From Traits to Empathy: Personality-Aware Multimodal Empathetic Response Generation

Anonymous ACL submission

Abstract

Empathetic dialogue systems improve the user experience across various domains. Existing approaches mainly focus on acquiring affective and cognitive information from text, often neglecting the unique personality traits of individuals and the inherently multimodal nature of human conversation. To this end, we propose enhancing dialogue systems with the ability to generate customized empathetic responses, considering the diverse personality traits of speakers, and we advocate for the incorporation of multimodal data analysis to gain a more detailed comprehension of speakers' emotional states and context. Specifically, we initially identify the speaker's trait across the context. The dialogue system then comprehends the speaker's emotion and situation by emotion perception through the analysis of multimodal inputs. Finally, the response generator models the correlations among the captured personality, emotion, and multimodal data, thereby generating empathetic responses. Extensive experiments are conducted utilizing the MELD dataset and the IEMOCAP dataset to investigate the influence of personality traits on empathetic response generation and validate the effectiveness of the proposed approach.

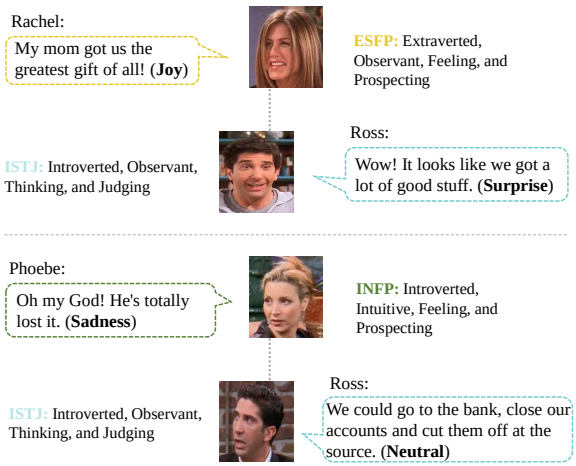

Figure 1: The examples illustrate humans' propensity to consider their conversational partners' personalities to achieve empathy. The individual with an ESFP personality is depicted as lively, extroverted, and sharing their joy with others. The individual with an INFP personality is portrayed as quiet and introverted, possessing a spirit of exploration and a tendency to approach problem-solving creatively. Upon analysis of Ross's responses to Rachel and Phoebe, it becomes apparent that Ross deliberately considers the distinct personality traits of each speaker in his interactions, which facilitates his ability to achieve empathy with them.

1 Introduction

Empathy is often defined as the ability to understand and potentially share and react to another person's feelings and experiences from their perspective (Macarov and David, 1978; Main, 2021; Liu and Picard, 2005). Research in psychology and mental health establish empathy as a crucial component in the development of dialogue systems that aim to provide more humanized interactions (Zech and Rimé, 2005).

The advent of the EmpatheticDialogue dataset (Rashkin et al., 2019) amplifies interest in empathetic response generation, underscoring its wide-ranging applicability across diverse fields (Zhou et al., 2020; Song et al., 2021b; Kulshreshtha et al., 2020). Predominantly, existing endeavors focus on discerning speakers' emotional states through emotion recognition and employing knowledge graphs to deduce implicit information within the dialogue context (Raamkumar and Yang, 2023; Ma et al., 2020). Some researchers propose to apprehend speakers' emotions at utterance level, including mixture of empathetic listeners (Lin et al., 2019), emotion mimicry (Ghosal et al., 2020), while others examine strategies to model speakers' feelings comprehensively, incorporating multi-task learning (Varshney et al., 2021), multi-resolution adversarial training (Li et al., 2020). Moreover, knowledge graphs are applied to infer broader contextual in-

formation directly from dialogues (Sabour et al., 2022; Wang et al., 2022; Zhou et al., 2023), which function as prior knowledge, thereby guiding dialogue systems in generating responses that are more relevant and consistent. Recently, the newly introduced large language models (LLMs), such as GPT 4 (OpenAI, 2023a) and Claude 3 (Anthropic, 2023), demonstrating proficiency in comprehending, inferring, and conveying empathy (Lee et al., 2024). Whereas, these models are expensive and not completely open-source, leaving the details of their development process somewhat opaque.

However, these studies often ignore the significant influence of speakers' personality traits, and train conversational models without adapting to differences in empathy expression, so that to generate standardized responses and struggle to engage users who may discern the mechanical nature of the dialogue system (WEN et al., 2021). In human interactions, the expression of empathy is not isolated from individuals' personality traits, such as those outlined by the Myers Briggs Type Indicator (MBTI) (Carlson, 1985). MBTI is a psychological assessment tool (Jung and Beebe, 2016) that categorizes individuals into 16 personality types based on four dichotomies: Extraversion (E) vs. Introversion (I), Sensing (S) vs. Intuition (N), Thinking (T) vs. Feeling (F), and Judging (J) vs. Perceiving (P). It is designed to help people understand personal preferences and improve interpersonal relationships (Cohen et al., 2013). During interactions, individuals not only resort to their habitual modes of expressing empathy but, more importantly, adapt tailor their empathetic responses to match the personality traits of their interlocutors (Chae, 2016). Despite considerable efforts dedicated to the development of persona-based dialogue models (Zhong et al., 2020a; Song et al., 2021a; Xu et al., 2022), the existing persona-related works still face several issues: the data volume is often insufficient (WEN et al., 2021), and the focus of persona information tends to be on users' demographic data rather than their deeper personality traits (Zhong et al., 2020b; Ahn et al., 2023).

