Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Situation-aware empathetic response generation

Zhou Yang [a,c,d], Zhaochun Ren [b], Yufeng Wang [a,c,d], Haizhou Sun [g], Xiaofei Zhu [f], Xiangwen Liao [a,c,d,e,*]

[a] *College of Computer and Data Science, Fuzhou University, Fuzhou, 350108, China*
[b] *Leiden University, Leiden, 2333 CA, Netherlands*
[c] *Fujian Provincial Key Laboratory of Networking Computing and Intelligent Information Processing, Fuzhou University, Fuzhou, 350108, China*
[d] *Digital Fujian Institute of Financial Big Data, Fuzhou University, Fuzhou, 350108, China*
[e] *Research Center for Cyberspace Security, Peng Cheng Laboratory, Shenzhen, 518000, China*
[f] *College of Computer Science and Technology, Chongqing University of Technology, Chongqing, 400054, China*
[g] *SmartMore, Beijing, 100006, China*

ARTICLE INFO

Keywords:
Empathetic response generation
Text generation
Emotional detection
Natural language processing

ABSTRACT

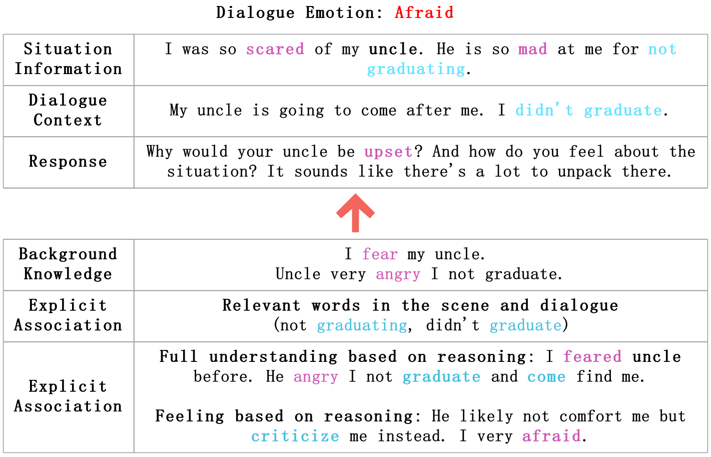
Empathetic response generation endeavours to perceive the interlocutor's emotional and cognitive states in the dialogue and express proper responses. Previous studies detect the interlocutor's states by understanding the immediate context of the dialogue. However, these methods are at an elementary/intermediate level of empathetic understanding due to the neglect of the broader context (i.e., the situation) and its associations with the dialogue, leading to inaccurate comprehension of the interlocutor's states. In this paper, we utilize the EMPATHETIC-DIALOGUES dataset consisting of 25k dialogues, and on this basis, we propose a Situation-Dialogue Association Model (SDAM). SDAM focuses on the broader context, i.e., the situation, and enhances the understanding of empathy from explicit and implicit associations. Regarding explicit associations, we propose a bidirectional filtering encoder. It selects relevant keywords between the situation and dialogue, learning their direct lexical relevance. For implicit associations, we use a knowledge-based hypergraph network grounded to learn convoluted connections between the situation and the dialogue. Moreover, we also introduce a simple fine-tuning approach that combines SDAM with large language models to further strengthen the empathetic understanding capability. Compared to the baseline, SDAM demonstrates superior empathetic ability. In terms of emotion accuracy, fluency, and response diversity (Distinct-1/Distinct-2), SDAM achieves improvements of 12.25 (a 30.47% increase), 0.3 (a 0.85% increase), and 0.86/1.23 (116.22% and 30.67% increases), respectively. Additionally, our variant model based on large language models exhibits better emotion recognition capability without compromising response quality, specifically achieving an improvement of 0.23 (a 0.37% increase) in emotion accuracy.

1. Introduction

Endowing machines with human-like traits is a crucial step in developing anthropomorphic conversational agents. Multiple studies have demonstrated that traits such as emotionality (Ghosh, Chollet, Laksana, Morency, & Scherer, 2017; Yuan, Wang, Yu, & Zhang, 2022), personalization (Capel & Brereton, 2023; Chen, Wang, Yu, & Zhang, 2023; Kim, Kim, & Kim, 2020; Lynn, Son,

[*] Corresponding author at: College of Computer and Data Science, Fuzhou University, Fuzhou, 350108, China.
E-mail address: liaoxw@fzu.edu.cn (X. Liao).

https://doi.org/10.1016/j.ipm.2024.103824
Received 9 March 2024; Received in revised form 17 May 2024; Accepted 20 June 2024
Available online 12 July 2024



Dialogue Emotion: Afraid

Situation Information	I was so scared of my uncle. He is so mad at me for not graduating.
Dialogue Context	My uncle is going to come after me. I didn't graduate.
Response	Why would your uncle be upset? And how do you feel about the situation? It sounds like there's a lot to unpack there.

Background Knowledge	I fear my uncle. Uncle very angry I not graduate.
Explicit Association	Relevant words in the scene and dialogue (not graduating, didn't graduate)
Explicit Association	Full understanding based on reasoning: I feared uncle before. He angry I not graduate and come find me. Feeling based on reasoning: He likely not comfort me but criticize me instead. I very afraid.

Fig. 1. An example from the EMPATHETIC-DIALOGUES dataset. The situation and its explicit and implicit associations with the dialogue facilitate an accurate and comprehensive understanding of the dialogue.

Kulkarni, Balasubramanian, & Schwartz, 2017; Zhang, Wang, Yu, Xu, & Zhang, 2024), and empathy (Fitzpatrick, Darcy, & Vierhile, 2017; Zhong et al., 2021) significantly enhance machine performance across various tasks. This paper focuses on imbuing machines with the trait of empathy, specifically addressing the empathetic response generation task. The task aims to perceive the emotional and cognitive states of interlocutors and generate appropriate responses accordingly (Curry & Curry, 2023; Rashkin, Smith, Li, & Boureau, 2019; Sabour, Zheng, & Huang, 2022).

According to the empathy development theory, the understanding of empathy is categorized into four stages, progressing from low to high (Hoffman, 1975, 1977, 1987, 2001). For the first three stages, understanding empathy only requires considering the immediate context in which the interlocutor is situated, i.e., the dialogue context. However, the highest stage of empathy necessitates additional attention to the broader context, i.e., the situation information. Existing work perceives emotional states (Li et al., 2020; Li, Li, Ren, Ren, & Chen, 2022; Lin, Madotto, Shin, Xu, & Fung, 2019; Majumder et al., 2020; Rashkin et al., 2019; Yang et al., 2023) or emotional and cognitive states (Sabour et al., 2022; Wang et al., 2022; Zhao, Zhao, Lu, & Qin, 2023; Zhou, Zheng, Wang, Zhang, & Huang, 2023) from the dialogue context while neglecting the situation information in which the dialogue takes place. Due to the neglect of background knowledge present in the situation information, these methods are more prone to misunderstanding emotional and cognitive states. As illustrated in Fig. 1, the dialogue context conveys the implication that "the speaker has not graduated, and their uncle is about to visit the school". Simultaneously, the situation information expresses the implication that "the speaker is afraid that their intimidating uncle will be angry due to their failure to graduate". Disregarding the situation information, previous methods cannot accurately comprehend the speaker's emotion as "fear". For instance, if the uncle were a considerate individual, he would be more likely to visit the school to console the speaker. Consequently, the speaker is more prone to express a sense of "care" rather than "fear". Conversely, associating relevant information from the situation information and the dialogue context, and making reasonable inferences, is more likely to accurately comprehend the dialogue. For example, linking "not graduating" and "didn't graduate" reveals the same implication of "failure to graduate". Simultaneously, connecting "my scary uncle who will be angry that I didn't graduate" and "my uncle is about to come see me" accurately reflects the speaker's states of "afraid" and "fear of being reprimanded by their uncle". Therefore, understanding the situation information and its associations with the context facilitates accurate comprehension of the user's state, yet this aspect remains unexplored.

In this paper, we propose a Situation-Dialogue Association Model (SDAM) for empathetic response generation. Inspired by the situation model (Pickering & Garrod, 2004), we introduce situation information and categorize its associations with the dialogue context into explicit and implicit associations. Explicit associations refer to direct word-level connections, while implicit associations require inference to uncover the underlying connections. To learn explicit associations, we propose a bidirectional filtering encoder, which selects and encodes keywords relevant to both the situation and the context. Since hypergraph neural network can effectively capture complex associations (Feng, You, Zhang, Ji, & Gao, 2019; Wang et al., 2019; Yu, Tao, & Wang, 2012), we use a reasoning knowledge-based hypergraph neural network to capture complex implicit associations in situational and dialogical reasoning knowledge. To flexibly convey empathy in the situation and the dialogue, we adopt an adjustable situation-dialogue decoder to generate empathetic responses.

Experiments on the EMPATHETICDIALOGUES dataset (Rashkin et al., 2019) demonstrate SDAM significantly outperforms state-of-the-art baselines with stronger perception and empathy. Further analyses verify that both explicit and implicit situation-dialogue associations substantially contribute to the understanding and empathy.

2. Research objectives

To clearly articulate the research objectives of this paper, we delineate the limitations of existing research, the problems to be addressed, and the contributions of our proposed approach.

Limitations of Existing Research. Existing studies detect the interlocutor's emotional and cognitive states by understanding the immediate context within the dialogue. As a broader context, the situation is closely associated with the dialogue and crucial for comprehending the interlocutor's states. However, existing methods neglect the situation, leading to inaccurate understanding of the interlocutor's states.

