CTSM: Combining Trait and State Emotions for Empathetic Response Model

Yufeng Wang[1,2], **Chao Chen**[3], **Zhou Yang**[1,2], **Shuhui Wang**[1,2], **Xiangwen Liao**[1,2*]
[1]College of Computer and Data Science, Fuzhou University, Fuzhou, China
[2]Digital Fujian Institute of Financial Big Data, Fuzhou, China
[3]School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), China
211027083@fzu.edu.cn, cha01nbox@gmail.com
{200310007, 221027214, liaoxw}@fzu.edu.cn

Abstract

Empathetic response generation endeavors to empower dialogue systems to perceive speakers' emotions and generate empathetic responses accordingly. Psychological research demonstrates that emotion, as an essential factor in empathy, encompasses trait emotions, which are static and context-independent, and state emotions, which are dynamic and context-dependent. However, previous studies treat them in isolation, leading to insufficient emotional perception of the context, and subsequently, less effective empathetic expression. To address this problem, we propose **C**ombining **T**rait and **S**tate emotions for Empathetic Response **M**odel (**CTSM**). Specifically, to sufficiently perceive emotions in dialogue, we first construct and encode trait and state emotion embeddings, and then we further enhance emotional perception capability through an emotion guidance module that guides emotion representation. In addition, we propose a cross-contrastive learning decoder to enhance the model's empathetic expression capability by aligning trait and state emotions between generated responses and contexts. Both automatic and manual evaluation results demonstrate that CTSM outperforms state-of-the-art baselines and can generate more empathetic responses. Our code is available at `https://github.com/wangyufeng-empty/CTSM`

Keywords: empathetic response generation, dialogue system, emotion recognition, contrastive learning

1. Introduction

Empathy is crucial in human conversation (Abu-Elrob, 2022) and human-like dialogue systems (Beredo and Ong, 2022; Zhao et al., 2023a). Central to this work is the empathetic response generation task (Rashkin et al., 2019), which aims to produce empathetic responses by profoundly comprehending speakers' emotions (Zhao et al., 2023a). Emotion, as an essential factor facilitating empathy (Lebowitz and Dovidio, 2015; Zaki, 2020; Krol and Bartz, 2022), bridges communicators and fosters understanding (Decety and Holvoet, 2021). Psychological studies (Rosenberg, 1998) differentiate between trait (static and context-independent (Goetz et al., 2015)) and state (dynamic and context-dependent (Goetz et al., 2015; Zheng et al., 2023)) emotions. Specifically, we regard *static* as the inherent emotional connotation of textual words, whereas *dynamic* corresponds to the emotion's variability and adaptability to contexts. Figure 1 illustrates the distinction between trait and state emotions. The bar chart highlights that the trait emotion of *excited* consistently embodies the fundamental emotional dimensions of Valence, Arousal, and Dominance (Mohammad, 2018) in context A and B. Trait emotions can be accurately quantified and remain context-independent, and overlooking them may miss inherent emotional connotations of words, weakening emotion understanding. Conversely, the heat map presents the diverse emotional expressions of *excited* across various contexts. In context A, it conveys positive feelings like *happiness* and *anticipation*, while in context B, *excited* tends towards negative emotions like *terrified* and *anxiety*. Ignoring state emotions could confuse semantics and emotional interpretation.

Existing approaches focus separately on perceiving only one type of emotion. Research targeting trait emotions often utilizes pre-trained classifiers (Rashkin et al., 2019) and external knowledge (Li et al., 2022; Sabour et al., 2022; Zhou et al., 2023; Zhao et al., 2023b). Conversely, approaches centered on state emotions employ multi-listener frameworks (Lin et al., 2019), emotion mimicry techniques (Majumder et al., 2020), and embedding adjustments (Agrawal et al., 2018; Mao et al., 2019; Wang and Meng, 2018). However, treating trait and state emotions in isolation comprised the completeness and intricacy of emotional expression in dialogue. Neglecting the emotional reaction (Elliott et al., 2018) stemming from the interaction between trait and state emotions can engender inaccurate emotion comprehension and categorization, generating inappropriate empathetic responses. Therefore, modeling both emotion types is imperative for empathetic response generation, but remains underexplored.

To this end, we propose a **C**ombining **T**rait and **S**tate Emotions for Empathetic Response **M**odel

*Corresponding author.

LREC-COLING 2024, pages 4214–4225

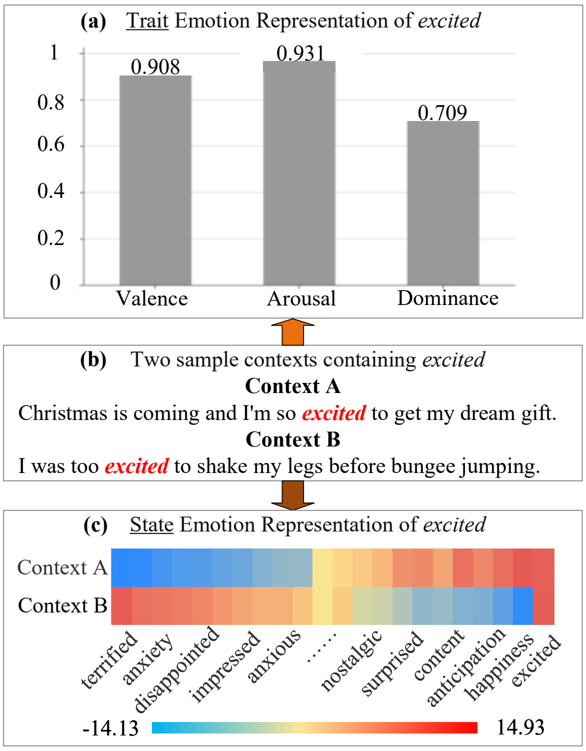

Figure 1: An example of trait and state emotions. (a) The bar chart illustrates the context-independent trait emotions of *excited*. (b) Two sample contexts contain the term *excited*. (c) The heat map displays the varying state emotions of *excited* across two different contexts, with warmer colors indicating stronger emotion inclination and cooler colors denoting lesser emotional intensity.