Therefore, we propose a multimodal dialogue system that is attentive to personality intricacies and can produce diverse, targeted empathetic responses. To achieve this, we utilize a pre-trained MBTI classifier (Ryan et al., 2023) to infer speakers' personalities from their dialogue history, going beyond the current scope of persona-based works. We employ multimodal emotion recognition to capture emotions, which are then combined with personality traits as control signals. For text processing, we use the GPT-2 model (Radford et al.) to extract features from the dialogue, and we leverage a pre-trained BLIP model for visual information (Li et al., 2022a). A cross-modal feature fusion module integrates these multimodal features, emphasizing relevant image aspects in the context of the dialogue, ensuring that the features are well-optimized for the final stage of response generation.

In summary, our work presents several significant contributions to the field:

(1) We propose integrating personality into the response generation process, enabling more empathetic interactions.

(2) We design a multimodal framework that ensures the generated empathetic responses are contextually coherent and emotionally attuned.

(3) Extensive experiments on the MELD and IEMOCAP datasets, using both machine and human evaluations, demonstrate the superior efficacy of our proposed method.

2 Related Work

2.1 Multimodal Emotion Recognition

Multimodal emotion recognition in conversation aims to recognize human emotions via multimodal data (Lian et al., 2023), which has seen extensive research. For instance, Makiuchi et al. (2021) propose using high-level features to improve emotion recognition. Li et al. (2022b) consider the emotional tendencies of utterances, and extract multimodal representations from various modalities. Chudasama et al. (2022) design a multimodal fusion network, complemented by an adaptive margin-based loss. Srivastava et al. (2023) embark on an endeavor to analyze the emotions and mental states of characters within cinematic narratives. Shi and Huang (2023) devise a focus-weighted focal contrastive loss to focus on emotions that are difficult to discern. The insights derived from the aforementioned works provide a valuable repository of knowledge that can be applied to enhance the capability of empathetic dialogue systems to comprehend speakers' feelings.

2.2 Empathetic Response Generation

Empathetic response generation necessitates that dialogue systems understand speakers' emotions and situations (Li et al., 2021), so that generate pertinent responses and achieve empathy with speakers.

The seminal work of Rashkin et al. (2019), which introduces the task and establishes the benchmark dataset, has catalyzed heightened interest in this area. Some works endeavours to endow dialogue systems with the capability to comprehend affective knowledge via emotion perception. Lin et al. (2019) (**MoEL**) employ n encoders to identifying emotions with a specific category. Ghosal et al. (2020) (**MIME**) divide emotions into two groups according to their polarity and integrate emotions with stochasticity. Li et al. (2020) identify emotions from both utterance level and token level, to capture the subtle emotions in dialogues (**EmpDG**). While other researchers (Li et al., 2021; Hwang et al., 2020) introduce knowledge graphs to infer speakers' circumstances. Sabour et al. (2022) feed the dialogue history to Comet (Bosselut et al., 2019), and obtain inferences from five distinct aspects (**CEM**). Wang et al. (2022) address the challenge of capturing dynamic emotional shifts in conversations, as well as the potential discrepancies between knowledge graph inferences and actual emotions expressed (**SEEK**). Zhao et al. (2023) propose a framework (**EmpSOA**), consisting of self-other differentiation and modulation, and a response generator. (Zhou et al., 2023) construct the cognition graph utilizing inferred knowledge and the emotional concept graph to align speakers' cognitive and affective information (**CASE**).

In summary, previous studies extract speakers' emotional and situational details from both affective and cognitive information through solely textual data but restrict the maximum depth of understanding that dialogue systems can reach regarding speakers.

3 Problem Statement

We denote a dialogue context as a sequence of n utterances, represented by the notation $(U, \hat{P})$, where $U = \{u_1, \ldots, u_n\}$, and $u_i = \{u_i^t, u_i^v\}, i \in [1, n]$. U indicates the utterances in the dialogue history, with u_i^t and u_i^v denoting textual and visual data of each utterance. $\hat{P} = \{p_1, \ldots, p_l\}$ represents the set of personality traits associated with l speakers engaged in a singular dialogue context. Besides, $u_i^t = \{w_i^1, \ldots, w_i^k\}$ elucidates that the utterance u_i^t consists of k words. l and k can vary from various contexts and utterances. The task is to train a model $P(u_{n+1}|u_{<n+1}, u_n^v, p_n; \theta)$ to generate empathetic responses u_{n+1} that are cognizant of the personality traits embedded within the dialogue context, where θ represents the parameters of the model.

4 Methodology

Our proposed personality-aware framework is present in Figure 2, which mainly incorporates a refine encoder, a cross-modal fusion encoder for multimodal emotion perception, an emotion recognizer, a personality classifier, and the response generator. Various special tokens are shown in token samples. For example, the BOS token and EOS token indicate the beginning and the end of a context, and the SEP token separates different speakers' utterances.

4.1 Cross-Modal Emotional Insights

To understand the speaker's emotional states from the dialogue history, we employ multimodal emotion recognition techniques. Specifically, for each multimodal input $\{u^t, u^v\}$, the pre-trained BLIP model with a projection linear layer and the pre-trained GPT-2 model act as feature extractors to obtain the visual representations $r^v \in \mathbb{R}^d$ and the textual representations $r^t \in \mathbb{R}^{k \times d}$ respectively, where k is the length of the utterance u^t and d is the dimension of the feature space.