Problems to be Addressed. The problem we face is how to leverage the situation to enhance the empathetic capability of the model, i.e., how to better understand the interlocutor's states by exploiting the close associations between the situation and the dialogue.

Contributions of Our Research. Our main contributions are as follows: (1) We introduce the situation and explore its explicit and implicit associations with the dialogue to accurately comprehend the dialogue and generate empathetic responses. (2) We propose a bidirectional filtering encoder and a reasoning knowledge-based hypergraph neural network to respectively focus on the explicit and implicit associations. (3) We introduce a simple fine-tuning approach that combines our small-scale model with large language models, enhancing the emotion perception capability of large language models without compromising response quality. (4) Experimental results demonstrate that our proposed approach improves the empathetic ability of the model and generates more appropriate responses.

3. Literature review

Endowing anthropomorphic conversational agents with key human-like traits such as emotionality (Ghosh et al., 2017; Yuan et al., 2022), personalization (Capel & Brereton, 2023; Chen et al., 2023; Kim et al., 2020; Lynn et al., 2017; Zhang et al., 2024), and empathy (Fitzpatrick et al., 2017; Zhong et al., 2021) have been demonstrated to significantly enhance machine performance across various tasks. We focus on imbuing machines with the trait of empathy, specifically addressing the empathetic response generation task (Rashkin et al., 2019). This task requires the automatic perception of the emotions and cognition of the interlocutor, and the generation of appropriate responses accordingly (Sabour et al., 2022; Zhao et al., 2023). To clearly present the relevant work, we elaborate from three aspects: emotion, cognition, and empathetic response generation models.

Emotion. Emotion is an important aspect of empathy, which has promoted the models' performance in various tasks (Chen, Wang, & Zhang, 2021; Fitzpatrick et al., 2017; Ghosh et al., 2017; Rashkin et al., 2019). Previous research has represented emotion as continuous or discrete representations to incorporate it into models. One part of the research views emotion as continuous vector representations, focusing on sentence-level emotion (Calvo & Mac Kim, 2013; Lee, Li, & Yu, 2022), multi-granularity emotion (Buechel & Hahn, 2017), and relations between multiple dimensions (Xie, Lin, Lin, Wang, & Yu, 2021). Another part views emotion as discrete emotion labels, generally categorized into 2 to 32 emotion classes (Rashkin et al., 2019; Zhong et al., 2021; Zhong, Zhang, Wang, Liu, & Miao, 2020). Compared to the former, the latter is more easily understood and interpreted by humans, and thus is commonly used for emotion labels. In this paper, 32 discrete emotion classes are adopted.

Cognition. Cognition is another important aspect of empathy (Sabour et al., 2022). Previous research enhances the cognitive ability of models by incorporating external knowledge. One line of research has augmented cognition by integrating ConceptNet knowledge (Li et al., 2022; Zhong et al., 2021). Other studies have introduced commonsense reasoning knowledge to strengthen cognitive capabilities (Sabour et al., 2022; Zhao et al., 2023; Zhou et al., 2023). As the latter approach considers the user's state more comprehensively, it has achieved superior performance.

Empathetic Response Generation Models. Early empathetic response generation models focus on the emotional aspect of empathy from multiple perspectives. These studies explore dialogue-level emotions (Rashkin et al., 2019), mixed emotions (Lin et al., 2019; Majumder et al., 2020), and word-level emotions (Li et al., 2020, 2022) respectively. As empathy also encompasses the cognitive aspect (Cuff, Brown, Taylor, & Howat, 2016; Davis, 1983; Elliott, Bohart, Watson, & Murphy, 2018), later research additionally considered cognitive factors. These studies enhance the cognitive abilities of models by focusing on commonsense knowledge (Sabour et al., 2022), emotion fluidity (Wang et al., 2022), self-other awareness (Zhao et al., 2023), dynamic fusion of commonsense knowledge (Bi et al., 2023), and alignment between emotion and cognition (Zhou et al., 2023).

However, existing methods are all based on the immediate context, namely the dialogue context. According to psychology study (Hoffman, 1975), more advanced empathy requires paying attention to broader context, such as the situation. Unlike the above methods, we introduce the situation, and consider both explicit and implicit associations between the situation and the dialogue to further enhance the model's empathic ability.

4. Method

4.1. Overview

Our proposed Situation-Dialogue Association Model (SDAM) considers situations, dialogues and their associations, which accurately and comprehensively understand dialogues and generate more empathetic responses. As shown in Fig. 2, SDAM is a transformer-based model composed of four parts: (1) context and situation encoders, which encode the context and the situation to fully comprehend the dialogue; (2) a bidirectional filtering encoder, which selects and encodes keywords relevant to the situation and the context to learn explicit associations; (3) a reasoning knowledge-based hypergraph neural network. It utilizes a hypergraph neural network to capture complex implicit associations in situational and dialogical reasoning knowledge. (4) emotion and response prediction module. It first uses aggregation attention to aggregate the situation, context, and their associations to predict the emotion category. Then, it adopts an adjustable situation-context decoder to predict the response.

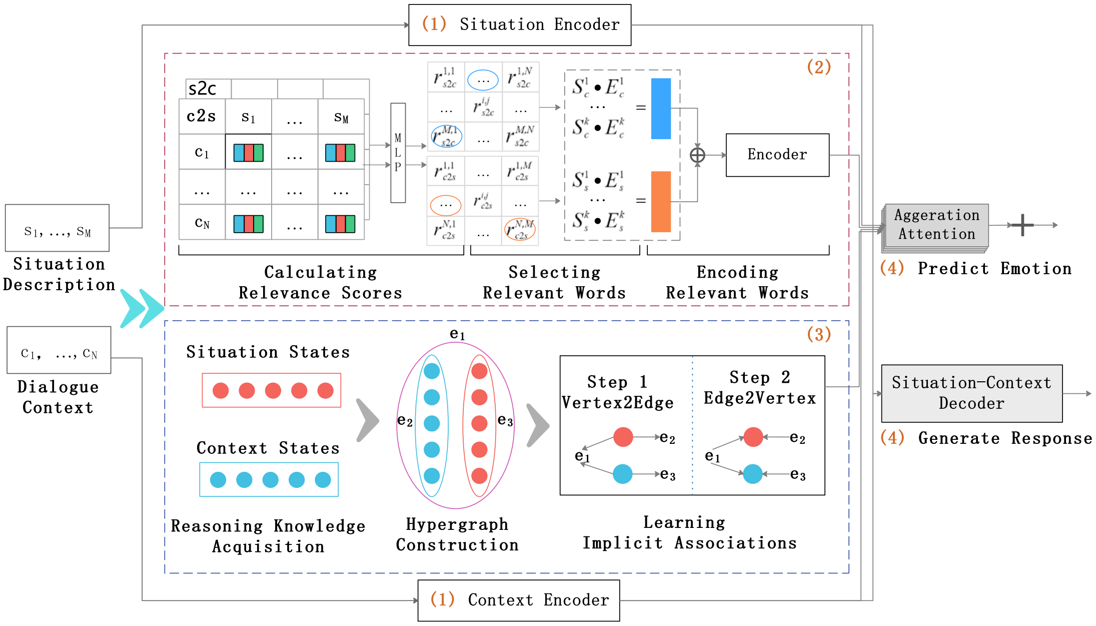


Fig. 2. An overview of SDAM. It consists of four parts: (1) context and situation encoders, which comprehensively understand the dialogue; (2) bidirectional filtering encoder, which selects and encodes keywords relevant to both the situation and the context to learn explicit associations; (3) a reasoning knowledge-based hypergraph neural network. It utilizes the hypergraph neural network to capture complex implicit associations in reasoning knowledge; (4) emotion and response prediction module, which employs aggregation attention to perceive the situation, context and associations to predict emotion categories. It then utilizes an adjustable situation-context decoder to generate empathetic responses.

4.2. Task formulation

Given the dialogue context $D = [U_1, U_2, \dots, U_{N_d}]$ and the situation information $S = [s_1, s_2, \dots, s_M]$, our goal is to understand the communicator's states in the dialogue, predict the dialogue emotion E, and generate an appropriate response $Y = [y_1, y_2, \dots, y_j, y_L]$. Here, $U_i = [w^i_1, w^i_2, \dots, w^i_{n_i}]$ represents the ith utterance with n_i words. S represents a sequence with M words. Y is a sequence with L words.

4.3. Context and situation encoders

We encode the context and the situation respectively to understand the immediate dialogue content and the background information.

Regarding the context, similar to previous approaches (Li et al., 2022; Sabour et al., 2022), we first concatenate the utterances in the dialogue context and prepend $[CLS]$ to form the dialogue input sequence $C = U_1 \oplus U_2 \oplus \cdots \oplus U_{N_d}$. $[CLS]$ represents the overall semantic token and serves as the start symbol, while $\oplus$ is the concatenation symbol. Based on the dialogue input sequence C, we convert it into word embeddings and sum them with the user type embeddings to form the context embeddings E_c. Subsequently, we sum the context embeddings E_c and position embeddings to obtain $\widetilde{E}_c$, which is then inputted into the context encoder $\mathbf{Enc}_{ctx}$ to obtain the context representation H_{ctx}. Here, the user type embeddings are used to differentiate between speakers and listeners and are randomly initialized.

$$H_{ctx} = \mathbf{Enc}_{ctx}(\widetilde{E}_c) \tag{1}$$

where $H_{ctx} \in R^{N \times d}$, N represents the length of the context sequence, i.e., dialogue input sequence. And d is the hidden size of the encoder.