(CTSM) to fully incorporate trait and state emotions, enabling more comprehensive perception and expression of contextual emotions. First, in addressing the distinction between static trait emotions and dynamic state emotions, we present two specialized embedding patterns to capture their unique characteristics at the token level. Building on this foundation, we introduce emotion encoders as the essential component to extract and refine the nuanced feature representations inherent in these embeddings. Subsequently, to enhance emotion perception capability, we propose an emotion guidance module with teacher and student components. The teacher guides the student through emotion labels to enhance the student's understanding of complex emotions. Additionally, we design a cross-contrastive learning decoder to enhance CTSM's empathetic expression capability that aligns the features of generated responses and contexts in terms of trait and state emotions.

Experiments with benchmark models on the EMPATHETICDIALOGUES (ED) dataset (Rashkin et al., 2019) demonstrate that CTSM outperforms benchmark models, particularly excelling in emotion accuracy and diversity metrics. Further studies verify that CTSM can accurately perceive both trait and state emotions.

Our contributions are summarized as follows:

- To the best of our knowledge, our work is the first to simultaneously model both trait and state emotions for *each* token within the dialogue text. This addresses the limitations in emotion perception methodologies of prior works.
- We augment the interplay between trait and state emotions utilizing an emotion guidance module, which improves the perception of intricate emotions through full emotion feature guidance. Furthermore, we employ a cross-contrastive learning decoder to enhance empathetic expression during empathetic response generation.
- The experimental results demonstrate that CTSM effectively combines trait and state emotions within dialogues, exhibiting enhanced empathetic capabilities.

2. Related Work

Empathetic response generation involves perceiving emotions conveyed by speakers in a dialogue to produce sympathetic responses (Rashkin et al., 2019). Early approaches directly simulated emotions using rules and statistics (Colby et al., 1972; Keshtkar and Inkpen, 2011; Adamopoulou and Moussiades, 2020), but faced challenges in scalability, flexibility, and cost. Recent methods, leveraging the power of deep neural networks (Mikolov et al., 2013; Pennington et al., 2014; Ma et al., 2020) and word embeddings (Jianqiang et al., 2018; Baali and Ghneim, 2019; Hamdi et al., 2019), have made significant strides in perceiving emotions in conversational text. Building on these advancements, state-of-the-art approaches can be broadly categorized into two groups: those targeting the perception of trait emotions and those focusing on state emotions within dialogue contexts.

The first category of methods emphasizes the perception of trait emotions to enhance emotional accuracy. Specifically, Rashkin et al. (2019) leveraged a pre-trained emotion classifier to capture trait emotions in context, and Li et al. (2020) adopted interactive discriminators to extract multi-resolution trait emotions. Further considering specific words like negation combined with word intensity (Zhong et al., 2019) and words causing emotional causality (Kim et al., 2021) can strengthen the model's perception of subtle trait emotions. However, the lack of external knowledge makes it challenging for models to perceive implicit emotions. Li et al. (2022) addressed this limitation by utilizing external knowledge to construct an emotion context graph, enhancing the expression of implicit trait emotions in the semantic space. Sabour et al. (2022) built COMET through external reasoning knowledge, reinforcing the perception of trait emotions. Building on this, Zhou et al. (2023) and Zhao et al. (2023b) integrate external knowledge through graph structure to enhance the model's empathetic capabilities. However, these models focus on perceiving static, context-independent trait emotions while neglecting dynamic state emotions, leading to a misalignment between contextual semantics and the emotions conveyed in the text.

The second type of method focuses on perceiving state emotions to enrich emotion understanding, such as modeling mixed emotions using multiple listeners (Lin et al., 2019) or mimicking user emotions considering emotion polarity and randomness (Majumder et al., 2020). However, directly modeling global context can cause semantically similar words to convey opposing emotions (Agrawal et al., 2018). Thus, some methods address this by constructing contextual word embeddings that capture emotional influences on individual words (Agrawal et al., 2018; Mao et al., 2019; Wang and Meng, 2018; Yang et al., 2023), enabling richer state emotion perception. However, these approaches overlook the inherent static trait emotions within the dialogue, leading to inaccurate discernment of contextual emotions and generating inappropriate empathetic responses.

3. Method

3.1. Overview

Figure 2 illustrates the overall architecture of CTSM. To effectively perceive and utilize the trait and state emotions in dialogues, we abstract the model into four primary components: **1) Inference-Enriched Context Encoder** encodes the context and integrates inference knowledge; **2) Emotion Encoding Module** constructs and encodes two emotion embeddings, enabling the model to perceive both trait and state emotions from context fully; **3) Emotion Guidance Module** facilitates CTSM's learning of emotion representations by utilizing the inference-enriched context to enhance its emotion perception capabilities; **4) Cross-Contrastive Learning Decoder** employs cross-contrastive learning after decoding process during training, allowing CTSM to generate more empathetic and appropriate responses.

3.2. Task Formulation

Given a dialogue history $D = [u_1, u_2, \ldots, u_n]$ with a context-level emotion label ε, as well as a set of emotion words e, our goal is to generate empathetic responses $R = [r_1, r_2, \cdots, r_m]$ whose semantics and emotions align with the context while conveying empathy. e is the union of all emotion labels and $\varepsilon \in e$. D is made up of n sentences, and R contains m words. The i-th sentence $u_i = \left[x_1^i, x_2^i, \cdots, x_{l_i}^i\right]$ consists of l_i words. For batches, sequences shorter than the maximum length L are padded to L with $[PAD]$ tokens.

3.3. Inference-Enriched Context Encoder

Following prior works (Li et al., 2022; Sabour et al., 2022), we flatten the dialogue context, and prepend a $[CLS]$ token to obtain the context sequence $C = [CLS] \oplus u_1 \oplus u_2 \oplus \cdots \oplus u_n$, where $\oplus$ represents concatenation. The context embedding $E_C \in \mathbb{R}^{L \times d}$ is the sum of the word embeddings, positional embeddings, and dialogue state embeddings. An encoder is used to extract the contextual hidden representation from E_C:

$$H_C = \mathbf{Encoder}_C\left(E_C\right), \tag{1}$$

where $H_C \in \mathbb{R}^{L \times d}$ and d is the dimension of the context encoding.