The refine encoder plays a pivotal role in distilling the features of visual representations pertinent to the task at hand. Specifically, the representations derived from visual data are mapped into query, key, and value domains as defined by Equation 1:

$$Q_{r^v},\ K_{r^v},\ V_{r^v} = W_q r^v,\ W_k r^v,\ W_v r^v \tag{1}$$

where $W_q, W_k, W_v \in \mathbb{R}^{d \times d}$ represents learnable parameter matrices, and $Q_{r^v}, K_{r^v}, V_{r^v}$ are the query, key and value matrices. Then, the self-attention encodes the visual features by matching their query and key matrices, which is calculated by Equation 2:

$$A_{r^v} = \sigma\left(\frac{Q_{r^v} K_{r^v}^T}{\sqrt{d}}\right) V_{r^v} \tag{2}$$

where $K_{r^v}^T$ is the transposed key matrix, $A_{r^v} \in \mathbb{R}^d$ is the refined visual features, and $\sigma(\cdot)$ denotes the softmax function.

Similar to the refine encoder, the cross-modal fusion encoder processes the textual representations via self-attention encoding, as described in Equation 2, resulting in an encoded matrix $A_{r^t} \in \mathbb{R}^d$. The cross-modal fusion encoder aims to model the correlation between pairwise features of visual and textual modalities. In this stage, the cross-modal

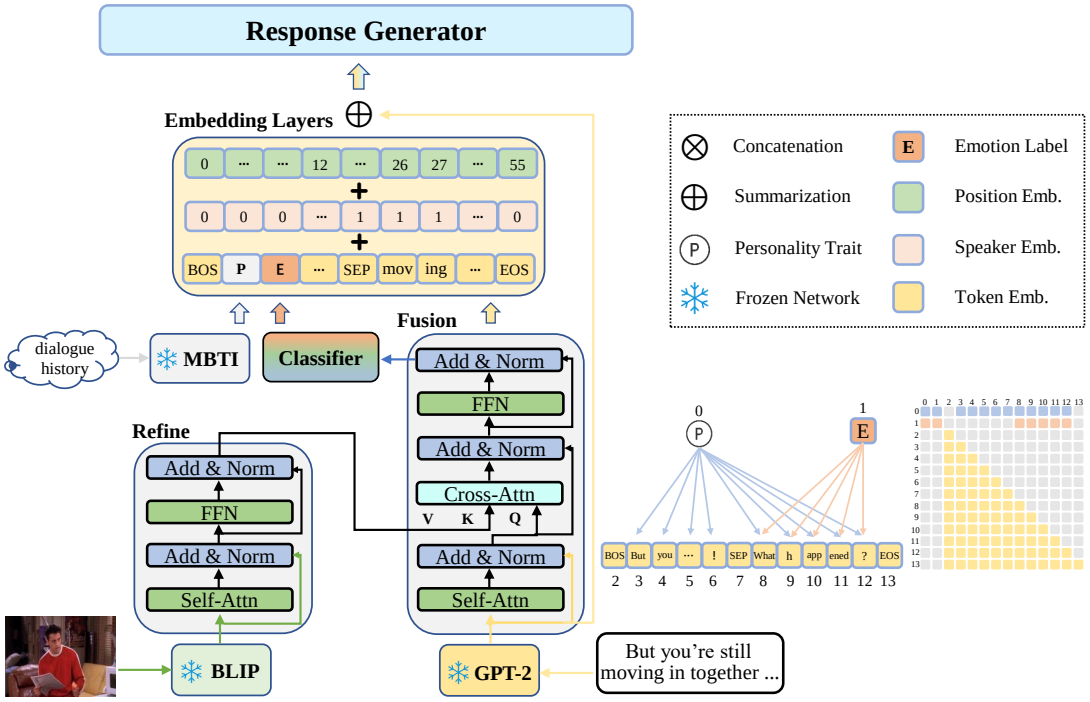

Figure 2: Overview of the proposed framework. The visual features refined by a specialized encoder, are integrated with textual features in a cross-modal fusion encoder for multimodal emotion recognition. The incorporation of personalities, emotional labels, and multimodal representations augments the response generator to produce responses that are not only contextually relevant but also empathetically and personally attuned.

attention matches the query matrix A_{r^t} of the textual modality with the key matrix A_{r^v} of the visual modality to learn the correlation, which can be formulated as:

$$A_{tv} = \sigma\left(\frac{A_{r^t} A_{r^v}^T}{\sqrt{d'}}\right) A_{r^v} \tag{3}$$

where $A_{r^v}^T$ is the transposed key matrix of A_{r^v}, and d' is the dimension of the attention heads. Subsequently, the combined data proceeds through the feed-forward layer and the residual normalization layer, we specify the output of the cross-modal fusion encoder as $H \in \mathbb{R}^{k \times d}$. After that, a linear classifier is applied to the output H and predicts the emotion label E, formalized by Equation 4:

$$E = argmax\left(LN\left(W_h H\right)\right) \tag{4}$$

where LN represents the linear layers within the classifier, W_h is learnable parameters, and E indicates the predicted emotion label. Therefore, we calculate the loss of the multimodal emotion perception by Equation 5:

$$\mathcal{L}_E = -\frac{1}{\sum_{h=1}^{m} f(h)} \sum_{j=1}^{m} \sum_{i=1}^{f(j)} y_{ji} \log(E) \tag{5}$$

where m is the total number of dialogues in the training set, $f(j)$ signifies the count of utterances within the j-th dialogue context, $\log(E)$ and y_{ij} represents the probability distribution of emotion label and the ground truth label respectively.

