overall performance of the model. ˜Integ. leads to a notable decrease in performance across all metrics, particularly in PPL and METEOR, which experience a substantial increase of 1.89 and a significant decrease of 2.45, respectively. These results demonstrate that the integration module can significantly integrate

knowledge and conversational contexts and enhance the quality of generated questions.

According to the second group of data in Table VIII, it is observed that ˜gate has a positive impact on the METEOR, B-2/3, and Acc metrics. However, the PPL value increases by 0.4 after its removal, which is only 0.4 lower than the second-ranked KEMP. This change reduces the advantage of the model over baselines in terms of PPL. In addition, TriKF significantly outperforms all baselines across all metrics. Although TriKF is slightly inferior to ˜Gate, it still meets our expectations. Therefore, to balance the results across multiple metrics, we finally chose the current scheme to maintain a low PPL value. ˜MLP leads to a decrease in the PPL, METEOR, and B-2 scores, but an increase in the B-3 and Acc scores. It indicates that the two-layer MLP network mainly improves the quality of generated questions but may hinder emotion prediction. ˜Aln. induces a certain degree of decline in performance across all evaluation metrics. Especially in METEOR and Acc, the scores of which decrease by 1.6% and 1.09%, respectively. These results suggest that the alignment module effectively aligns knowledge with dialogue contexts, acquiring a comprehensive contextual representation that combines multiple types of knowledge. Furthermore, they also indicate that the combined utilization of the gate mechanism and the MLP network can have a more comprehensive and positive effect compared to their independent utilization, which may result in underperformance in certain metrics, as previously mentioned.

After removing Pol., the model's performance exhibits a slight decline in metrics such as PPL and METEOR, indicating that polarity prediction contributes weakly to the quality of generated questions. However, the Acc score decreases by 1.75% after its removal, suggesting that the introduction of emotional polarity indeed improves the Acc of emotion prediction as intended.

C. Case Study

To intuitively demonstrate TriKF's capability in empathetic QG, two representative dialogues are selected from the EQT test set for case studies, as shown in Table IX. In the first dialogue, many baselines as well as TriKF capture the speaker's joyful emotion and express empathy through phrases like "that is great." However, all the baselines misunderstand the central event of "I saw the picture of my unborn baby" and result in highly inappropriate responses, such as KEMP's "What kind of car?." Despite TriKF's response differing from the standard one, it indeed understands the speaker's joy in the central event. By suggesting "I bet you were looking forward to it?," TriKF identifies the event as something worth anticipating and encourages the speaker to share more about his feelings and stories. In the second example, the speaker expresses his confidence in his interview performance. TriKF identifies the central event as the "job interview." It first generates an empathetic sentence "That is great!," then proposes a probing question "What job is it?." This response closely aligns with the standard response, indicating that TriKF can accurately generate empathetic questions. However, most baseline models except for CEM, still demonstrate issues such as misunderstanding the central event. These two examples demonstrate the superiority of the proposed TriKF in generating suitable empathetic questions.

TABLE IX
RESULTS OF THE CASE STUDY ON THE EQT TEST SET

Context	I came home, and my wife showed me a picture in 4-D of our future baby, my heart speeded so fast!
Emotion	Joyful
Trs	Wow, that is a good thing. I am so happy for you! how did you do?
Multi-Trs	Wow, that is so cool! how did you do?
MoEL	That is a great surprise! Did you enjoy it?
MIME	Wow, that is awesome. what was it?
EmpDG	Oh wow, that is so exciting! How old is she?
KEMP	That is so sweet. What kind of car?
CEM	Wow! That is so exciting! how did it go?
TriKF	I bet you were looking forward to it?
Gold	Oh wow. Do you know whether it will be a boy or girl yet?
Context	I know I shouldn't count my chicks before they hatch, but I am feeling pretty good after that interview today!
Emotion	Confident
Trs	That is so exciting! What kind of things do you plan to do with your birthday?
Multi-Trs	I am sure you will not be! What subject is it?
MoEL	That is great! I am happy for you. where do you think the went?
MIME	That is good. What is your favorite?
EmpDG	What is your job?
KEMP	What kind of restaurant is it?
CEM	That is great. What is the job?
TriKF	That is great! What job is it?
Gold	Oh what job was the interview for?