Referring to the approach of Sabour et al. (2022) for the fusion of context and inferential knowledge, we establish the inference-enriched context teacher H_C^{tchr} and similarly the student H_C^{stu}. These two contexts will be used in the emotion guidance module (in Sec. 3.5) and cross-contrastive learning decoder (in Sec. 3.6).

3.4. Emotion Encoding Module

In this subsection, we illustrate how to perceive trait and state emotions by considering their unique characteristics. We also present how we integrate and encode them with external knowledge and the importance of words.

3.4.1. Trait Emotions Encoding

To ensure that the trait emotion embedding V_t effectively captures the context-independent emotion, importance of words, and contextual semantics. V_t integrates static emotion knowledge from the VAD emotion lexicon (Mohammad, 2018) V_{VAD}, Inverse Document Frequency (IDF) (Sparck Jones, 1988) V_{IDF} and condensed contextual semantics $\widetilde{H}_C$. Formally,

$$V_t = V_{VAD} \oplus V_{IDF} \oplus \widetilde{H}_C, \tag{2}$$

$$\widetilde{H}_C = W_C H_C, \tag{3}$$

where $V_{VAD} \in \mathbb{R}^{L \times 3}, V_{IDF} \in \mathbb{R}^{L \times 1}, \tilde{H}_C \in \mathbb{R}^{L \times d_{cs}}$ and d_{cs} is the dimension after semantic compression. The VAD lexicon delineates emotions into three dimensions: Valence (negativity or positivity),

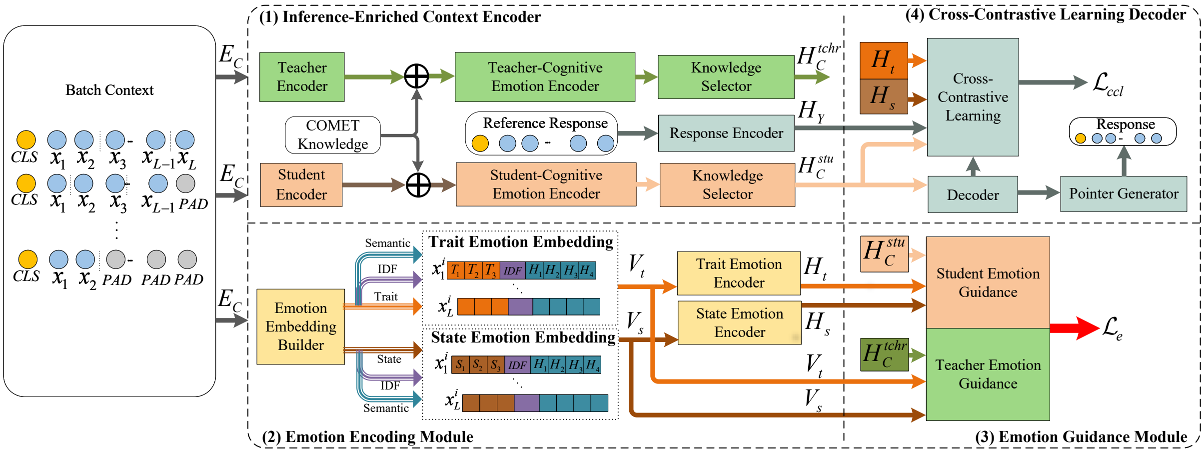

Figure 2: An overview architecture of CTSM. It consists of four parts: **1)** Inference-Enriched Context Encoder; **2)** Emotion Encoding Module; **3)** Emotion Guidance Module; **4)** Cross-Contrastive Learning Decoder.

Arousal (calmness or excitement), and Dominance (weak or strong control), with each value ranging from [0,1]. The words excluded in VAD are set with neutral default values [0.00, 0.50, 0.00]. Then, $V_t \in \mathbb{R}^{L \times d_t}$ is encoded to encapsulate the trait emotions representation within the context:

$$H_t = \mathbf{Encoder}_t\left(V_t\right), \tag{4}$$

where $\mathbf{Encoder}_t\left(\cdot\right)$ is the trait emotion encoder and $H_t \in \mathbb{R}^{L \times d_t}$.

3.4.2. State Emotion Encoding

Regarding the state emotion embedding V_s, we initially define the dynamic state inclination $V_{\cos}$, whose entries with higher values indicate a stronger emotional inclination. Specifically, $V_{\cos}$ is the cosine similarity between the linear embeddings of emotion words $\tilde{E}_e$ and context $\tilde{E}_C$.

$$\widetilde{E}_e = W_1 \times \mathbf{Embedding}(e) + b_1, \tag{5}$$

$$\widetilde{E}_C = W_2 E_C + b_2, \tag{6}$$

$$V_{\cos} = \mathbf{cos}\left(\widetilde{E}_C, \widetilde{E}_e^{\top}\right), \tag{7}$$

where $\widetilde{E}_e \in \mathbb{R}^{32 \times d}$, $\widetilde{E}_C \in \mathbb{R}^{L \times d}$ and $V_{cos} \in \mathbb{R}^{L \times 32}$. The 32 dimensions in $\widetilde{E}_e$ correspond to the 32 emotion categories in the ED dataset, used for prediction classification. Notably, these emotion labels are *not* fed into the emotion encoder. $\mathbf{cos}(\cdot)$ is the cosine similarity function. W_1, W_2, b_1, b_2 are all trainable parameters. Next, V_s consolidates the state inclination, IDF vector, and semantics $\widetilde{H}_C$ to understand the token's state emotions comprehensively.

$$V_s = V_{\cos} \oplus V_{IDF} \oplus \widetilde{H}_C, \tag{8}$$

where $V_s \in \mathbb{R}^{L \times d_s}$, d_s is the dimensionality of the state emotion embedding. Finally, we encode V_s to capture the state emotion representation:

$$H_s = \mathbf{Encoder}_s\left(V_s\right), \tag{9}$$

where $\mathbf{Encoder}_s\left(\cdot\right)$ is the state emotion encoder and $H_s \in \mathbb{R}^{L \times d_s}$.