In terms of the situation, we transform the sequence of situation informations S into situation embeddings E_s. Subsequently, we sum the situation embeddings and position embeddings to obtain $\widetilde{E}_s$, which is then fed into the situation encoder $\mathbf{Enc}_{sit}$ to learn the situation and obtain the situation representation H_{sit}.

$$H_{sit} = \mathbf{Enc}_{sit}(\widetilde{E}_s) \tag{2}$$

where $H_{sit} \in R^{M \times d}$, M represents the length of the situation information sequence.

4.4. Bidirectional filtering encoder

Explicit associations refer to direct connections in the situation and the dialogue (Pickering & Garrod, 2004). Since in this task, both the situation and dialogue are described through words, the most direct connection is related keywords. To capture explicit associations embodied in words, we propose a bidirectional filtering encoder, which consists of three steps: calculating relevance scores, selecting relevant words, and encoding relevant words.

Calculating Relevance Scores. We design a relevance function **Func** to calculate the relevance between the situation and context. We input the situation embedding E_s and the context embedding E_c into the relevance function **Func** in different orders to compute two relevance scores, i.e., the context-to-situation relevance score $r_{c2s} \in R^{M\times N}$ and the situation-to-context relevance score $r_{s2c} \in R^{N\times M}$.

$$r_{c2s} = \mathbf{Func}(E_c, E_s) \tag{3}$$

$$r_{s2c} = \mathbf{Func}(E_s, E_c) \tag{4}$$

To describe the relevance function clearly and concisely, we take the context-to-situation order as an example to illustrate the computation process. Specifically, we feed the situation embedding E_s and context embedding E_c separately into two linear layers and calculate their relevance degree E_{dot} using dot product. Additionally, we use the learnable weight parameters w_{com} to compress the context and situation embeddings, resulting in compressed context embedding E_θ and situation embedding E_ε. The relevance degree indicates the strength of the association between the situation and context words, while the compressed embeddings indicate the meaning of the words.

$$E_{dot} = (w_\theta^{dot} E_c + b_\theta^{dot}) \cdot (w_\varepsilon^{dot} E_s + b_\varepsilon^{dot})^T \tag{5}$$

$$E_\theta = w_{com} E_c, E_\varepsilon = w_{com} E_s \tag{6}$$

where w_θ^{dot}, w_ε^{dot}, b_θ^{dot}, b_ε^{dot}, w_{com} are learnable weights. w_θ^{dot}, $w_\varepsilon^{dot} \in R^{d\times d}$, b_θ^{dot}, $b_\varepsilon^{dot} \in R^d$, $w_{com} \in R^{d\times d_{com}}$. Meanwhile, $E_{dot} \in R^{N\times M}$, $E_\theta \in R^{N\times d_{com}}$, $E_\varepsilon \in R^{M\times d_{com}}$, where d_{com} is a hyperparameter for the compressed dimension.

To consider both association strength and word meaning, we concatenate two types of vectors and input them into a linear layer with activation functions to obtain the context-to-situation relevance score r_{c2s}.

$$E_{dc}^{i,j} = E_\theta^i \oplus E_\varepsilon^j \oplus E_{dot}^{i,j}, i \in [1, n]; j \in [1, m] \tag{7}$$

$$r_{c2s} = \sigma_2(w_{ea}\sigma_1(E_{dc}) + b_{ea}) \tag{8}$$

where $E_{dc}^{i,j} \in R^{N\times M\times(2d_{com}+1)}$, $r_{c2s} \in R^{N\times M}$, σ_1, σ_2 are the Sigmoid and ReLU activation functions, respectively.

Selecting Relevant Words. Based on the context-to-situation relevance score, we select the top k situation words with the highest relevance scores. By multiplying the relevant scores $S_c^{k_i}$ with the word embeddings $E_c^{k_i}$, we obtain the relevant situation embedding of k words. In the same way, we process the situation-to-context relevance scores to obtain the relevant context embeddings of k words. Afterwards, we concatenate the above relevant embeddings of the situation and context words to obtain the bidirectional relevant embedding $\widetilde{r}_{ea}$.

$$S_s^{k_i}, E_s^{k_i} = \mathbf{TopK}(r_{d2s}), k_i \in [1, k] \tag{9}$$

$$S_c^{k_i}, E_c^{k_i} = \mathbf{TopK}(r_{s2d}), k_i \in [1, k] \tag{10}$$

$$\widetilde{r}_{ea} = \bigoplus_{k_i\in[1,k]}^{type\in\{c,s\}} S_{type}^{k_i} \cdot E_{type}^{k_i} \tag{11}$$

where $S_s^{k_i}$, $S_c^{k_i} \in R^{k\times 1}$ represent the top k scores for the situation and context. $E_s^{k_i}$, $E_c^{k_i} \in R^{k\times d}$ are the k word embeddings of the situation and context corresponding to the top k scores, where k is a hyperparameter. Moreover, $\widetilde{r}_{ea} \in R^{2k\times d}$.

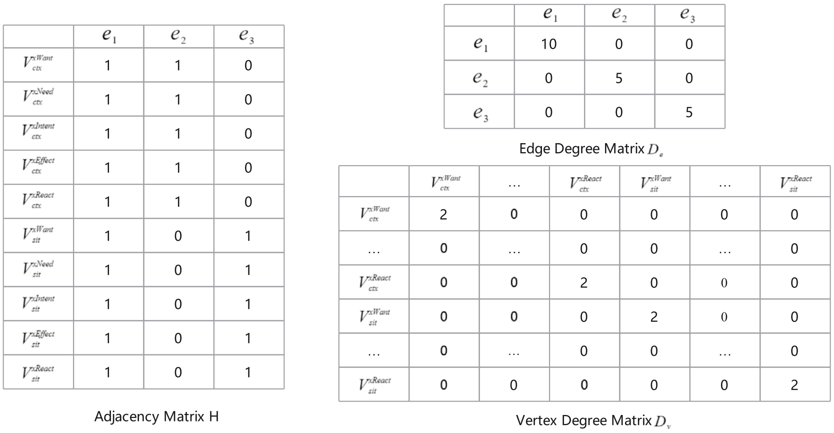
Encoding Relevant Words. By encoding the bidirectional relevant embedding $\widetilde{r}_{ea}$, we obtain explicit association representations $r_{ea} \in R^{2k\times d}$ between the situation and the context.

$$r_{ea} = \mathbf{Enc}_{ea}(\widetilde{r}_{ea}) \tag{12}$$

4.5. Reasoning knowledge-based hypergraph neural network

Since the complex implicit associations often manifested in reasoning knowledge (Pickering & Garrod, 2004), we employ the reasoning model COMET (Hwang et al., 2021) to acquire reasoning knowledge of the situation and the context. Based on the reasoning knowledge, we construct a hypergraph G and utilize a hypergraph neural network (Feng et al., 2019) to learn the complex implicit associations.

Reasoning Knowledge Acquisition. Similar to Sabour et al. (2022), we obtain reasoning knowledge from the last sentence U of the dialogue context. We respectively append ([xWant], [xNeed], [xIntent], [xEffect], [xReact]) to the utterance U and use the COMET model to generate reasoning knowledge $[ck_1, \dots, ck_i, \dots, ck_5]$, where ck_i represents the textual description of the corresponding knowledge. Among them, $cog \in [[xWant], [xNeed], [xIntent], [xEffect]]$ represent cognitive states, while $emo = [xReact]$ represents emotional states. We then convert these textual descriptions by adding the starting symbol [CLS] and learn



	e_1	e_2	e_3
V_{ctx}^{xWant}	1	1	0
V_{ctx}^{xNeed}	1	1	0
$V_{ctx}^{xIntent}$	1	1	0
$V_{ctx}^{xEffect}$	1	1	0
V_{ctx}^{xReact}	1	1	0
V_{sit}^{xWant}	1	0	1
V_{sit}^{xNeed}	1	0	1
$V_{sit}^{xIntent}$	1	0	1
$V_{sit}^{xEffect}$	1	0	1
V_{sit}^{xReact}	1	0	1

Adjacency Matrix H

	e_1	e_2	e_3
e_1	10	0	0
e_2	0	5	0
e_3	0	0	5

Edge Degree Matrix D_e

	V_{ctx}^{xWant}	…	V_{ctx}^{xReact}	V_{sit}^{xWant}	…	V_{sit}^{xReact}
V_{ctx}^{xWant}	2	0	0	0	0	0
…	0	…	0	0	…	0
V_{ctx}^{xReact}	0	0	2	0	0	0
V_{sit}^{xWant}	0	0	0	2	0	0
…	0	…	0	0	…	0
V_{sit}^{xReact}	0	0	0	0	0	2

Vertex Degree Matrix D_v

Fig. 3. Regarding the matrices of hypergraphs, these include the adjacency matrix, edge degree matrix, and vertex degree matrix. The adjacency matrix of a hypergraph, where the rows of the matrix represent the vertices and the columns represent the edges. The edge degree matrix, which is a diagonal matrix with the degrees of the edges along the diagonal. The vertex degree matrix. It is a diagonal matrix with the degrees of the vertices along the diagonal.

them using the cognitive state encoder $\mathbf{Enc}_{cs}$ and emotional state encoder $\mathbf{Enc}_{es}$, respectively. By encoding cognitive and emotional states, we obtain the cognitive state representation H_c^{cog} and the emotional state representation H_c^{emo}. Moreover, we process the situation information S in the same way and obtain the cognitive state representation H_s^{cog} and emotional state representation H_s^{emo} for the situation. For convenience, we generalize the symbols for the situation or context as $type$, where $type \in \{s, c\}$.

$$H_{type}^{cog} = \mathbf{Enc}_{cs}(E_{type}^{cog}) \tag{13}$$

$$H_{type}^{emo} = \mathbf{Enc}_{es}(E_{type}^{emo}) \tag{14}$$

where $H_{type}^{cs} \in R^{d_{cs} \times d}$, $H_{type}^{es} \in R^{d_{es} \times d}$. d_{cs}, d_{es} represent the textual description lengths for the cognitive state and emotive state, respectively.