VIII. CONCLUSION

In this article, we introduced the task of empathetic QG and proposed a novel framework named TriKF. TriKF first derives CSK from the perspectives of events, cognition, and affection, and incorporates these three types of knowledge into dialogue contexts to obtain knowledge-aware contextual representations. It then aligns these representations with the original contexts to ensure consistency between knowledge and contexts and reduce external noise. Subsequently, TriKF can effectively predict users' emotions and generate tailored empathetic questions with these fused contextual representations. To promote further research in this area, we construct an empathetic question dataset named EQ-EMAC, which consists of 4213 dialogue examples. Extensive experiments and studies have demonstrated the superiority of TriKF across various metrics on the EQT and EQ-EMAC datasets.

DATA AVAILABILITY STATEMENT

The source code and data are available at https://github.com/slptongji/TriKF.

REFERENCES

[1] H. Riess, "The science of empathy," *J. Patient Experience*, vol. 4, no. 2, pp. 74–77, 2017.
[2] T. Singer and O. M. Klimecki, "Empathy and compassion," *Current Biol.*, vol. 24, no. 18, pp. R875–R878, 2014.

[3] A. Sharma, A. Miner, D. Atkins, and T. Althoff, "A computational approach to understanding empathy expressed in text-based mental health support," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 5263–5276.

[4] L. Zhou, J. Gao, D. Li, and H.-Y. Shum, "The design and implementation of Xiaoice, an empathetic social chatbot," *Comput. Linguistics*, vol. 46, no. 1, p. 53–93, 2020.

[5] K. Mishra, A. M. Samad, P. Totala, and A. Ekbal, "PEPDS: A polite and empathetic persuasive dialogue system for charity donation," in *Proc. Int. Conf. Comput. Linguistics*, 2022, pp. 424–440.

[6] A. Wecker, T. Kuflik, P. Mulholland, B. Diaz-Agudo, and T. Pedersen, "Introducing empathy into recommender systems as a tool for promoting social cohesion," in *Proc. Workshop Social Cult. Integr. Personalized Interfaces*, 2021, pp. 1–6.

[7] A. Salehi-Abari, C. Boutilier, and K. Larson, "Empathetic decision making in social networks," *Artif. Intell.*, vol. 275, pp. 174–203, Oct. 2019.

[8] S. Liu et al., "Towards emotional support dialog systems," in *Proc. Annu. Meeting Assoc. Comput. Linguistics Int. Joint Conf. Natural Lang. Process.*, Online: Association for Computational Linguistics, 2021, pp. 3469–3483.

[9] A. Welivita and P. Pu, "A taxonomy of empathetic response intents in human social conversations," in *Proc. Int. Conf. Comput. Linguistics*, D. Scott, N. Bel, and C. Zong, eds., 2020, pp. 4886–4899.

[10] I. A. James and R. Morse, "The use of questions in cognitive behaviour therapy: Identification of question type, function and structure," *Behav. Cogn. Psychotherapy*, vol. 35, pp. 507–511, Jul. 2007.

[11] K. Huang, M. Yeomans, A. W. Brooks, J. Minson, and F. Gino, "It doesn't hurt to ask: Question-asking increases liking," *J. Personality Social Psychol.*, vol. 113, no. 3, 2017, Art. no. 430.

[12] G. I. Clark and S. J. Egan, "The Socratic method in cognitive behavioural therapy: A narrative review," *Cogn. Therapy Res.*, vol. 39, pp. 863–879, Dec. 2015.

[13] I. A. James, R. Morse, and A. Howarth, "The science and art of asking questions in cognitive therapy," *Behav. Cogn. Psychotherapy*, vol. 38, pp. 83–93, 2009.

[14] E. Svikhnushina, I. Voinea, A. Welivita, and P. Pu, "A taxonomy of empathetic questions in social dialogs," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 2952–2973.

[15] Q. Li, P. Li, Z. Ren, P. Ren, and Z. Chen, "Knowledge bridging for empathetic dialogue generation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 10, 2022, pp. 10993–11001.

[16] S. Sabour, C. Zheng, and M. Huang, "CEM: Commonsense-aware empathetic response generation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 10, 2022, pp. 11229–11237.

[17] L. Wang et al., "Empathetic dialogue generation via sensitive emotion recognition and sensible knowledge selection," in *Proc. Findings Assoc. Comput. Linguistics (EMNLP)*, 2022, pp. 4634–4645.

[18] J. Zhou, C. Zheng, B. Wang, Z. Zhang, and M. Huang, "CASE: Aligning coarse-to-fine cognition and affection for empathetic response generation," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2023, pp. 8223–8237.