3.5. Emotion Guidance Module

In light of the intricacies of textual emotions in dialogues, a hard label may not encompass the blend of trait and state emotions. Drawing from research (Xu et al., 2020; Jafari et al., 2021; Nguyen et al., 2022), we design an emotion guidance module that enhances the model's capacity to perceive and comprehend intricate emotions. Specifically, we employ a teacher component to extract soft labels encapsulating the *dark knowledge* (Hinton et al., 2015), which comprises hidden knowledge crossing both trait and state emotions. The student model assimilates the dark knowledge by training with soft labels, leading to enhanced generalization and performance (Hahn and Choi, 2019). Armed with augmented capabilities, the student is employed in both the emotion prediction (in Sec. 3.5.2) and decoding phases (in Sec. 3.6).

3.5.1. Teacher Emotion Guidance

After getting the embedding of trait and state emotions as well as the teacher inference-enriched context, we concatenate them to V_{tchr}. Subsequently, the teacher's semantic-emotion context C_{tchr} is derived by weighting V_{tchr} with the emotion intensities I (Li et al., 2022).

$$V_{tchr} = H_C^{tchr} \oplus V_t \oplus V_s, \tag{10}$$

$$C_{tchr} = \mathbf{SUM}_{d=1}\left(\boldsymbol{\sigma}\left(I\right) \times V_{tchr}\right), \tag{11}$$

where $V_{tchr} \in \mathbb{R}^{L \times d_s}$, $C_{tchr} \in \mathbb{R}^{L \times d_s}$, $H_C^{tchr} \in \mathbb{R}^{L \times d}$. $\mathbf{SUM}_{d=1}(\cdot)$ represents a summation over the first dimension to aggregate contextual semantics representations. $\boldsymbol{\sigma}(\cdot)$ is the softmax function. The emotion distribution P_e^{tchr} predicted by the teacher, which captures the characteristics of both trait and state emotions, is given by:

$$S = \boldsymbol{\sigma}\left(W_3^s\left(\mathbf{Tanh}\left(W_3^c C_{tchr} + b_3\right)\right)\right), \tag{12}$$

$$S' = \mathbf{Tanh}\left(W_4\left(C_{tchr} S\right) + b_4\right), \tag{13}$$

$$P_e^{tchr} = \boldsymbol{\sigma}\left(W_{out} S' + b_{out}\right), \tag{14}$$

where $S \in \mathbb{R}^{L \times d}$, $S' \in \mathbb{R}^{L \times d_s}$, $P_e^{tchr} \in \mathbb{R}^{32}$. W_3^s, W_3^c, W_4, b_3, b_4, b_{out}, and W_{out} are all trainable parameters. The teacher's parameters are optimized by the Cross-Entropy Loss between teacher emotion prediction P_e^{tchr} and the ground truth label e^*.

$$\mathcal{L}_{tchr} = -\log\left(P_e^{tchr}\left(e^*\right)\right). \tag{15}$$

3.5.2. Student Emotion Guidance

The student shares a similar model structure with the teacher. However, the student concatenates the student inference-enriched context H_C^{stu} with emotion representations H_t and H_s, rather than V_t and V_s. Specifically:

$$V_{stu} = H_C^{stu} \oplus H_t \oplus H_s, \tag{16}$$

$$C_{stu} = \mathbf{SUM_{d=1}}\left(\boldsymbol{\sigma}\left(I\right) \times V_{stu}\right), \tag{17}$$

where $V_{stu} \in \mathbb{R}^{L \times d_s}$. $C_{stu} \in \mathbb{R}^{L \times d_s}$ is the student semantic-emotion context. Analogous to how P_e^{tchr} is computed by Eqs. (12) - (14), we obtain the student's emotion prediction $P_e^{stu} \in \mathbb{R}^{32}$ and employ it for predicting dialogue emotion, formally represented by the equation: $\hat{e} = \mathbf{argmax}\left(P_e^{stu}\right)$. To learn hidden knowledge from the teacher, the student is trained with soft labels:

$$\mathcal{L}_{stu} = -\log\left(P_e^{stu}\left(P_e^{tchr}\right)\right). \tag{18}$$

Ultimately, the objective of the teacher and student components concerning the emotion perception is:

$$\mathcal{L}_e = \mathcal{L}_{tchr} + \mathcal{L}_{stu}. \tag{19}$$

3.6. Cross-Contrastive Learning Decoder

Contrastive learning (Chen et al., 2020; Sun et al., 2023) minimizes distances between positive samples while maximizing distances between negative samples. Inspired by its application in feature alignment (Zhou et al., 2022), we incorporate cross-contrastive learning into the decoding process. Specifically, we align dialogue emotions between responses and contexts by minimizing the distance among generated responses, target responses, contextual semantics, as well as trait and state emotions. Consequently, this enhances the student's contextual semantic representation and teacher-guided performance, further improving the model's ability for empathetic expression.

3.6.1. Response Generation

As mentioned in Sec. 3.3, the student inference-enriched context H_C^{stu} is used to make prediction for word distribution P_w:

$$\begin{aligned} P_w &= P\left(R_j \mid E_{R<j}, C, H_t, H_s\right) \\ &= \mathbf{PoGen}\left(\mathbf{Decoder}\left(E_Y, H_C^{stu}\right)\right), \end{aligned} \tag{20}$$

where $E_{R<j}$ is the embedding of the generated responses up to time step $j-1$. $E_Y = \mathbf{Embedding}(Y)$ is the embedding of the target response Y, and $\mathbf{PoGen}(\cdot)$ signifies the pointer generator network module (See et al., 2017).

Ultimately, the generation loss of the model is defined by a standard negative log-likelihood:

$$\mathcal{L}_g = -\sum_{j=1}^{T} \log P\left(R_j \mid E_{R<j}, C, H_t, H_s\right), \tag{21}$$

3.6.2. Cross-Contrastive Learning

We adopt Contrastive Learning to align dialogue emotion representations between responses and contexts for the same context within a batch. Besides aligning representations of contextual semantics H_C^{stu} and generated responses P_w, we are also interested in the hidden representation of target response $H_Y = \mathbf{Encoder}_Y(E_Y)$, and the combined representation of trait and state emotions $H_{ts} = H_t \oplus H_s$. Then, by *crossly* pairing these representations with each other, the set of positive sample pairs denoted as $\mathcal{H}^+$, consists of the five sample pairs for each context. To be specific:

$$\begin{aligned} \mathcal{H}^+ = \{&(H_Y, H_{ts}), (H_C^{stu}, P_w), (H_C^{stu}, H_Y), \\ &(H_{ts}, P_w), (H_Y, P_w)\}, \end{aligned} \tag{22}$$

Notably, considering the tight correlations between emotions and semantics, our model not only aligns them within the current context but also captures the emotion correlations across various contexts in the batch. To avoid potential misalignment of emotions with semantics from other contexts, we exclude the pair (H_C^{stu}, H_{ts}). Conversely, the representations of the different contexts are regarded as negative pairs, and the set of negative sample pairs $\mathcal{H}^-$ contains all such negative pairs.