4.2 Personality Indicator

We employ a pre-trained personality classifier $\mathcal{C}$, which achieves an average classification accuracy of 84.34% on Kaggle's MBTI dataset[1] (Ryan et al., 2023), to infer personality traits for each speaker. We begin by grouping the utterances in the dialogue context by speaker. For a given speaker s, we concatenate the utterances to form a set $U_s = \{u_{s1}, u_{s2}, \cdots\}$, which serves as input to the classifier $\mathcal{C}$, yielding the personality type $p = \mathcal{C}(U_s)$. Each personality type p is associated with a corresponding text description $\mathcal{R}$, we provide the specific 16 descriptions in the appendix A. In our experiments, we prepend a CLS token to each description, creating $\widetilde{\mathcal{R}} = [CLS] \oplus \mathcal{R}$. We then input $\widetilde{\mathcal{R}}$ into the GPT-2 model to obtain the representation p_s of the CLS token, which we use as the representative embedding for the personality p.

Subsequently, the emotion E and the personality

[1] https://www.kaggle.com/

p_s collaboratively control the generation process. We differentiate between tokens that serve as control signals and those that constitute dialogues. We model their relationship with a mask matrix W_m during the self-attention operation. Concretely, if token$_i$ controls token$_j$, the value at position (i, j) in W_m is 0, otherwise is negative infinity:

$$W_m(i,j) = \begin{cases} 0, i \Rightarrow j \\ -inf, i \nRightarrow j \end{cases} \quad (6)$$

This mechanism allows us to use the mask matrix to guide the generation of each response token using signals from various perspectives, representing diver factors for expressing empathy.

4.3 Empathetic Response Generator

We aggregate all utterances and control signals within a dialogue, and integrate special tokens to indicate the start and the end of the dialogue. The construction of input embeddings is a multifaceted process, encompassing token embeddings, speaker type embeddings, and position embeddings, which results in the formation of input context demoted as $X = x_1, \cdots, x_s$, with the ground truth response delineated as $Y = x_{s+1}, \cdots, x_N$, thus the conditional probabilities of $P(Y|X)$ can be formulated as:

$$P(Y|X) = \prod_{n=s+1}^{N} p(x_n|x_1, \cdots, x_s; p_s, E, \theta) \quad (7)$$

where θ represents the parameters of the model, p_s and E denote the control signals. Specifically, as depicted in Figure 2, p_s controls both the speaker's utterances and the response, while E only controls the response, and they also control and interact with each other. Besides, to capitalize on the advanced language processing capabilities of the pre-trained model, we introduce an efficient residual connection to integrate the output of the cross-modal fusion encoder with the hidden states from the GPT-2 model, which can be formulated as:

$$I = W^G h^G + W^H H \quad (8)$$

where W^G and W^H correspond to the linear projections of the language model and the fusion encoder respectively, and h^G represents the hidden states derived from the language model. Generally, one would use the cross-modal representation for generation, but such approach overlooks the GPT-2 model's exceptional skills in language, which provides a language-only generation perspective.

Moreover, when considering a multi-turn dialogue $D_1, \cdots, D_w$, the probability of generating a dialogue sequence can be reformulated as $P(D_w, \cdots, D_2|D_1)$, which can be computed through the multiplication of conditional probabilities of $P(D_i|D_1, \cdots, D_{i-1})$, taking into account all preceding dialogue contexts and their corresponding ground truth responses.

Consequently, to train the response generator, we opt for the standard negative log-likelihood (NLL) loss applied to the target responses, which is represented by:

$$\mathcal{L}_Y = \mathbb{E}_{(D,Y)} \left[-\log P(Y) \right] \quad (9)$$

where D is the dialogue context. During the training phase, the refine encoder, the cross-modal fusion encoder, the emotion recognizer, and the response generator concurrently update their parameters, enabling the seamless integration of multimodal features with textual features in the embedding space, and enhancing the model's capacity to capture the complex semantic information inherent in multimodal data. Considering the above components, an aggregated loss function is employed as the comprehensive optimization objective, facilitating an end-to-end training paradigm, expressed as:

$$\mathcal{L} = \lambda \mathcal{L}_Y + \gamma \mathcal{L}_E \quad (10)$$

where $\lambda = 1$ and $\gamma = 0.5$ are hype parameters, functioning to equilibrate the contributions of multimodal emotion recognition and empathetic response generation within the overall framework.

5 Experiments

5.1 Datasets

Our experiments utilize the MELD dataset (Poria et al., 2019) and the IEMOCAP dataset (Busso et al., 2008), both of them include multiple daily conversations annotated with emotional labels, as well as multimodal data. We use the original partitioning of the MELD dataset for training, validation and testing. We follow the previous work (Makiuchi et al., 2021) that choose Session 1 to Session 4 for training and use Session 5 for testing. Particularly, we randomly pick up 10% of the training data for validation.

5.2 Implementation Details

All codes are implemented with PyTorch. To build the framework, we incorporate the pre-trained

Table 1: P represents personalities, V is the visual data, mask and residual indicate the mask matrix and the residual connection. Acc is the average accuracy of emotion recognition.

Datasets	Ablation	PPL↓	Dist-1	Dist-2	Acc (%)
MELD	Ours	**35.38**	**2.12**	**9.83**	**67.05**
	w/o P	36.92	1.54	6.38	-
	w/o V	38.14	1.47	6.04	61.98
	w/o P&V	40.25	1.04	4.75	-
	w/o mask	40.09	1.61	6.24	-
	w/o residual	46.58	1.72	6.46	66.21
IEMOCAP	Ours	**30.64**	**5.63**	**20.46**	**64.12**
	w/o P	30.91	3.95	14.52	-
	w/o V	31.23	3.76	14.08	60.16
	w/o P&V	33.95	3.48	12.83	-
	w/o mask	33.46	3.99	14.55	-
	w/o residual	38.28	4.21	15.26	61.95

Table 2: Results of aspect-based pair comparisons (%). Ties are not shown. $0.4 < \kappa < 0.6$ indicates moderate agreement. †, ‡ denote significant improvement with $p-$value $< 0.1/0.05$.