[19] Z. Lin, A. Madotto, J. Shin, P. Xu, and P. Fung, "Moel: Mixture of empathetic listeners," in *Proc. Conf. Empirical Methods Natural Lang. Process. Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 121–132.

[20] N. Majumder et al., "MIME: MIMicking emotions for empathetic response generation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 8968–8979.

[21] Y. Su et al., "RLCA: Reinforcement learning model integrating cognition and affection for empathetic response generation," *IEEE Trans. Comput. Social Syst.*, vol. 11, no. 1, pp. 1158–1168, Feb. 2024.

[22] J. S. Beck, *Cognitive Behavior Therapy: Basics and Beyond*. New York, NY, USA: The Guilford Publications, 2020.

[23] J. A. Cully and A. L. Teten, *A Therapist's Guide to Brief Cognitive Behavioral Therapy*. Houston, TX, USA: Department of Veterans Affairs South Central MIRECC, 2008.

[24] F. B. Siddique, O. Kampman, Y. Yang, A. Dey, and P. Fung, "Zara returns: Improved personality induction and adaptation by an empathetic virtual agent," in *Proc. ACL Syst. Demonstrations*, 2017, pp. 121–126.

[25] P. Zhong, C. Zhang, H. Wang, Y. Liu, and C. Miao, "Towards persona-based empathetic conversational models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, B. Webber, T. Cohn, Y. He, and Y. Liu, eds., 2020, pp. 6556–6566.

[26] 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.*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, eds., 2021, pp. 2227–2240.

[27] Q. Li, H. Chen, Z. Ren, P. Ren, Z. Tu, and Z. Chen, "EmpDG: Multi-resolution interactive empathetic dialogue generation," in *Proc. Int. Conf. Comput. Linguistics*, 2020, pp. 4454–4466.

[28] C. Zheng, Y. Liu, W. Chen, Y. Leng, and M. Huang, "CoMAE: A multi-factor hierarchical framework for empathetic response generation," in *Proc. Findings Assoc. Comput. Linguistics: ACL-IJCNLP*, 2021, pp. 813–824.

[29] M. Y. Chen, S. Li, and Y. Yang, "EmpHi: Generating empathetic responses with human-like intents," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2022, pp. 1063–1074.

[30] F. Fu, L. Zhang, Q. Wang, and Z. Mao, "E-CORE: Emotion correlation enhanced empathetic dialogue generation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2023, pp. 10568–10586.

[31] X. Pang, Y. Wang, S. Fan, L. Chen, S. Shang, and P. Han, "EmpMFF: A multi-factor sequence fusion framework for empathetic response generation," in *Proc. ACM Web Conf.*, 2023, pp. 1754–1764.

[32] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 5370–5381.

[33] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, and Y. Choi, "COMET: Commonsense transformers for automatic knowledge graph construction," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 4762–4779.

[34] J. Amidei, P. Piwek, and A. Willis, "Evaluation methodologies in automatic question generation 2013-2018," in *Proc. Int. Conf. Natural Lang. Gener.*, 2018, pp. 307–317.

[35] R. Zhang, J. Guo, L. Chen, Y. Fan, and X. Cheng, "A review on question generation from natural language text," *ACM Trans. Inf. Syst.*, vol. 40, no. 1, pp. 1–43, 2021.

[36] N. Mulla and P. Gharpure, "Automatic question generation: A review of methodologies, datasets, evaluation metrics, and applications," *Prog. Artif. Intell.*, vol. 12, no. 1, pp. 1–32, 2023.

[37] N. Duan, D. Tang, P. Chen, and M. Zhou, "Question generation for question answering," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 866–874.

[38] C. Lyu, L. Shang, Y. Graham, J. Foster, X. Jiang, and Q. Liu, "Improving unsupervised question answering via summarization-informed question generation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 4134–4148.

[39] X. Du, J. Shao, and C. Cardie, "Learning to ask: Neural question generation for reading comprehension," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 1342–1352.

[40] L. Murakhovs'ka, C.-S. Wu, P. Laban, T. Niu, W. Liu, and C. Xiong, "MixQG: Neural question generation with mixed answer types," in *Proc. Findings Assoc. Comput. Linguistics: NAACL*, 2022, pp. 1486–1497.

[41] J. Gu, M. Mirshekari, Z. Yu, and A. Sisto, "ChainCQG: Flow-aware conversational question generation," in *Proc. Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2021, pp. 2061–2070.

[42] G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. Al-Emari, "A systematic review of automatic question generation for educational purposes," *Int. J. Artif. Intell. Educ.*, vol. 30, pp. 121–204, Mar. 2020.