The training objective for any given positive sample pair is to minimize the distance between their representations while maximizing the distance for negative pairs, expressed as:

$$\mathcal{L}_{cl}(h_p, h_q) = -\log \frac{e^{\mathbf{sim}(h_p, h_q)/\tau}}{\sum_{(h_p, h_k)} e^{\mathbf{sim}(h_p, h_k)/\tau}}, \tag{23}$$

where $(h_p, h_q) \in \mathcal{H}^+$, $(h_p, h_k) \in \mathcal{H}^-$. $\mathbf{sim}(\cdot)$ computes similarity using the dot product. τ is a temperature parameter adjusting the scale of similarity scores. The loss $\mathcal{L}_{ccl}$ is then computed as the average loss across the five positive sample pairs:

$$\mathcal{L}_{ccl} = \frac{1}{5} \sum_{(h_p, h_q)} \mathcal{L}_{cl}(h_p, h_q). \tag{24}$$

Integrating the diversity loss $\mathcal{L}_{div}$ suggested by Sabour et al. (2022), the total loss of our model is the weighted sum of the four mentioned losses:

$$\mathcal{L} = \gamma_1 \mathcal{L}_e + \gamma_2 \mathcal{L}_g + \gamma_3 \mathcal{L}_{ccl} + \gamma_4 \mathcal{L}_{div}, \tag{25}$$

where γ_1, γ_2, γ_3, and γ_4 are hyperparameters that can be manually set.

4. Experimental Settings

4.1. Baselines for Comparison

We compare the proposed model with six state-of-the-art (SOTA) benchmark models:

- **Transformer** (Vaswani et al., 2017): is a vanilla Transformer-based model with encoder-decoder architecture for generation.
- **MoEL** (Lin et al., 2019): is a Transformer-based empathetic response generation model using separate emotion decoders and a global contextual decoder to combine emotions softly.
- **MIME** (Majumder et al., 2020): is a Transformer-based model that considers emotion clustering based on polarity and emotion mimicry to generate empathetic responses.
- **EmpDG** (Li et al., 2020): combines dialog-level and token-level emotions through a multi-resolution adversarial model with multi-granularity emotion modeling and user feedback.
- **KEMP** (Li et al., 2022): uses ConceptNet (Speer et al., 2017) to construct an emotion context graph, capturing implicit emotions to enrich representations for appropriate response generation.
- **CEM** (Sabour et al., 2022): incorporates affection and cognition, and uses reasoning knowledge about the user's situation to enhance its ability to perceive and express emotions.
- **CASE** (Zhou et al., 2023): introduces commonsense reasoning and emotional concepts, aligning with the user's cognition and emotions at both coarse-grained and fine-grained levels to generate empathetic responses rich in information.

4.2. Implementation Details

We conduct experiments on EMPATHETICDIALOGUES dataset, using the 8:1:1 train/validation/test split as in (Rashkin et al., 2019). CTSM uses 300-dimensional pre-trained GloVe vectors (Pennington et al., 2014) for embedding. The dynamic state inclination is 32-dimensional, VAD vectors are 3-dimensional, and the compressed semantic dimensions d_{cs} are 10-dimensional. We set the cross-contrastive learning temperature τ to 0.07 and loss weights γ_1 to γ_4 as 1, 1, 1, and 1.5, respectively. When trained on a Tesla T4 GPU with a batch size of 16, our model utilizes the Adam optimizer (Kinga et al., 2015) combined with the NoamOpt (Vaswani et al., 2017) learning rate schedule. The model converges after roughly 17,250 iterations.

4.3. Evaluation Metrics

4.3.1. Automatic Evaluations

We adopt four automated metrics for evaluation: Emotion Accuracy (**Acc**), Perplexity (**PPL**) (Serban et al., 2015) and Distinct metrics (**Dist-1** and **Dist-2**) (Li et al., 2016). Lower perplexity indicates a higher quality of the generated responses. Higher emotion accuracy indicates more precise contextual emotion perception. Larger distinct shows a greater diversity of responses.

4.3.2. Human Evaluations

For human evaluation, we employ A/B testing between model response pairs (Li et al., 2020, 2022; Sabour et al., 2022) concerning Empathy (**Emp.**), Relevance (**Rel.**), and Fluency (**Flu.**). Empathy measures the emotional alignment between responses and contexts. Relevance assesses the coherence of generated responses with contexts. Fluency assesses readability and grammar. Three professional annotators compare responses from the CTSM against those from the baselines. Rather than using absolute 1-5 scales, which can be prone to subjective differences (Sabour et al., 2022), annotators label CTSM responses as Win, Tie, or Lose relative to the baseline for the same context.

5. Results and Analysis

5.1. Automatic Evaluation Results

Table 1 shows the performance of CTSM and baselines concerning automatic metrics. We find that models such as MoEL and MIME focus primarily on recognizing state emotions and exploring the relationship between context and various emotion categories. However, they tend to overlook trait emotions. This oversight decreases emotion detection accuracy (Acc) and compromises response quality (PPL). On the other hand, models like EmpDG, KEMP, CEM and CASE, which only focus on trait emotions, detect emotions more precisely but do not substantially improve response diversity (Dist-1 and Dist-2) or quality (PPL).

Models	Acc(%) ↑	PPL ↓	Dist-1 ↑	Dist-2 ↑
Transformer	-	37.73	0.47	2.04
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	40.20	35.37	0.74	4.01
CTSM	**43.41**	**34.56**	**2.00**	**7.34**

Table 1: Comparison of CTSM against baseline models on automatic evaluation metrics. The best results are bolded.