Hypergraph Construction. Based on the aforementioned cognitive and emotional state representations, we construct an implicit association hypergraph G, which consists of vertices and hyperedges. We take the [CLS] representation of the state representations, i.e., $V_{type}^{emo} = H_{type}^{emo}[0]$ and $V_{type}^{cog} = H_{type}^{cog}[0]$, as the hypergraph vertices V. To reflect the complex associations between situations and contexts, we construct three hyperedges e_1, e_2, and e_3. e_1 connects all the vertices of the situation and context, representing the global association between the situation and context; e_2 connects the vertices of context states, which is the local association of context states; e_3 connects the vertices of situation states, representing the local association of situation states. Based on the above vertices and edges, we construct the adjacency matrix $H \in R^{10\times 3}$, the edge degree matrix $D_e \in R^{3\times 3}$, and the vertex degree matrix $D_v \in R^{10\times 10}$ for the hypergraph. To clearly describe the structures of hypergraphs, we elaborate on the adjacency matrix, edge degree matrix, and vertex degree matrix respectively. See Fig. 3 for details. **Adjacency Matrix**. Different from a normal graph, in the adjacency matrix of the hypergraph, rows represent vertices and columns represent hyperedges. If a vertex is linked with a hyperedge, the value is 1, otherwise it is 0. **Edge Degree Matrix**. The degree matrix is a diagonal matrix, and the diagonal elements represent the degree of edges. **Vertex Degree Matrix**. We present the vertex degree matrix. This matrix is also a diagonal matrix, and the diagonal elements represent the degree of vertices.

Learning Implicit Associations. Since hypergraph neural networks can effectively capture complex correlations (Feng et al., 2019), we employ a hypergraph neural network (abbreviated as HGNN) to capture the complex implicit associations between the context and the situation. In order to capture these associations, we utilize two layer networks:

$$V_1 = \sigma_2(\mathbf{HGNN}_1(V)) \tag{15}$$

$$r_{ia} = V_2 = \mathbf{HGNN}_2(V_1) \tag{16}$$

where $V_1 \in R^{10\times d_1}$, $r_{ia}, V_2 \in R^{10\times d_2}$. We alias V_2 as r_{ia} to indicate that the vertices have implicit associations.

To illustrate the learning process of implicit associations clearly, we elaborate on the hypergraph neural network layer $HGNN_l$, where l denotes the ith layer of the network. Given the adjacency matrix H, the edge degree matrix D_e, and the vertex degree matrix D_v, we first normalize the degree matrices of edges and vertices separately to obtain $\overline{D}_e \in R^{10\times 3}$ and $\overline{D}_v \in R^{3\times 10}$.

$$\overline{D}_v = {D_v}^{-\frac{1}{2}} H, \overline{D}_e = {D_e}^{-1} H^T \tag{17}$$

Then, we aggregate the vertex representations V into hyperedge representations E_v^l. Afterwards, we further aggregate the association information from the hyperedges into the vertex representations V_l. Through the two aggregation processes of vertex-to-edge and edge-to-vertex, the model is able to learn more complex associations.

$$E_v^l = {D_v}^{-\frac{1}{2}}(w_h^l * V + b_h^l) \tag{18}$$

$$V_l = \overline{D}_v w_{v2e} \overline{D}_e E_v^l \tag{19}$$

where $w_h^l \in R^{d \times d_l}$, $b_h^l \in R^{d_l}$ are trainable parameters. Meanwhile, $E_v^l \in R^{3 \times d_l}$, $V_l \in R^{10 \times d_l}$. d_l denotes the hidden size of the lth layer and is a hyperparameter. And $w_{v2e} \in R^{3 \times 3}$ is a fixed identity matrix.

4.6. Emotion and response predicting

Using the learned situation and context information, we predict the emotion category and generate empathetic responses.

Emotion Predicting. To aggregate effective information from the situation, context, and their associations, we use four aggregation attention networks with identical structures but different parameters:

$$P_{ctx}^e = \mathbf{Att}_{ctx}(H_{ctx}), P_{sit}^e = \mathbf{Att}_{sit}(H_{sit}) \tag{20}$$

$$P_{ea}^e = \mathbf{Att}_{ea}(r_{ea}), P_{ia}^e = \mathbf{Att}_{ia}(r_{ia}) \tag{21}$$

where $\mathbf{Att}_{ctx}$, $\mathbf{Att}_{sit}$, $\mathbf{Att}_{ea}$, and $\mathbf{Att}_{ia}$ are aggregation networks for the context, situation, explicit associations, and implicit associations, respectively, and P_{ctx}^e, P_{sit}^e, P_{ea}^e, P_{ia}^e are the corresponding emotion probabilities.

To illustrate the structure of the aggregation attention network, we take the aggregation attention $\mathbf{Att}_{ctx}$ regarding the dialogue context as an example. Firstly, we calculate the word probabilities of context representation H_{ctx} and sum them up to obtain the attention hidden representation H_2.

$$H_1^a = \sigma_3(w_1^a H_{ctx} + b_1^a) \tag{22}$$

$$P_s = Softmax(w_1^s H_1^a + b_1^s) \tag{23}$$

$$H_2 = \sum_{j=1}^{L} P_s[j] \cdot H_{ctx}[j] \tag{24}$$

where $H_1^a \in R^{N \times d}$, $P_s \in R^N$, $H^2 \in R^d$. $w_1^a, b_1^a, w_1^s, b_1^s$ are learnable parameters, and σ_3 is the Tanh activation function.

Then, we employ a non-linear layer and a linear layer to learn and predict the context emotion probabilities.

$$H_2^a = \sigma_3(w_2^a H_2 + b_2^a) \tag{25}$$

$$P_{ctx}^e = Softmax(w_2^s H_2^a + b_2^s) \tag{26}$$

where $H_2^a \in R^d$, $P_{ctx}^e \in R^{d_e}$. d_e denotes the number of emotion categories, which is equal to 32. $w_2^a, b_2^a, w_2^s, b_2^s$ are learnable parameters.

Subsequently, we sum up the emotion probabilities of context, situation, and two types of associations to obtain the overall emotion probability P_e. Using the emotion probability P_e, we predict the emotion category of the conversation. During the training process, we separately calculate the log-likelihood losses between the predicted emotion categories and the ground truth label e^*, and combine them as the overall emotion loss $\mathcal{L}_e$. Based on the overall emotion loss, we optimize the model.

$$P_e = P_{ctx}^e + P_{sit}^e + P_{ea}^e + P_{ia}^e \tag{27}$$

$$\mathcal{L}_e = -log(P_{ctx}^e(e^*) \cdot P_{sit}^e(e^*) \cdot P_{ea}^e(e^*) \cdot P_{ia}^e(e^*)) \tag{28}$$

Response Predicting. We employ an adjustable situation-context decoder to flexibly generate responses associated with the situation or context. To jointly attend to the situation and the context, we employ a situation decoder $\mathbf{Dec}_{sit}$ and a context decoder $\mathbf{Dec}_{ctx}$:

$$\widetilde{H}_{sit} = \mathbf{Dec}_{sit}(H_{sit}) \tag{29}$$

$$\widetilde{H}_{ctx} = \mathbf{Dec}_{ctx}(H_{ctx}) \tag{30}$$

where $\widetilde{H}_{sit} \in R^{d_t \times d}$, $\widetilde{H}_{ctx} \in R^{d_t \times d}$. d_t denotes the length of the shifted response at the t-th decode step.

In order to flexibly adjust the influence of situation and context information, we design a gating network. Through the gating network, we integrate the two types of information and obtain the hidden representation H. Similar to Li et al. (2022), we feed the hidden representation H into a pointer generation network (See, Liu, & Manning, 2017) to generate responses.

$$H = \widetilde{H}_{ctx} \oplus \widetilde{H}_{sit} \tag{31}$$

$$g = \sigma_1(w_{dec} H + b_{dec}) \tag{32}$$

$$\widetilde{H} = g \cdot \widetilde{H}_{ctx} + (1 - g) \cdot \widetilde{H}_{sit} \tag{33}$$

$$P(y_t | y < t, C, S) = \mathbf{Generator}(\widetilde{H}) \tag{34}$$

Where $H \in R^{d_t \times 2d}$, $\widetilde{H} \in R^{d_t \times d}$. $w_{dec} \in R^{d \times 1}$, $b_{dec} \in R^1$ are learnable parameters.

We subsequently adopt cross-entropy loss as the generation loss to optimize the model.

$$\mathcal{L}_{gen}(y_t) = -\sum_{t=1}^{T} log(P(y_t | y < t, C, S)) \tag{35}$$

Table 1
The details of the Empathetic-Dialogues dataset. The dataset is a 25k open-domain multi-turn dialogue dataset collected on the Amazon Mechanical Turk platform.