Comparisons	Aspects	Win	Lose	κ
Ours vs. MoEL	Emp.	**56.2**†	33.5	0.46
	Coh.	**52.8**‡	30.4	0.53
	Flu.	**46.4**‡	35.7	0.49
Ours vs. MIME	Emp.	**57.8**‡	33.2	0.51
	Coh.	**52.9**‡	31.4	0.54
	Flu.	**46.2**‡	34.6	0.48
Ours vs. EmpDG	Emp.	**51.7**‡	35.8	0.55
	Coh.	**49.1**‡	34.5	0.52
	Flu.	**48.3**‡	30.0	0.54
Ours vs. CEM	Emp.	**48.4**‡	32.3	0.53
	Coh.	**53.3**‡	37.4	0.51
	Flu.	**47.2**†	40.2	0.44
Ours vs. SEEK	Emp.	**52.6**‡	30.9	0.52
	Coh.	**50.4**‡	38.7	0.53
	Flu.	**48.6**‡	41.6	0.49
Ours vs. EmpSOA	Emp.	**49.5**‡	31.1	0.48
	Coh.	**52.1**‡	36.5	0.56
	Flu.	**50.7**‡	39.7	0.54
Ours vs. CASE	Emp.	**49.5**‡	31.1	0.48
	Coh.	**52.1**‡	36.5	0.56
	Flu.	**50.7**‡	39.7	0.54

BLIP model (Li et al., 2022a) and the pre-trained GPT-2 model (Radford et al.) for pre-processing. The response generator is a decoder-only model built upon transformer blocks (Vaswani et al., 2017), consisting of 24 blocks with a multi-head self-attention layer (12 heads) and a feed-forward layer each. It is initially pre-trained on the EmpatheticDialogues dataset (Rashkin et al., 2019) for 10 epochs with a batch size of 8, enhancing its capacity for empathetic expression, and then fine-tuned on the two datasets, respectively. For inference, we employ a batch size of 1 and limit the decoding process to 30 steps, along with the nucleus sampling strategy with $p = 0.8$. We adopt the Adam optimizer with a learning rate of 1e-5. For the entirety of our training and fine-tuning phases, we utilize two NVIDIA Geforce RTX 3090 GPU cards equipped with 24 GB RAM of each, and we maintain the fine-tuning state until it becomes apparent that there is no additional decrease in loss achievable. For comparative analysis, we adhere to the original settings of official codes from all methods under consideration. All baselines follow the same experimental procedure as ours. Whereas, for text-only baselines, we use only the text portions of the datasets. For multimodal methods, we utilize their released model weights and fine-tune their models on the data used in our approach.

5.3 Ablation Study

Following the previous works (Sabour et al., 2022; Wang et al., 2022; Zhao et al., 2023), our evaluation employs automatic metrics: (1) Perplexity (**PPL**), assessing the overall quality of responses, where lower values denote higher quality. (2) Distinct-n (**Dist-1**, **Dist-2**), reflecting response diversity, with higher scores indicating greater diversity. Additionally, we evaluate our model's ability to accurately perceive speakers' emotions, using the average accuracy metric (**Acc.**), which complements the primary metrics by highlighting the model's emotional intelligence in dialogue contexts.

As illustrated in Table 1, we perform ablation studies to substantiate the essential roles of the components in our framework. Removing personality or visual data significantly reduces the diversity of responses, especially when visual data is removed, which greatly decreases emotion recognition accuracy. The masking operation and the residual connection help enhance the diversity of generated responses. Besides, removing the residual connection increases the model's perplexity and slightly decreases emotion recognition accuracy.

5.4 Automatic Evaluation

Table 3 provides an extensive experimental analysis, comparing the performance of our method with the contemporary state-of-the-art approaches using automatic evaluation metrics. Due to the absence of prior work on multimodal empathetic response generation, for fairness, we select two multimodal dialogue generation works (Han et al., 2023; Li et al., 2023) for comparison.

Illustrated by the results, Pace (Li et al., 2023)

Table 3: Evaluations of our method and the baselines. Acc is the average accuracy of emotion recognition.