Overall, CTSM outperforms all other baselines concerning all the automatic evaluation metrics. Specifically, CTSM achieves a **7.99%** relatively higher accuracy than CASE. We attribute this to our model's exceptional capability to combine trait and state emotions, perceiving a more comprehensive range of emotions. Subsequently, through the emotion guidance module, the teacher guides the student to learn features from soft labels that cover both trait and state emotions. This process enhances the emotion encoders' capacity and generates deeper emotional representations. Furthermore, CTSM achieves a relative improvement over CASE in Dist-1, Dist-2, and PPL by **170.27%**, **83.04%**, and **2.29%** respectively. We attribute the performance improvement to the enhanced integration of the student context and inference knowledge through the emotion guidance module, which enriches response diversity. Furthermore, cross-contrastive learning reduces the distance among generated responses, target responses, and contextual representations regarding both trait and state emotions. This module significantly strengthens the representation of contextual semantics and emotions, enhancing the model's response quality and capacity for empathetic expression.

5.2. Human Evaluation Results

Based on the results shown in Table 1, we select three competitive models as benchmarks for human evaluation. As shown in Table 2, CTSM achieves state-of-the-art performance on the human evaluation metrics of Empathy, Relevance, and Fluency compared to others. The high scores in Empathy underscore CTSM's proficiency in perceiving both trait and state emotions within contextual dialogues and in expressing appropriate emotions in the generated responses. Additionally, high scores in Relevance attest to CTSM's capability to comprehend contextual semantics effectively, generate topically coherent responses, and extract and convey pertinent information from diverse contexts. Finally, the outstanding Fluency highlights CTSM's superior decoding capabilities to generate more natural, human-like responses.

Comparison	Aspects	Win	Lose	κ
CTSM vs. KEMP	Emp.	**48.6**	8.1	0.64
	Rel.	**52.6**	13.3	0.60
	Flu.	**35.1**	14.1	0.56
CTSM vs. CEM	Emp.	**38.5**	10.9	0.62
	Rel.	**47.9**	17.0	0.68
	Flu.	**28.9**	15.8	0.44
CTSM vs. CASE	Emp.	**36.8**	14.6	0.57
	Rel.	**49.1**	16.5	0.53
	Flu.	**28.1**	13.8	0.48

Table 2: CTSM's human A/B evaluation results(%). The best results are bolded. κ is the label consistency measured by Fleiss' kappa (Fleiss and Cohen, 1973), with 0.41 $\leq \kappa \leq$ 0.60 and 0.61 $\leq \kappa \leq$ 0.80 indicating moderate and substantial agreement respectively.

Models	Acc(%)	PPL	Dist-1	Dist-2
CTSM	**43.41**	34.56	**2.00**	**7.34**
w/o TEE	42.58	34.68	1.18	4.28
w/o SEE	42.63	35.01	1.42	5.06
w/o EGM	39.66	35.08	1.72	6.50
w/o CCL	42.97	**34.42**	1.73	6.27
w/o EGM & CCL	41.88	36.35	1.57	5.42

Table 3: Ablation studies results of CTSM. The best results are bolded.

5.3. Ablation Studies

We design four variants for ablation studies to verify the effectiveness of the key components in our model: **1) w/o TEE**: Without the **t**rait **e**motion embedding and corresponding emotion **e**ncoder. **2) w/o SEE**: Without the **s**tate **e**motion embedding and corresponding emotion **e**ncoder. **3) w/o EGM**: Without the **e**motional **g**uidance **m**odule, which contains teacher and student emotion guidance. **4) w/o CCL**: Without the **c**ross-**c**ontrastive **l**earning component in the decoding and generating process, and removing the contrastive loss. **5) w/o EGM & CCL**: Simultaneously eliminating the aforementioned EGM and CCL components.

The results are shown in Table 3. Omitting **TEE** from CTSM leads to a notable decline in Acc and Dist, indicating TEE enhances emotion perception accuracy and quality of response. On the other hand, the exclusion of **SEE** results in lower emotion accuracy and response quality, highlighting SEE's role in improving alignment between contextual semantics and conveyed emotions, thereby enhancing the model's comprehensive emotion perception and semantic understanding. Moreover, the superior Acc in the w/o **EGM & CCL** relative to CASE suggests that capturing trait and state emotions alone enhances emotional perception capabilities.

Furthermore, detaching **EGM** from CTSM significantly reduces all metrics, especially Acc and PPL. It emphasizes that EGM markedly enhances emotion encoders' efficiency and the inference-enriched context's semantic representation.

Finally, CTSM without **CCL** achieves worse accuracy and diversity. The results demonstrate that CCL enhances the model's ability to fully perceive comprehensive contextual emotions and produce diverse responses, ultimately optimizing empathetic expression. However, its PPL value is better than CTSM. We attribute it to two factors: Firstly, the negative samples contain noise introduced by the dataset, which could be amplified during the feature alignment (Zhou et al., 2022). Secondly, a small batch size and relatively short dialogue sentences in the dataset may cause the model to overfit during training (Wang et al., 2022), generating overly simplistic responses and subsequently deteriorating the PPL.

5.4. Deeper Analysis on Trait and State Emotions

In this subsection, we delve deeper into the emotional polarities of trait and state, emphasizing the importance of considering them concurrently by showing the discrepancy between them.

Specifically, the word's *trait* emotion polarity $\mathcal{P}_t$ is determined by the Valence in the VAD lexicon. Specifically, a word exhibits a negative trait polarity $\mathcal{P}_t = 0$ when $0 \leq$ Valence ≤ 0.5, and a positive polarity $\mathcal{P}_t = 1$ when $0.5 <$ Valence ≤ 1. Then, we divide all words into positive and negative groups based on their $\mathcal{P}_t$. For the *state* emotion polarity $\mathcal{P}_s$, we utilize the GloVe word embeddings, which encapsulate rich contextual semantics. We calculate the centroid of each of the two groups and the cosine similarity (as in Sec. 3.4.2) between each word and the two centroids. A word is deemed to have a negative state polarity $\mathcal{P}_s = 0$ when it exhibits a more substantial similarity to the negative centroid than to the positive one, and vice versa.