Aspects	Quantity
Number of dialogues	24,850
Average turns per dialogue	4.31
Average words per situation	19.8
Average words per utterance	15.2
Minimum/maximum utterances per dialogue	4/8
Number of emotion categories	32
Train/Valid/Test data split	19 533/2770/2547

Table 2
The emotion categories of the Empathetic-Dialogues dataset. These emotion categories encompass a variety of both positive and negative emotions.

Emotion categories
surprised,excited,annoyed,proud,angry,sad,grateful,lonely,impressed,afraid,disgusted,confident,terrified, hopeful,anxious,disappointed,joyful,prepared,guilty,furious,nostalgic,jealous,anticipating,embarrassed,content, devastated,sentimental,caring,trusting,ashamed,apprehensive,faithful

Total Loss. Finally, we take the sum of the emotion loss $\mathcal{L}_e$ and the generation loss $\mathcal{L}_{gen}(y_t)$ as the total loss.

$$\mathcal{L} = \mathcal{L}_{gen}(y_t) + \mathcal{L}_e \tag{36}$$

5. Experiments

5.1. Dataset

We conduct experiments on the empathetic dialogue dataset Empathetic-Dialogues (Rashkin et al., 2019). This dataset is a 25k open-domain multi-turn dialogue dataset. In this dataset, each dialogue contains a situation information (abbreviated as a situation), multiple utterances, and an emotion label for the entire dialogue. The emotion label belongs to one of 32 emotion types. To construct the experiments, we follow prior methods (Li et al., 2022; Lin et al., 2019; Sabour et al., 2022) and use a data split with a ratio of 8:1:1 for train/valid/test. The details of the dataset are shown in Table 1. Furthermore, we also list the 32 emotion types involved in the Empathetic-Dialogues dataset, as shown in Table 2.

5.2. Baselines

To compare with our proposed model SDAM, we select the following state-of-the-art baselines: **(1) Transformer** (Vaswani et al., 2017): A vanilla seq2seq model with encoder and decoder; **(2) EmoPrend-1** (Rashkin et al., 2019): A Transformer model that incorporates emotion labels from a pretrained classifier to enhance empathy; **(3) MoEL** (Lin et al., 2019): A Transformer model that softly combines emotions using multiple decoders to generate empathetic responses; **(4) MIME** (Majumder et al., 2020): A Transformer model considering polarity-based emotion clusters and mimicry for empathy; **(5) EmpDG** (Li et al., 2020): This model emphasizes the importance of user feedback and multi-resolution emotion modelling for empathetic response generation; **(6) KEMP** (Li et al., 2022): This model Employs ConceptNet knowledge to enrich implicit emotion representations for appropriate responses; **(7) CEM** (Sabour et al., 2022): The model accounts for emotional and cognitive aspects of empathy using reasoning knowledge for enhanced perception and expression. **(8) CASE** (Zhou et al., 2023): The model enhances the understanding and expression of empathy by aligning emotions and cognition from coarse to fine levels.

It is noteworthy that, for fairness, we conduct our experiments according to the following criteria: (1) We do not utilize pre-trained models. Pre-trained models can exert a significant impact on the model. Our model is not built upon pre-trained models. Therefore, we choose baselines that do not employ pre-trained models. (2) We maintain consistent dataset splitting. Different dataset splitting can lead to substantial variations in the metrics. Consequently, we directly utilize the dataset splitting methods employed by previous approaches, thereby avoiding the impact introduced by different splitting methods. (3) We keep crucial parameters consistent. To mitigate the influence of parameters, we maintain consistency with the baseline for crucial parameters, such as the batch size.

5.3. Implementation details

We conduct experiments on the EMPATHETICDIA-LOGUES dataset (Rashkin et al., 2019), which contains dialogue context and situation informations. During model initialization, we initialize the word embeddings of the situation and dialogue through Glove embeddings (Pennington, Socher, & Manning, 2014). For model hyperparameters, we first set the multi-head transformer as a network with one layer and two heads. In the bidirectional filtering encoder, we set the compressed dimension d_{com} to 10 and

Table 3
Results of automatic evaluation, where the bold numbers represent the optimal metrics. Emotion accuracy refers to the accuracy of 32 emotions. Perplexity, Distinct-1/Distinct-2 are fluency and diversity metrics for the generated responses, respectively.

Models	Emotion accuracy (Acc) ↑	Perplexity (PPL) ↓	Distinct-1 ↑	Distinct-2 ↑
Transformer	–	37.73	0.47	2.04
EmoPrend-1	33.28	38.30	0.46	2.08
MoEL	32.00	38.04	0.44	2.10
MIME	34.24	37.09	0.47	1.91
EmpDG	34.31	37.29	0.46	2.02
KEMP	39.31	36.89	0.55	2.29
CEM	39.11	36.11	0.66	2.99
CASE	40.2	35.37	0.74	4.01
SDAM	**52.45**	**35.07**	**1.6**	**5.24**

the number of filtered words k to 5. Subsequently, we set the output dimensions of the hypergraph neural networks $HGNN_1$ and $HGNN_2$ to $d_1 = 300$ and $d_2 = 50$, respectively. During model training, the Adam optimizer (Kingma & Ba, 2015) with $\beta 1 = 0.9$ and $\beta 2 = 0.98$ was used. Additionally, our model converged after 16,000 iterations on an NVIDIA Tesla T4 GPU.

5.4. Evaluation metrics

To assess model performance, we utilize both automatic metrics and human evaluations.

Automatic Evaluation Metrics. Following prior work (Li et al., 2022), we use perplexity (PPL), emotion accuracy (Acc), and two distinct metrics (Distinct-1 and Distinct-2) (Li, Galley, Brockett, Gao, & Dolan, 2016). Perplexity measures the fluency of the language. Lower perplexity indicates greater language fluency. Emotion accuracy assesses the model's capability of emotion perception. Higher emotion accuracy reflects better perception of emotions. Distinct-1/Distinct-2 measures the degree of informativeness in the generated responses. Greater distinct-1/distinct-2 shows increased response informativeness.

Human Evaluation Metrics. For human assessment, rather than using absolute 1–5 scales prone to subjective criteria differences (Sabour et al., 2022), we employ A/B testing between model response pairs (Lin et al., 2019; Majumder et al., 2020). Three professional annotators compare responses by our proposed SDAM model versus baselines for the same dialogues. If the annotator judges SDAM's response to be better, it gets a win point. If it is worse, it gets a loss point. Otherwise it is a tie. Specifically, we evaluate the performance of the responses in terms of empathy, relevance, and fluency. Empathy evaluates whether the generated response conveys an emotionally appropriate reaction. Relevance assesses the topical relevance of the generated response to the dialogue context. Fluency measures whether the response adheres to natural language expression conventions.

5.5. Main results

Automatic Evaluation Results. Table 3 shows the results of the automatic evaluation. Overall, recent models focusing on emotion and cognitive states (CEM) generally outperform earlier models focusing on emotions only (EmoPrend-1, MoEL, MIME, EmpDG, KEMP). Meanwhile, our proposed SDAM outperforms the above models and achieves the state-of-the-art. In terms of emotion accuracy, SDAM significantly outperforms the baselines. This is mainly because our model attends to the background information in the situation and reasoning knowledge in the implicit associations. In terms of fluency, SDAM also exceeds the baselines. This is primarily because the bidirectional filtering encoder captures keywords conducive to expression, and the situation decoder also provides advantages for fluent language expression. In terms of diversity, SDAM surpasses the baselines. This is mainly because the situation provides rich background information. Meanwhile, capturing explicit and implicit associations prompts the model to attend to situation and dialogue relevant information, which further promotes the generation of informative responses.

Human Evaluation Results. As shown in Table 4, SDAM also exceeds the three strongest baselines in empathy, relevancy, and fluency assessments. The improved empathy indicates our model accurately perceives the broader feelings according to the situation and the associations, while expressing suitable responses through the decoder. The superiority in relevancy shows that after considering the situation and the associations, our model comprehensively understands the topical scope and background knowledge of the dialogue. The improvement in fluency demonstrates that SDAM understands the situation's background and association information, and expresses natural responses based on this information.

5.6. Ablation study

As shown in Table 5, we conduct ablation experiments to verify the effectiveness of each module. (1) w/o $\mathbf{Enc}_{bf}$: Without the bidirectional filtering encoder, i.e., ignoring explicit associations; (2) w/o $\mathbf{HGNN}_{rk}$: Without the reasoning knowledge-based hypergraph neural network, i.e., ignoring implicit associations; (3) w/o **E&H**: Without the bidirectional filtering encoder and reasoning knowledge-based hypergraph neural network, i.e., ignoring explicit and implicit associations while retaining the background information of situation; (4) w/o $\mathbf{Dec}_{ctx}$: Removing the context decoder; (5) w/o $\mathbf{Dec}_{sit}$: Removing the situation decoder.

Table 4
Results of human evaluation. Where κ is the inter-rater agreement measured by Fleiss's kappa (Fleiss & Cohen, 1973), and $0.4 < \kappa \leq 0.6$ indicates moderate agreement of the evaluation results.