Datasets	Methods	Automotic Evaluation				Human / GPT-4 Evaluation		
		PPL ↓	**Dist-1**	**Dist-2**	**Acc (%)**	**Emp.**	**Coh.**	**Flu.**
MELD	Pace(Li et al., 2023)	**26.19**	1.86	6.97	-	1.92 / 1.85	3.31 / 3.42	3.55 / 3.95
	CHAMPAGNE(Han et al., 2023)	36.25	1.73	6.42	-	1.88 / 1.80	3.22 / 3.26	3.42 / 3.86
	MoEL(Lin et al., 2019)	50.41	0.71	3.22	57.93	2.91 / 2.68	3.09 / 3.19	3.37 / 3.68
	MIME(Ghosal et al., 2020)	48.50	0.64	2.88	56.90	2.88 / 2.83	3.14 / 3.22	3.34 / 3.77
	EmpDG(Li et al., 2020)	50.51	0.89	4.05	57.62	2.95 / 2.82	3.22 / 3.28	3.42 / 3.70
	CEM(Sabour et al., 2022)	54.00	0.97	4.36	57.55	3.02 / 2.85	3.27 / 3.30	3.65 / 3.79
	SEEK(Wang et al., 2022)	54.72	1.01	4.54	58.95	3.11 / 2.84	3.24 / 3.32	3.58 / 3.89
	EmpSOA(Zhao et al., 2023)	53.33	1.02	4.60	59.69	3.13 / 2.76	3.28 / 3.33	3.61 / 3.95
	CASE(Zhou et al., 2023)	55.27	1.05	4.68	58.84	3.12 / 2.80	3.25 / 3.36	3.63 / 3.92
	Ours	35.38	**2.12**	**9.83**	**67.05**	**3.26 / 2.99**	**3.43 / 3.49**	**3.71 / 4.00**
IEMOCAP	Pace(Li et al., 2023)	**28.33**	4.65	17.46	-	1.95 / 1.82	3.25 / 3.40	3.60 / 3.97
	CHAMPAGNE(Han et al., 2023)	30.62	4.37	15.54	-	1.87 / 1.79	3.18 / 3.25	3.46 / 3.80
	MoEL(Lin et al., 2019)	36.86	2.82	9.66	54.18	3.01 / 2.76	3.02 / 3.18	3.34 / 3.71
	MIME(Ghosal et al., 2020)	36.48	2.33	8.27	53.52	3.10 / 2.82	3.09 / 3.26	3.30 / 3.69
	EmpDG(Li et al., 2020)	35.80	2.14	8.12	54.27	3.02 / 2.75	3.17 / 3.24	3.39 / 3.72
	CEM(Sabour et al., 2022)	36.17	3.15	11.35	56.83	3.13 / 2.82	3.21 / 3.33	3.50 / 3.83
	SEEK(Wang et al., 2022)	36.91	3.78	13.61	58.40	3.17 / 2.89	3.19 / 3.29	3.53 / 3.88
	EmpSOA(Zhao et al., 2023)	34.56	3.90	14.15	58.35	3.20 / 2.94	3.29 / 3.35	3.58 / 3.94
	CASE(Zhou et al., 2023)	36.02	3.85	14.30	57.66	3.23 / 2.90	3.26 / 3.38	3.57 / 3.94
	Ours	30.64	**5.63**	**20.46**	**64.12**	**3.36 / 3.03**	**3.35 / 3.45**	**3.62 / 3.98**

attains the lowest PPL scores on both datasets, which is likely attributable to its robust pre-training process. Our method obtains the competitive perplexity score, reflecting its overall response quality. Meanwhile, our model also surpasses all the compared models significantly in Dist-1 and Dist-2 metrics, demonstrating its capacity to generate a wider kind of empathetic responses, thereby catering to user needs across diverse multimodal contexts. The superior performance of our response generator can be attributed to its decoder-only architecture, the masking operation, and the utilization of a large-scale, multi-turn dialogue dataset for pre-training. Besides, our approach excels in multimodal emotion perception accuracy, benefiting from the specialized refine and cross-modal fusion encoders, as well as the efficacy of the employed feature extractors.

5.5 Human and GPT-4 Evaluation

To evaluate the quality of the generated empathetic responses from humans' perspective, following the previous works (Li et al., 2020; Zhao et al., 2023), we conduct human evaluations on 100 randomly selected dialogue context-response pairs generated by our model and the baselines. These evaluations assess the empathetic quality of responses from the following aspects:(1) Empathy (**Emp.**): assessing the response's ability to reflect an understanding of the speaker's emotions and situation; (2) Coherence (**Coh.**): evaluating the response's consistency with the preceding dialogue and its relevance to the topic; (3) Fluency (**Flu.**): determining the naturalness and smoothness of the response.

To facilitate human evaluations, we enlist five independent graduate researchers, ensuring no conflicts of interest, to rate the context-response pairs on a scale from 1 (lowest) to 5 (highest) across empathy, coherence, and fluency dimensions. The average scores calculated from these ratings are summarized in Table 3. We also utilize the GPT-4 API to evaluate the randomly selected responses as a reference by configuring the API parameters with a temperature setting of 0, which ensures that the same inputs to GPT-4 will produce consistent outputs (OpenAI, 2023b).

Furthermore, to account for individual variations among annotators, we conduct aspect-based pairwise comparisons to directly evaluate the response quality between our model and the baselines, focusing on empathy, coherence, and fluency. Given any two generated responses, the annotators are instructed to make a preferred choice by choosing the "Win" or "Lose" option. If the annotators find it hard to choose a better one in both responses, they could choose the "Tie" option. However, we encourage them to make their preferences. Besides, we utilize Fleiss's kappa (Fleiss, 2006) to measure the agreement of the annotators. The outcomes, detailed in Table 2, reveal a clear preference for responses generated by our model, underscoring its empathetic response capabilities.

The results presented in Table 3 and Table 2 demonstrate that our approach not only attains the highest scores compared to other state-of-the-art empathetic methods, but also excels in empathy, coherence, and fluency aspects, which underscores our method's superior ability to generate responses that more effectively express empathy and align with speakers' perspectives.

Generated Responses:
MoEL**:** Oh, I'm sorry to hear that.
MIME**:** I'm sorry to hear that.
EmpDG**:** Oh, no, I'm sorry to hear that.
CEM**:** Oh, What happened?
SEEK**:** I hope you feel better.
EmpSOA: That is sad. I'm sorry to hear that.
Ours: I'm sorry to hear that. How can I make your **foot** more **comfortable**?

Image

Context:
I'm sorry. It's just my **foot itches** like crazy. (INTJ & Sadness)
Ground Truth:
Oh, I'll get it.

Image

Context:
Everything's ruined. My bed. My clothes. Look at my favorite blue sweater. (ESFP & Sadness)
Ground Truth:
Isn't that mine?

Generated Responses:
MoEL**:** That is terrible.
MIME**:** I'm sorry to hear that.
EmpDG**:** I'm sorry to hear that. I'm sure you'll be fine.
CEM**:** Oh, no! What happened?
SEEK**:** I'm sorry to hear that. I'm glad you are okay.
EmpSOA: That is sad. I'm sorry to hear that.
Ours: It's okay. **Everything's** gonna be alright.