We visualize the trait and state emotion polarities of 32 emotion words in the VAD three-dimensional space, as illustrated in Figure 3. Notably, while *surprised* manifests positive trait emotions, its state emotions inclination towards the negative in contextual semantics (whose $\mathcal{P}_t \neq \mathcal{P}_s$), results in emotion and semantic divergence. Further analysis of 23,712 words from the ED dataset reveals a 48.75% proportion (11,559 words) exhibiting such discrepancies. Therefore, combining the word's trait and state emotions and analyzing their interplay is necessary to diminish the model's emotion and semantic discrepancies.

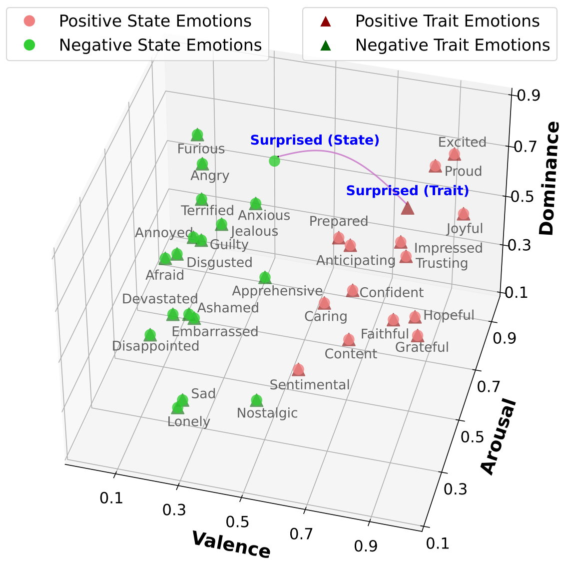

Figure 3: Visualization of trait and state emotion polarities in the VAD space. When a word's trait and state emotional polarities align, they overlap; otherwise, an offset occurs (highlighted in blue).

Emotion	Nostalgic
Context	Last week when Toys"R"Us closed, it really made me start thinking of the 90s and my childhood.
KEMP **CEM** **CASE**	I am sure you will do great again. I bet you were happy for you. What kind of memories?
CTSM	That is awesome. I am sure it was a good thing to have a good time to go back to the 90s and the best of us.
Golden	I have heard a lot of people say something similar. Did you go there a lot?
Emotion	Impressed
Context	My 6 year old tried to play a video game but could not understand how the controller worked. A few days later she was playing no problem!
KEMP **CEM** **CASE**	I am sure you will do great! That is so sweet of her! Oh no! Did she have a good time?
CTSM	That is awesome! I love video games too!
Golden	Kids pick those things up quickly. And it'll help with her hand-eye coordination, reading - all sorts of things!

Table 4: Comparison of responses generated by CTSM and three baselines.

5.5. Case Study

In Table 4, two case analyses compare CTSM against three prominent baselines: KEMP, CEM and CASE. In the first sample, these baselines fail to integrate the semantics of words such as *closed* and *childhood* with both trait and state emotions, resulting in irrelevant and incoherent responses. In contrast, CTSM effectively merges *nostalgic* trait emotions and state emotions with semantics, inferring that the speaker is reminiscing about the *1990s era*. In the second sample, the baselines either convey inaccurate emotions (as with CASE) or express weak feelings (as seen with KEMP and CEM). These shortcomings are coupled with a limited grasp of semantics, resulting in generic responses. Given words like *video game* and *not understand* carry negative trait emotions, but *tried* and *no problem* express positive state emotions, CTSM can combine emotions with semantics to understand the feelings conveyed by *impressed*.

6. Conclusions and Future Work

In this paper, we propose CTSM that combines trait and state emotions for comprehensive dialogue emotion perception. By encoding trait and state emotion embeddings, CTSM captures dialogue emotions fully. Then, the emotion guidance module further augments emotion perception capability. Lastly, the cross-contrastive learning module enhances the model's empathetic expression capability. The automatic and human evaluation results validate the efficacy of CTSM on the empathetic response generation task. In the future, we will emphasize refining the trait emotion embedding and exploring more methods for state inclination.

7. Limitations

Our work primarily has two limitations as follows:

Firstly, we employ a cross-contrastive learning module in CTSM. This process constructs multiple positive sample pairs by cross combining features for contrastive learning. However, it might overlook other strongly correlated positive feature pairs.

The second limitation is the inconsistency between automatic evaluation metrics and human evaluation scores (Liu et al., 2016). Automated metrics struggle to assess the degree of empathy in responses, and solely relying on existing metrics makes generating empathetic dialogues challenging.

8. Ethical Considerations

The data we use is sourced from EMPATHETICDIALOGUES (Rashkin et al., 2019), an open-source dataset that does not contain any personal privacy information. Our human evaluations are conducted by three professional annotators, ensuring no involvement of personal privacy, with reasonable wages paid.

9. Acknowledgements

We are grateful for the support and assistance from the backers of this project and the related laboratories. Yufeng Wang, Zhou Yang, Shuhui Wang, and Xiangwen Liao were supported by the Natural Science Foundation of China (61976054).

Bibliographical References

Rula Ahmad Mahmoud Abu-Elrob. 2022. The role of empathy in jordanian medical encounters. *Health communication*, 37(14):1850–1859. Epub 2022 Oct 2.

Eleni Adamopoulou and Lefteris Moussiades. 2020. Chatbots: History, technology, and applications. *Machine Learning with Applications*, 2:100006.

Ameeta Agrawal, Aijun An, and Manos Papagelis. 2018. Learning emotion-enriched word representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 950–961, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Massa Baali and Nada Ghneim. 2019. Emotion analysis of arabic tweets using deep learning approach. *Journal of Big Data*, 6:1–12.

Jackylyn L. Beredo and Ethel C. Ong. 2022. A hybrid response generation model for an empathetic conversational agent. In *2022 International Conference on Asian Language Processing (IALP)*, pages 300–305.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Kenneth Mark Colby, Franklin Dennis Hilf, Sylvia Weber, and Helena C Kraemer. 1972. Turing-like indistinguishability tests for the validation of a computer simulation of paranoid processes. *Artificial Intelligence*, 3:199–221.

Jean Decety and Claire Holvoet. 2021. The emergence of empathy: A developmental neuroscience perspective. *Developmental Review*, 62:100999.

Robert Elliott, Arthur C Bohart, Jeanne C Watson, and David Murphy. 2018. Therapist empathy and client outcome: An updated meta-analysis. *Psychotherapy*, 55(4):399.