Comparisons	Aspects	Win	Tie	Lose	κ
SDAM vs. EmpDG	Empathy	**0.35**	0.46	0.19	0.47
	Relevance	**0.40**	0.40	0.20	0.42
	Fluency	**0.38**	0.49	0.13	0.42
SDAM vs. KEMP	Empathy	**0.29**	0.55	0.16	0.45
	Relevance	**0.37**	0.49	0.14	0.57
	Fluency	**0.32**	0.57	0.11	0.47
SDAM vs. CEM	Empathy	**0.37**	0.39	0.24	0.48
	Relevance	**0.40**	0.36	0.24	0.51
	Fluency	**0.32**	0.54	0.14	0.44

Table 5
Results of ablation study. w/o $\mathbf{Enc}_{bf}$, w/o $\mathbf{HGNN}_{rk}$, w/o **E&H** refers to ablations of the Bidirectional Filtering Encoder, Reasoning Knowledge-based Hypergraph Neural Network, and both models. w/o $\mathbf{Dec}_{ctx}$, w/o $\mathbf{Dec}_{sit}$ refer to variants with removal of the context decoder and situation decoder, respectively.

Models	Emotion accuracy (Acc) ↑	Perplexity (PPL) ↓	Distinct-1 ↑	Distinct-2 ↑
SDAM	**52.45**	**35.07**	1.6	5.24
w/o $\mathbf{Enc}_{bf}$	51.74	35.3	1.33	4.31
w/o $\mathbf{HGNN}_{rk}$	48.03	35.16	1.42	4.7
w/o **E&H**	46.03	35.22	1.34	4.43
w/o $\mathbf{Dec}_{ctx}$	50.71	37.27	0.47	1.49
w/o $\mathbf{Dec}_{sit}$	52.08	35.87	**1.63**	**5.36**

After removing the bidirectional filtering encoder $\mathbf{Enc}_{bf}$, the diversity decreases significantly. This is mainly because focusing on explicit associations helps capture words with important associations between the situation and the dialogue. The model generates more informative utterances after attending to these words.

Removing the reasoning knowledge-based hypergraph neural network $\mathbf{HGNN}_{rk}$ leads to a sharp decline in emotion accuracy. This indicates that the implicit emotional and cognitive associations between the situation and the dialogue greatly influence the accurate judgment of emotions.

When removing both $\mathbf{Enc}_{bf}$ and $\mathbf{HGNN}_{rk}$, i.e., keeping only the situation encoder $\mathbf{Enc}_{sit}$, the accuracy of emotion and diversity decrease sharply. This demonstrates that: the two association information greatly influence the accurate and comprehensive understanding of the dialogue, which enables the model to generate better responses. And encoding background information in the situation is necessary.

Subsequently, we remove the context decoder $\mathbf{Dec}_{ctx}$ and situation decoder $\mathbf{dec}_{sit}$ respectively. The results show that the context decoder has a huge impact on fluency and diversity, while the situation decoder has a significant impact on fluency. This indicates that the expression of responses mainly relies on contextual information, while situation information can facilitate natural and informative language expression.

5.7. Explicit association analysis

We conduct two experiments to further understand explicit associations: we first construct variant models with different explicit association words and verify their metrics on the dataset. Subsequently, we also analyse the characteristics of the optimal model's explicit association words in terms of emotion intensity.

For the variant models, we use different numbers of explicit association words (see Formulas (9) and (10) for details). The results are shown in Fig. 4. The x-axis represents the number of explicit association words, and the y-axis represents the metrics. We find that as the number of words increases, the metrics of the model first increase to an optimal point, then decrease continuously. This shows that when the number of correlation words is too small, the model cannot find words with stronger relevance well. At the same time, when the number of association words is too large, too much noise is introduced.

To validate the characteristics of explicit association words, we sort the association words by the model's attention weights (or frequency of model's attention) from high to low and calculate their emotion intensity. As shown in Fig. 5, the x-axis represents the top n words by weight, and the y-axis represents the average emotion intensity of words. The blue line sorts words by model attention weights, while the red line sorts by model attention frequency. The experimental results show that the model pays more attention to common words, but does not give higher weights to them, such as "her". At the same time, the model gives more weight to words with high emotion intensity, such as "miss". This is mainly because the model associates common words and high emotion intensity words to understand the dialogue more deeply, for example "miss her".

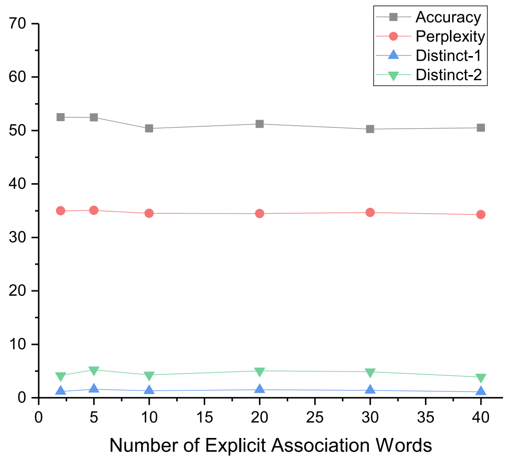


Fig. 4. Variant models of SDAM that focus on different numbers of explicit association words. As the number of words increases, the metrics of the variant models first increase then decrease.

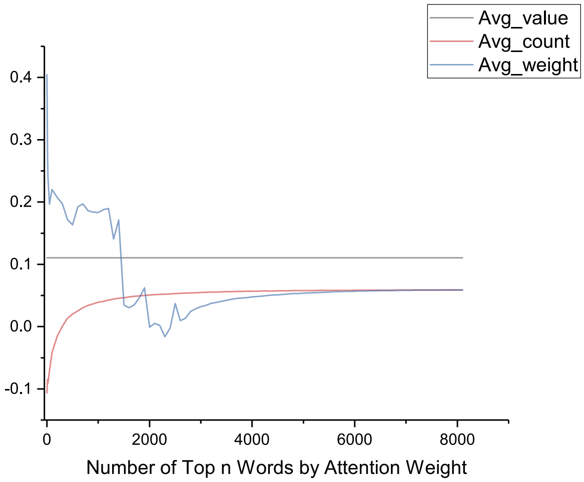


Fig. 5. Emotion intensity of words focused on by the SDAM model. The blue line shows the emotion intensity of the top k words with the highest attention weights, while the orange line represents the emotion intensity of the top k words most frequently focused on by the model. The grey line is the average emotion intensity of words in the dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5.8. Implicit association analysis

In the process for modelling implicit associations, we construct global and local association hyperedges for the hypergraph. To validate the effects of both types of edges, we conduct variant experiments. As shown in Fig. 6, the x-axis represents the metric type, and the y-axis represents the metric value. Yellow, green, and purple represent the local association model, the global association based model, and the model focusing on both associations, respectively. The experimental results show that focusing only on local or global associations will result in responses with better fluency but lower diversity. Such responses tend to be more generic, such as “I am so sorry”. This is mainly because the model needs to combine both association information to infer important keywords. At the same time, the emotion accuracy of these models is relatively close. This is mainly because emotional inference information can be obtained from both the dialogue context and situation information.

5.9. Large language model based experiments

Large language models have achieved excellent metrics on various tasks. To verify the effectiveness of association information, we also construct a variant model based on the large language model ChatGLM3 (Du et al., 2022; Zeng et al., 2023). We first use the SDAM model to extract association words from the dialogue context and situation information. Subsequently, we input these association words into the large language model, and use the LoRA-based Instruct-tuning method (Chung, Hou, Longpre, Zoph, et al., 2022; Hu et al., 2022) to train the model. $ChatGLM3_{lora}$ refers to the model using only the dialogue context. $ChatGLM3_{lora}^{sit}$

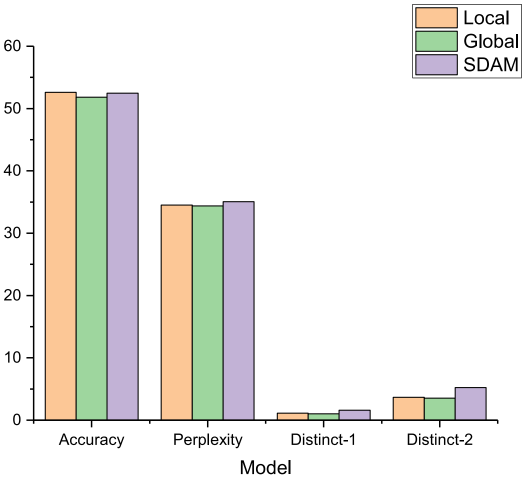


Fig. 6. Comparison for models focusing on local associations, global associations, and both. "Local" refers to the model focusing on local implicit associations, while "Global" refers to the model focusing on global implicit associations. SDAM denotes the model focusing on both. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 6
The instructions of the large language model, which consists of three parts: task definition, keyword explanation, and response formatting.

Aspects	Instructions
Task definition	You are an empathetic conversational agent. You need to simulate a listener based on the dialogue context and situation information to converse with the speaker. – For better empathetic dialogue, infer emotions from dialogue context. – Choose inferred emotions only from the provided "Emotion labels", not outside of them. – Generate appropriate responses based on dialogue history and inferred emotions. – Pay attention to the associated words, if applicable. [Optional]
Keyword explanation	– Dialogue context: The conversation history between speaker and listener, with utterances separated by </s>. – Associated words: Words that connect dialogue context and scene information. [Optional] – Situation information: Text that describes the scenario in which the dialogue takes place. – Emotion labels: surprised,excited,annoyed,proud,angry,sad,grateful,lonely,impressed, afraid,disgusted,confident,terrified,hopeful,anxious,disappointed,joyful,prepared, guilty,furious,nostalgic,jealous,anticipating,embarrassed,content,devastated, sentimental,caring,trusting,ashamed,apprehensive,faithful
Input content	– Dialogue context: [The dialogue context] – Situation information: [The Situation information] – Associated words: [The associated words between dialogue and scenario that SDAM focuses on.] [Optional]
Response formatting	– Response format: Emotion: "inferred emotion" Response: "reply utterance"

incorporates the situation information as background information. $ChatGLM3^{sit,kw}_{lora}$ considers both the situation and association words emphasized by SDAM. The detailed instructions are shown in Table 6.