[43] Z. Zhao, Y. Hou, D. Wang, M. Yu, C. Liu, and X. Ma, "Educational question generation of children storybooks via question type distribution learning and event-centric summarization," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 5073–5085.

[44] Z. Wang, A. S. Lan, W. Nie, A. E. Waters, P. J. Grimaldi, and R. G. Baraniuk, "QG-Net: A data-driven question generation model for educational content," in *Proc. Annu. ACM Conf. Learn. Scale*, 2018, pp. 1–10.

[45] Y. Wang, C. Liu, M. Huang, and L. Nie, "Learning to ask questions in open-domain conversational systems with typed decoders," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2193–2203.

[46] B. Pan, H. Li, Z. Yao, D. Cai, and H. Sun, "Reinforced dynamic reasoning for conversational question generation," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 2114–2124.

[47] Y. Ling, F. Cai, H. Chen, and M. de Rijke, "Leveraging context for neural question generation in open-domain dialogue systems," in *Proc. Web Conf.*, 2020, pp. 2486–2492.

[48] M. Aliannejadi, H. Zamani, F. Crestani, and W. B. Croft, "Asking clarifying questions in open-domain information-seeking conversations," in *Proc. Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval*, 2019, pp. 475–484.

[49] L. Shen, F. Meng, J. Zhang, Y. Feng, and J. Zhou, "GTM: A generative triple-wise model for conversational question generation," in *Proc. Annu. Meeting Assoc. Comput. Linguistics Int. Joint Conf. Natural Lang. Proc.*, 2021, pp. 3495–3506.
[50] Y. Gao, P. Li, I. King, and M. R. Lyu, "Interconnected question generation with coreference alignment and conversation flow modeling," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 4853–4862.
[51] J. D. Hwang et al., "(comet-) Atomic 2020: On symbolic and neural commonsense knowledge graphs," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 7, 2021, pp. 6384–6392.
[52] M. Lewis et al., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 7871–7880.
[53] Y. Li et al., "Towards an online empathetic chatbot with emotion causes," in *Proc. Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval*, 2021, pp. 2041–2045.
[54] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543.
[55] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.
[56] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *Proc. AAAI Conf. Artif. Intell.*, vol. 30, no. 1, 2016, pp. 3776–3783.
[57] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.
[58] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization*, 2005, pp. 65–72.

Tiantian Chen received the B.S. degree in software engineering from the School of Software Engineering, Tongji University, Shanghai, China, in 2021, where she is currently working toward the Ph.D. degree in software engineering.

Her research interests include dialogue systems, sentiment analysis, and natural language processing.

Ying Shen received the B.S. and M.S. degrees in software engineering from the School of Software, Shanghai Jiao Tong University, Shanghai, China, in 2006 and 2009, respectively, and the Ph.D. degree in computer science 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, China, where she is currently an Associate Professor. Her research interests include speech and language processing, and sentiment analysis.

Xuri Chen received the M.S. degree in developmental and educational psychology from the School of Psychology, Beijing Normal University, Beijing, China, in 2004. She is currently working toward the Ph.D. degree in psychology with the School of Humanities, Tongji University, Shanghai, China.

She joined Tongji University, in 2004, where she is engaged in student management and mental health education. She is a national level 2 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.

Lin Zhang (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in computer science from the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2003, and 2006, respectively, and the Ph.D. degree in computer science from the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, in 2011.

From March 2011 to August 2011, he was a Research Associate with the Department of Computing, The 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 research interests include environment perception of intelligent vehicle, pattern recognition, computer vision, and perceptual image/video quality assessment. He serves as an Associate Editor for IEEE ROBOTICS AND AUTOMATION LETTERS and *Journal of Visual Communication and Image Representation*.

Dr. Zhang was awarded as a Young Scholar of Changjiang Scholars Program, Ministry of Education, China.

Shengjie Zhao (Senior Member, IEEE) received the B.S. degree in electrical engineering from the University of Science and Technology of China, Hefei, China, in 1988, the M.S. degree in electrical and computer engineering from the China Aerospace Institute, Beijing, China, in 1991, and the Ph.D. degree in electrical and computer engineering from Texas A&M University, College Station, TX, USA, in 2004.

He is currently the Dean of 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, China. His research interests include artificial intelligence, big data, wireless communications, image processing, and signal processing.

Prof. Zhao is a Fellow of the Thousand Talents Program of China and an Academician of the International Eurasian Academy of Sciences.