Generated Responses:
MoEL**:** I do not know.
MIME**:** I do not know what to do.
EmpDG**:** Okay, I'm sorry!
CEM**:** Oh, okay.
SEEK**:** Okay, I'll do that.
EmpSOA: Oh, that's a good idea.
Ours: Well, I'm going to **tell her**. I'll go do that.

Image

Context:
Well, Why don't you **tell her** to stop being silly! (ENTP & Anger)
Ground Truth:
Okay, okay!

Image

Context:
Why do all you're coffee mugs have numbers on the bottom? (ISTJ & Surprise)
Ground Truth:
Oh. That's so Monica can keep track.

Generated Responses:
MoEL**:** I do not know.
MIME**:** I do not know.
EmpDG**:** I'm sorry. I do not know.
CEM**:** I have no idea what you mean.
SEEK**:** I do not know what you mean.
EmpSOA: I have no idea what that is.
Ours: What is that? That might be interesting to look up.

Figure 3: The cases generated by our model and the baselines. We highlight those words or responses that illustrate the priority of our model in understanding the speaker's situation and showing much more empathy.

6 Case Study

We exhibit cases across four distinct scenarios in Table 3, showing empathetic responses generated by our model alongside the baselines, which underscores the superiority of the proposed approach in facilitating empathetic interaction.

Specifically, in the top-left example, the speaker is characterized by the INTJ personality type, marked by a reluctance to express sentiments. Our model empathizes towards the speaker's itchy condition and introverted nature, and also proposes to alleviate the discomfort. In the top-right example, the speaker is identified with the ESFP personality, demonstrating a willingness to share feelings. In response to the sadness expressed by the speaker, the baselines produce general and safe comforting replies, but our model responds with more relevant information. In the bottom-left example, the speaker is exemplified as embodying the ENTP personality, characterized by tenacity to achieve goals irrespective of encountered challenges. Among the generated responses, only SEEK and our model produce responses congruent with the speaker's aspirations. In the bottom-right example, the speaker is portrayed as embodying the ISTJ personality type, known for their thoughtful and inquisitive trait. The baselines simply respond with "I do not know.", showcasing a lack of engagement. In contrast, our model follows the cue of questioning by proposing to look up the number, which is very much in tune with the speaker's personality. These cases demonstrate that our model generates empathetic responses that align with the distinct personalities of the dialogue participants.

7 Conclusion

In this paper, we endeavor to tackle the challenge inherent in empathetic response generation, identifying a gap in current state-of-the-art methods, especially their limitations in incorporating multimodality and personality dimensions. We propose a multimodal framework that capitalizes on the integration of multimodal information and personality traits to attain a comprehensive understanding of the speaker's emotional state and situational context, aiming to generate empathetic responses that are not only pertinent to the context but also resonate on a personal level with the speaker. Our study not only advances the empathetic response generation field but also underscores the significance of multimodal data and personality awareness in creating more meaningful and effective empathetic interactions. However, there are some works that endow large language models (LLMs) with defined personalities (Cui et al., 2023). Considering the substantial expenses associated with training LLMs, our future work will focus on employing knowledge distillation to utilize the outputs from such LLMs to facilitate the development of a more human-like yet cost-effective model.

8 Limitations

In human conversation, each individual has a distinct personality type. However, the dialogue system we have developed does not incorporate a specific personality, resulting in a system that lacks sufficient anthropomorphism.

In addition, we investigate the limitations of our approach through two illustrative examples from Figure 5. In the left example of Figure 5, the speaker's personality type is ISTP, which constitutes only 0.3% of the datasets. Our model has not learned well to generate responses tailored to such infrequently occurring personality traits, consequently producing more generalized but less empathetic responses. Meanwhile, it is challenging to extract information related to the speaker's emotions from the visual input of this example. As a result, our model's multimodal emotion recognition technique does not perform effectively in such scenarios. In the right example of Figure 5, our method

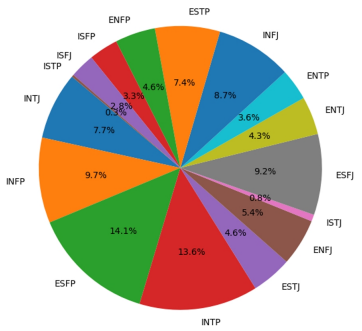

Figure 4: Analysis of personality type distribution within samples from the datasets we used.

is limited by the absence of visual input, evident in two primary ways: First, our model is unable to acquire additional insights about the speaker's situation through multimodal representations; second, it fails to accurately discern the speaker's emotional states using multimodal emotion recognition techniques. Indeed, in this instance, our model wrongly identifies the speaker's emotion as neutral, leading to a response that lacks empathy. However, human conversation in its natural form is inherently multimodal (Poria et al., 2019), suggesting that multimodal inputs are essential for achieving empathetic dialogue. Therefore, our future work will aim to overcome challenges associated with minority personality types and to develop techniques effective across both multimodal and purely textual contexts.

Context:
My apartment isn't there anymore. (ISTP & Sadness)
Generated Responses:
MoEL: I'm sorry to hear that.
MIME: I'm sorry to hear that.
EmpDG: That's terrible!
CEM: I'm sorry to hear that.
SEEK: That's a tough situation.
EmpSOA: That happens to me.
Ours: Oh! What happened?

No Visual Modality

Context:
She's gonna help us take care of the baby. (ESFP & Surprise)
Generated Responses:
MoEL: I'm glad to hear that.
MIME: That sounds great.
EmpDG: I'm glad to hear that.
CEM: I'm sure she will do it well.
SEEK: Oh, that is so great!
EmpSOA: That is so nice of her.
Ours: Why would she do that?