As shown in Table 7, the experimental results show that the model using only context, $ChatGLM3_{lora}$, expresses more abundant responses. Since its ability to understand emotion is very weak, it cannot produce appropriate emotional responses. After incorporating the situation information as background information, the emotion perception ability of model $ChatGLM3^{sit}_{lora}$ greatly improves, but the richness of the responses decreases. On the basis of $ChatGLM3^{sit}_{lora}$, considering association words can improve emotion perception and language abilities. This shows that paying attention to keywords in the situation and dialogue can promote emotional understanding and enrich language expression.

5.10. Case study

To further analyse SDAM and the baselines, we select the three strongest baselines for case analysis. The experimental results demonstrate that SDAM can comprehensively and accurately understand dialogues and generate more empathetic responses. The cases are shown in Table 8.

In the first case, all three baselines understand the dialogue's positive emotion and express this emotion using "glad". However, because they do not pay attention to the key phrases "risk pregnancy" expressed by the speaker, they can not provide responses that are highly relevant to the dialogue context. Among them, EmpDG expresses an incorrect meaning, CEM expresses an incorrect referent with "them". The KEMP model, on the other hand, generates repetitive and generic sentences. In contrast, SDAM pays

Table 7
Results of automatic evaluation, where the bold numbers represent the optimal metrics. “lora” indicates that the model is fine-tuned on the EmpatheticDialogues dataset using the lora method. “sit” indicates that the model uses the situation information as background knowledge. “kw” indicates that the model uses the key words focused on by SDAM.

Models	Emotion accuracy (Acc) ↑	Distinct-1 ↑	Distinct-2 ↑
$ChatGLM3_{lora}$	54.25	**3.6**	**21.8**
$ChatGLM3_{lora}^{sit}$	62.16	3.51	21.56
$ChatGLM3_{lora}^{sit,kw}$	**62.39**	**3.6**	**21.8**

Table 8
Case study of SDAM and benchmarks. The words highlighted in orange are the words that SDAM pays attention to, which often have associations with the situation and context. The words in blue are response words that have strong associations with the context or situation.

Emotion	Grateful
Situation description	Healthy baby after a high risk pregnancy.
Context	[CLS] I had a high risk pregnancy up until a few months ago. I felt the most grateful I would ever been when my baby turned out healthy.
Gold	Congrats, that must have been a very stressful time leading up the birth.
EmpDG	I am so glad you have a lot of people.
KEMP	I am glad you are okay. I am glad you are okay.
CEM	That is so sweet of them.
SDAM	That is great. I am glad you were able to get a baby.
Emotion	Apprehensive
Situation description	Last summer I went skydiving with my partner. I was scared to death!
Context	I went skydiving last summer with my partner. It was so scary!
Gold	Wow, that is scary. Were you nervous?
EmpDG	Did you go check it out?
KEMP	That is great! I am sure you will do great again.
CEM	Oh no! What happened?
SDAM	That is scary! I am so scared of that.
Emotion	Trusting
Situation description	I lost my wallet and trusted to have it turned in.
Context	I lost my wallet last week.
Gold	Did you find it at any point?
EmpDG	Oh no! I am so sorry. I hope you get it back.
KEMP	Oh no! I am so sorry to hear that.
CEM	Oh no! what happened?
SDAM	Oh no! Did you find out who did it?

attention to the key phrases “risk pregnancy” and the important emotion of “grateful”, thereby expressing a relevant and appropriate response.

In the second case, EmpDG generates an unemotional and irrelevant response. KEMP interprets the speaker’s apprehensive and scared emotions as the positive emotion “great”. CEM does not understand the dialogue context, thereby expressing an irrelevant sentence. In contrast, SDAM pays attention to the key event, i.e., “skydiving”, from the situation information and the dialogue context. And it also focuses to the speaker’s emotion of “scared”. Therefore, SDAM understands the key points and emotions of the dialogue, and utilizes the emotional keyword “scared” to express an appropriate response.

In the third case, CEM does not understand the speaker’s meaning. KEMP generates a generic response. The response generated by EmpDG contradicts the background information, since it does not consider the background information in the situation information. In contrast, SDAM introduces and understands the background information, and pays attention to the emotion “trusted” in the situation information. It therefore produces a high-quality response that approaches the gold response.

6. Theoretical and practical significance

We elucidate the significance of our work from both theoretical and practical perspectives.

Theoretical Significance. (1) A novel empathetic generation approach is proposed by introducing situation information and exploring explicit and implicit associations between situations and dialogues, laying the foundation for deeper understanding of emotional and cognitive states in dialogues, and advancing empathetic generation research towards a more advanced stage. (2) A bidirectional filtering encoder and a reasoning knowledge-based hypergraph neural network are designed to capture explicit and implicit associations between situations and dialogues respectively, extending the methods for mining complex semantic associations between texts. (3) A fine-tuning method that combines a small model with large language models is established, enhancing the emotion understanding capability of large models and providing new insights for integrating large models with other modules.

Practical Significance. (1) Experimental results demonstrate significant improvements in emotion recognition accuracy, response fluency, and diversity, validating the effectiveness of introducing situational and associative information for enhancing empathetic generation capability. (2) Through case studies, it is proven that this method can comprehensively and accurately understand dialogues and generate high-quality empathetic responses, showing promising application prospects. (3) This method provides a new technical approach for building more human-like dialogue systems, contributing to the naturalness and friendliness of human–machine interactions.

7. Conclusion

In this paper, we have proposed a Situation-Dialogue Association Model (SDAM) that considered the situation, dialogue, and their explicit and implicit associations to enhance advanced empathetic understanding and expression. SDAM encoded the situation to obtain background information and used a bidirectional filtering encoder and a reasoning knowledge-based hypergraph neural network to capture explicit and implicit associations, respectively. Based on the above information, SDAM employed an adjustable situation-dialogue decoder to generate empathetic responses. Automation and human evaluations have demonstrated that SDAM accurately and comprehensively perceives empathetic information in the dialogues and expresses more empathetic responses. In addition, our method has the following limitations: (1) Although situations widely exist and are important, high-quality situation information is still lacking in some tasks. (2) In real dialogue scenes, multimodality and personalization are important factors for perceiving and expressing empathy. Due to dataset limitations, we did not explore them.

To compensate the limitations above, we will conduct the following future work: (1) we will explore models to generate high-quality situation information. (2) we will explore other factors that promote advanced empathetic understanding and expression, such as personalization and multimodality.

CRediT authorship contribution statement

Zhou Yang: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Zhaochun Ren:** Writing – review & editing, Supervision, Resources. **Yufeng Wang:** Validation, Investigation, Data curation. **Haizhou Sun:** Resources, Investigation. **Xiaofei Zhu:** Supervision. **Xiangwen Liao:** Supervision, Resources, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This research was Supported by National Natural Science Foundation of China (No. 61976054).

References

Bi, G., Shen, L., Cao, Y., Chen, M., Xie, Y., Lin, Z., et al. (2023). DiffusEmp: A diffusion model-based framework with multi-grained control for empathetic response generation. In A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics* (pp. 2812–2831). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.acl-long.158, https://aclanthology.org/2023.acl-long.158.

Buechel, S., & Hahn, U. (2017). EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In M. Lapata, P. Blunsom, & A. Koller (Eds.), *Proceedings of the 15th conference of the European chapter of the association for computational linguistics: volume 2, short papers* (pp. 578–585). Association for Computational Linguistics, http://dx.doi.org/10.48550/arXiv.2205.01996, https://aclanthology.org/E17-2092.

Calvo, R. A., & Mac Kim, S. (2013). Emotions in text: dimensional and categorical models. *Computational Intelligence*, *29*(3), 527–543. http://dx.doi.org/10.1016/j.asoc.2022.108842.

Capel, T., & Brereton, M. (2023). What is human-centered about human-centered AI? A map of the research landscape. *CHI Conference on Human Factors in Computing Systems*, http://dx.doi.org/10.1145/3544548.3580959, https://api.semanticscholar.org/CorpusID:258217892.

Chen, R., Wang, J., Yu, L.-C., & Zhang, X. (2023). Learning to memorize entailment and discourse relations for persona-consistent dialogues. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*, 12653–12661. http://dx.doi.org/10.1609/aaai.v37i11.26489, https://arxiv.org/abs/2301.04871.

Chen, R., Wang, J., & Zhang, X. (2021). Variational autoencoder with interactive attention for affective text generation. In L. Wang, Y. Feng, Y. Hong, & R. He (Eds.), *Proceedings of the natural language processing and Chinese computing* (pp. 111–123). Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-88483-3_9.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., et al. (2022). Scaling instruction-finetuned language models. http://dx.doi.org/10.48550/ARXIV.2210.11416, CoRR abs/2210.11416.

Cuff, B. M., Brown, S. J., Taylor, L., & Howat, D. J. (2016). Empathy: A review of the concept. *Emotion Review*, *8*(2), 144–153. http://dx.doi.org/10.1177/1754073914558466, https://api.semanticscholar.org/CorpusID:147125172.

Curry, A., & Curry, A. C. (2023). Computer says "No": The case against empathetic conversational AI. In A. Rogers, J. L. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics* (pp. 8123–8130). Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/2023.FINDINGS-ACL.515.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, *44*(1), 113. http://dx.doi.org/10.1037/0022-3514.44.1.113.

Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., et al. (2022). GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th annual meeting of the association for computational linguistics* (pp. 320–335). http://dx.doi.org/10.18653/v1/2022.acl-long.26, https://aclanthology.org/2022.acl-long.26.

Elliott, R., Bohart, A. C., Watson, J. C., & Murphy, D. (2018). Therapist empathy and client outcome: An updated meta-analysis. *Psychotherapy*, *55*(4), 399. http://dx.doi.org/10.1037/pst0000175.

Feng, Y., You, H., Zhang, Z., Ji, R., & Gao, Y. (2019). Hypergraph neural networks. *Vol. 33*, In *Proceedings of the AAAI conference on artificial intelligence* (pp. 3558–3565). AAAI Press, http://dx.doi.org/10.48550/arXiv.1809.09401, https://arxiv.org/abs/1809.09401.

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent: a randomized controlled trial. *JMIR Mental Health*, *4*(2), Article e7785. http://dx.doi.org/10.2196/mental.7785.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, *33*(3), 613–619. http://dx.doi.org/10.1186/s12885-023-11325-z.

Ghosh, S., Chollet, M., Laksana, E., Morency, L.-P., & Scherer, S. (2017). *Affect-LM: A neural language model for customizable affective text generation*. Association for computational linguistics, http://dx.doi.org/10.18653/v1/P17-1059, https://api.semanticscholar.org/CorpusID:18999401.

Hoffman, M. L. (1975). Developmental synthesis of affect and cognition and its implications for altruistic motivation. *Developmental Psychology*, *11*(5), 607. http://dx.doi.org/10.1037/0012-1649.11.5.607.

Hoffman, M. L. (1977). Sex differences in empathy and related behaviors. *Psychological Bulletin*, *84*(4), 712. http://dx.doi.org/10.1037/0033-2909.84.4.712.

Hoffman, M. L. (Ed.), (1987). *The contribution of empathy to justice and moral judgment*. Cambridge University Press.

Hoffman, M. L. (Ed.), (2001). *Empathy and moral development: Implications for caring and justice*. Cambridge University Press, http://dx.doi.org/10.1017/CBO9780511805851.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., et al. (2022). LoRA: Low-rank adaptation of large language models. In *The tenth international conference on learning representations*. OpenReview.net, https://openreview.net/forum?id=nZeVKeeFYf9.

Hwang, J. D., Bhagavatula, C., Bras, R. L., Da, J., Sakaguchi, K., Bosselut, A., et al. (2021). (Comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Thirty-third conference on innovative applications of artificial intelligence* (pp. 6384–6392). AAAI Press, http://dx.doi.org/10.1609/AAAI.V35I7.16792.

Kim, H., Kim, B., & Kim, G. (2020). Will I sound like me? Improving persona consistency in dialogues through pragmatic self-consciousness. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 904–916). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.emnlp-main.65, https://aclanthology.org/2020.emnlp-main.65.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Y. Bengio, & Y. LeCun (Eds.), *Proceedings of the international conference on learning representations*. http://arxiv.org/abs/1412.6980.

Lee, L.-H., Li, J.-H., & Yu, L.-C. (2022). Chinese EmoBank: Building valence-arousal resources for dimensional sentiment analysis. *Transactions on Asian and Low-Resource Language Information Processing*, *21*(4), 1–18. http://dx.doi.org/10.1145/3489141.

Li, Q., Chen, H., Ren, Z., Ren, P., Tu, Z., & Chen, Z. (2020). EmpDG: Multi-resolution interactive empathetic dialogue generation. In *Proceedings of the 28th international conference on computational linguistics* (pp. 4454–4466). International Committee on Computational Linguistics, http://dx.doi.org/10.18653/V1/2020.COLING-MAIN.394.

Li, J., Galley, M., Brockett, C., Gao, J., & Dolan, B. (2016). A diversity-promoting objective function for neural conversation models. In K. Knight, A. Nenkova, & O. Rambow (Eds.), *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics* (pp. 110–119). The Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/N16-1014.

Li, Q., Li, P., Ren, Z., Ren, P., & Chen, Z. (2022). Knowledge bridging for empathetic dialogue generation. In *Thirty-fourth conference on innovative applications of artificial intelligence* (pp. 10993–11001). AAAI Press, http://dx.doi.org/10.1609/AAAI.V36I10.21347.

Lin, Z., Madotto, A., Shin, J., Xu, P., & Fung, P. (2019). Moel: Mixture of empathetic listeners. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* (pp. 121–132). Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/D19-1012.

Lynn, V., Son, Y., Kulkarni, V., Balasubramanian, N., & Schwartz, H. A. (2017). Human centered NLP with user-factor adaptation. In M. Palmer, R. Hwa, & S. Riedel (Eds.), *Empirical methods in natural language processing* (pp. 1146–1155). Copenhagen, Denmark: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D17-1119, https://aclanthology.org/D17-1119.

Majumder, N., Hong, P., Peng, S., Lu, J., Ghosal, D., Gelbukh, A. F., et al. (2020). MIME: MIMicking emotions for empathetic response generation. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 8968–8979). Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/2020.EMNLP-MAIN.721.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1532–1543). ACL, http://dx.doi.org/10.3115/V1/D14-1162.

Pickering, M. J., & Garrod, S. (2004). The interactive-alignment model: Developments and refinements. *Behavioral and Brain Sciences*, *27*(2), 212–225. http://dx.doi.org/10.1017/S0140525X04450055.

Rashkin, H., Smith, E. M., Li, M., & Boureau, Y. (2019). Towards empathetic open-domain conversation models: A new benchmark and dataset. In A. Korhonen, D. R. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th conference of the association for computational linguistics* (pp. 5370–5381). Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/P19-1534.

Sabour, S., Zheng, C., & Huang, M. (2022). CEM: commonsense-aware empathetic response generation. In *Thirty-fourth conference on innovative applications of artificial intelligence* (pp. 11229–11237). AAAI Press, http://dx.doi.org/10.1609/AAAI.V36I10.21373.

See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th annual meeting of the association for computational linguistics* (pp. 1073–1083). Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/P17-1099.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Annual conference on neural information processing systems* (pp. 5998–6008). https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Wang, X., Ji, H., Shi, C., Wang, B., Ye, Y., Cui, P., et al. (2019). Heterogeneous graph attention network. In *The world wide web conference* (pp. 2022–2032). https://arxiv.org/pdf/1903.07293.

Wang, L., Li, J., Lin, Z., Meng, F., Yang, C., Wang, W., et al. (2022). Empathetic dialogue generation via sensitive emotion recognition and sensible knowledge selection. In Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), *Findings of the association for computational linguistics* (pp. 4634–4645). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.findings-emnlp.340, https://aclanthology.org/2022.findings-emnlp.340.

Xie, H., Lin, W., Lin, S., Wang, J., & Yu, L.-C. (2021). A multi-dimensional relation model for dimensional sentiment analysis. *Information Sciences*, *579*, 832–844. http://dx.doi.org/10.1016/j.ins.2021.08.052.

Yang, Z., Ren, Z., Yufeng, W., Zhu, X., Chen, Z., Cai, T., et al. (2023). Exploiting emotion-semantic correlations for empathetic response generation. In *Proceedings of the 2023 conference on empirical methods in natural language processing*. https://openreview.net/forum?id=ilCMZV0Qdl.

Yu, J., Tao, D., & Wang, M. (2012). Adaptive hypergraph learning and its application in image classification. *IEEE Transactions on Image Processing*, *21*(7), 3262–3272. http://dx.doi.org/10.1109/TIP.2012.2190083.

Yuan, L., Wang, J., Yu, L.-C., & Zhang, X. (2022). Hierarchical template transformer for fine-grained sentiment controllable generation. *Information Processing & Management*, *59*(5), Article 103048. http://dx.doi.org/10.1016/j.ipm.2022.103048, https://www.sciencedirect.com/science/article/pii/S0306457322001546.

Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., et al. (2023). GLM-130B: an open bilingual pre-trained model. In *The eleventh international conference on learning representations*. OpenReview.net, https://openreview.net/pdf?id=-Aw0rrrPUF.

Zhang, Y., Wang, J., Yu, L., Xu, D., & Zhang, X. (2024). Personalized LoRA for human-centered text understanding. http://dx.doi.org/10.48550/ARXIV.2403.06208, CoRR abs/2403.06208.

Zhao, W., Zhao, Y., Lu, X., & Qin, B. (2023). Don't lose yourself! Empathetic response generation via explicit self-other awareness. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics* (pp. 13331–13344). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.findings-acl.843, https://aclanthology.org/2023.findings-acl.843.

Zhong, P., Wang, D., Li, P., Zhang, C., Wang, H., & Miao, C. (2021). CARE: commonsense-aware emotional response generation with latent concepts. In *Thirty-third conference on innovative applications of artificial intelligence* (pp. 14577–14585). AAAI Press, http://dx.doi.org/10.1609/AAAI.V35I16.17713.

Zhong, P., Zhang, C., Wang, H., Liu, Y., & Miao, C. (2020). Towards persona-based empathetic conversational models. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 6556–6566). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.emnlp-main.531, https://aclanthology.org/2020.emnlp-main.531.

Zhou, J., Zheng, C., Wang, B., Zhang, Z., & Huang, M. (2023). CASE: Aligning coarse-to-fine cognition and affection for empathetic response generation. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics* (pp. 8223–8237). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.acl-long.457, https://aclanthology.org/2023.acl-long.457.