Joseph L. Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3):613–619.

Thomas Goetz, Eva S. Becker, Madeleine Bieg, Melanie M. Keller, Anne C. Frenzel, and Nathan C. Hall. 2015. The glass half empty: How emotional exhaustion affects the state-trait discrepancy in self-reports of teaching emotions. *PLOS ONE*, 10(9):1–14.

Sangchul Hahn and Heeyoul Choi. 2019. Self-knowledge distillation in natural language processing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 423–430, Varna, Bulgaria. INCOMA Ltd.

Eman Hamdi, Sherine Rady, and Mostafa Aref. 2019. A convolutional neural network model for emotion detection from tweets. In *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2018*, pages 337–346, Cham. Springer International Publishing.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Aref Jafari, Mehdi Rezagholizadeh, Pranav Sharma, and Ali Ghodsi. 2021. Annealing knowledge distillation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2493–2504, Online. Association for Computational Linguistics.

Zhao Jianqiang, Gui Xiaolin, and Zhang Xuejun. 2018. Deep convolution neural networks for twitter sentiment analysis. *IEEE Access*, 6:23253–23260.

Fazel Keshtkar and Diana Inkpen. 2011. A pattern-based model for generating text to express emotion. In *Affective Computing and Intelligent Interaction*, pages 11–21, Berlin, Heidelberg. Springer Berlin Heidelberg.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2021. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2227–2240, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

D Kinga, Jimmy Ba Adam, et al. 2015. A method for stochastic optimization. In *International conference on learning representations (ICLR)*, volume 5, page 6. San Diego, California;.

Sonia A Krol and Jennifer A Bartz. 2022. The self and empathy: Lacking a clear and stable sense of self undermines empathy and helping behavior. *Emotion (Washington, D.C.)*, 22(7):1554—1571.

M. S. Lebowitz and J. F. Dovidio. 2015. Implications of emotion regulation strategies for empathic concern, social attitudes, and helping behavior. *Emotion*, 15(2):187–194.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020. EmpDG: Multi-resolution interactive empathetic dialogue generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4454–4466, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. Knowledge bridging for empathetic dialogue generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10993–11001.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. MoEL: Mixture of empathetic listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 121–132, Hong Kong, China. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau.

2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Yukun Ma, Khanh Linh Nguyen, Frank Z. Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Information Fusion*, 64:50–70.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIME: MIMicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979, Online. Association for Computational Linguistics.

Xingliang Mao, Shuai Chang, Jinjing Shi, Fangfang Li, and Ronghua Shi. 2019. Sentiment-aware word embedding for emotion classification. *Applied Sciences*, 9(7):1334.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Chuong H. Nguyen, Thuy C. Nguyen, Tuan N. Tang, and Nam L. H. Phan. 2022. Improving object detection by label assignment distillation. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1322–1331.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Erika L. Rosenberg. 1998. Levels of analysis and the organization of affect. *Review of General Psychology*, 2(3):247–270.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. Cem: Commonsense-aware empathetic response generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11229–11237.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2015. Hierarchical neural network generative models for movie dialogues. *ArXiv*, abs/1507.04808.

Karen Sparck Jones. 1988. *A Statistical Interpretation of Term Specificity and Its Application in Retrieval*, page 132–142. Taylor Graham Publishing, GBR.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press.

Weiwei Sun, Zhengliang Shi, Shen Gao, Pengjie Ren, Maarten de Rijke, and Zhaochun Ren. 2023. Contrastive learning reduces hallucination in conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13618–13626.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Haoqing Wang, Xun Guo, Zhi-Hong Deng, and Yan Lu. 2022. Rethinking minimal sufficient representation in contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16041–16050.

Shuo Wang and Xiaofeng Meng. 2018. Multi-emotion category improving embedding for sentiment classification. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 1719–1722, New York, NY, USA. Association for Computing Machinery.

Guodong Xu, Ziwei Liu, Xiaoxiao Li, and Chen Change Loy. 2020. Knowledge distillation meets self-supervision. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX*, page 588–604, Berlin, Heidelberg. Springer-Verlag.

Zhou Yang, Zhaochun Ren, Wang Yufeng, Xiaofei Zhu, Zhihao Chen, Tiecheng Cai, Wu Yunbing, Yisong Su, Sibo Ju, and Xiangwen Liao. 2023. Exploiting emotion-semantic correlations for empathetic response generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4826–4837, Singapore. Association for Computational Linguistics.

Jamil Zaki. 2020. Integrating empathy and interpersonal emotion regulation. *Annual Review of Psychology*, 71(1):517–540. PMID: 31553672.

Weixiang Zhao, Yanyan Zhao, Xin Lu, and Bing Qin. 2023a. Don't lose yourself! empathetic response generation via explicit self-other awareness. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13331–13344, Toronto, Canada. Association for Computational Linguistics.

Weixiang Zhao, Yanyan Zhao, Xin Lu, and Bing Qin. 2023b. Don't lose yourself! empathetic response generation via explicit self-other awareness. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13331–13344, Toronto, Canada. Association for Computational Linguistics.

Juan Zheng, Susanne Lajoie, and Shan Li. 2023. Emotions in self-regulated learning: A critical literature review and meta-analysis. *Frontiers in Psychology*, 14.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. An affect-rich neural conversational model with biased attention and weighted cross-entropy loss. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7492–7500.

Jinfeng Zhou, Chujie Zheng, Bo Wang, Zheng Zhang, and Minlie Huang. 2023. CASE: Aligning coarse-to-fine cognition and affection for empathetic response generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8223–8237, Toronto, Canada. Association for Computational Linguistics.

Yuanhang Zhou, Kun Zhou, Wayne Xin Zhao, Cheng Wang, Peng Jiang, and He Hu. 2022. C^2-crs: Coarse-to-fine contrastive learning for conversational recommender system. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 1488–1496, New York, NY, USA. Association for Computing Machinery.

Language Resource References

Mohammad, Saif. 2018. *Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words*. Association for Computational Linguistics. PID https://aclanthology.org/P18-1017.

Rashkin, Hannah and Smith, Eric Michael and Li, Margaret and Boureau, Y-Lan. 2019. *Towards empathetic open-domain conversation models: A new benchmark and dataset*. Association for Computational Linguistics.