Figure 5: The limited instances. The left is a failure example, and the right is a case of missing visual modality.

References

Jaewoo Ahn, Yeda Song, Sangdoo Yun, and Gunhee Kim. 2023. MPCHAT: Towards multimodal persona-grounded conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3354–3377, Toronto, Canada. Association for Computational Linguistics.

Anthropic. 2023. Introducing the claude 3 family. https://www.anthropic.com/news/claude-3-family.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

John G Carlson. 1985. Recent assessments of the myers-briggs type indicator. *Journal of personality assessment*, 49(4):356–365.

Myung-Ock Chae. 2016. Empathic ability and communication ability according to myers-briggs type indicator (mbti) personality type in nursing students. *Journal of the Korea Academia-Industrial Cooperation Society*, 17(4):303–311.

Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe. 2022. M2fnet: Multi-modal fusion network for emotion recognition in conversation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4651–4660.

Yuval Cohen, Hana Ornoy, and Baruch Keren. 2013. Mbti personality types of project managers and their success: A field survey. *Project Management Journal*, 44(3):78–87.

Jiaxi Cui, Liuzhenghao Lv, Jing Wen, Rongsheng Wang, Jing Tang, YongHong Tian, and Li Yuan. 2023. Machine mindset: An mbti exploration of large language models. *Preprint*, arXiv:2312.12999.

Joseph L. Fleiss. 2006. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, page 378–382.

Debanjan Ghosal, Bodhisattwa Prasad Majumder, Soujanya Poria, Alexander Gelbukh, and Erik Cambria. 2020. Mime: Mimicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7645–7655.

Seungju Han, Jack Hessel, Nouha Dziri, Yejin Choi, and Youngjae Yu. 2023. Champagne: Learning real-world conversation from large-scale web videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15498–15509.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2020. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI Conference on Artificial Intelligence*.

Carl Jung and John Beebe. 2016. *Psychological types*. Routledge.

Apoorv Kulshreshtha, Daniel De Freitas Adiwardana, David Richard So, Gaurav Nemade, Jamie Hall, Noah Fiedel, Quoc V. Le, Romal Thoppilan, Thang Luong, Yifeng Lu, and Zi Yang. 2020. Towards a human-like open-domain chatbot. In *arXiv*.

Yoon Kyung Lee, Jina Suh, Hongli Zhan, Junyi Jessy Li, and Desmond C. Ong. 2024. Large language models produce responses perceived to be empathic.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022a. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

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, Yizhe Zhang, Chenyang Liang, Nan Li, and Jianfeng Li. 2021. Knowledge bridging for empathetic dialogue generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 15727–15735.

Yunshui Li, Binyuan Hui, ZhiChao Yin, Min Yang, Fei Huang, and Yongbin Li. 2023. PaCE: Unified multi-modal dialogue pre-training with progressive and compositional experts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13402–13416, Toronto, Canada. Association for Computational Linguistics.

Zaijing Li, Fengxiao Tang, Ming Zhao, and Yusen Zhu. 2022b. EmoCaps: Emotion capsule based model for conversational emotion recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1610–1618, Dublin, Ireland. Association for Computational Linguistics.

Hailun Lian, Cheng Lu, Sunan Li, Yan Zhao, Chuangao Tang, and Yuan Zong. 2023. A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. *Entropy*, 25(10).

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*.

Karen K. Liu and Rosalind W. Picard. 2005. Embedded empathy in continuous, interactive health assessment.

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

Macarov and David. 1978. Empathy: The charismatic chimera. *Journal of Education for Social Work*, 14(3):86–92.

Alexandra Main. 2021. Cultivating empathy. *APA Monitor on Psychology*, 52(10):44–49.

Mariana Rodrigues Makiuchi, Kuniaki Uto, and Koichi Shinoda. 2021. Multimodal emotion recognition with high-level speech and text features. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 350–357.

OpenAI. 2023a. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2023b. Reproducible outputs. `https://platform.openai.com/docs/guides/text-generation/reproducible-outputs`.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Aravind Sesagiri Raamkumar and Yinping Yang. 2023. Empathetic conversational systems: A review of current advances, gaps, and opportunities. *IEEE Transactions on Affective Computing*, 14(4):2722–2739.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners.

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.

Gregorius Ryan, Pricillia Katarina, and Derwin Suhartono. 2023. Mbti personality prediction using machine learning and smote for balancing data based on statement sentences. *Information*, 14(4).

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

Tao Shi and Shao-Lun Huang. 2023. MultiEMO: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14752–14766, Toronto, Canada. Association for Computational Linguistics.

Haoyu Song, Yan Wang, Kaiyan Zhang, Wei-Nan Zhang, and Ting Liu. 2021a. BoB: BERT over BERT for training persona-based dialogue models from limited personalized data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–177, Online. Association for Computational Linguistics.

Shuangyong Song, Chao Wang, Haiqing Chen, and Huan Chen. 2021b. An emotional comfort framework for improving user satisfaction in E-commerce customer service chatbots. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 130–137, Online. Association for Computational Linguistics.

Dhruv Srivastava, Aditya Kumar Singh, and Makarand Tapaswi. 2023. How you feelin'? learning emotions and mental states in movie scenes. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2517–2528.

Deeksha Varshney, Asif Ekbal, and Pushpak Bhattacharyya. 2021. Modelling context emotions using multi-task learning for emotion controlled dialog generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2919–2931, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Lanrui Wang, Jiangnan Li, Zheng Lin, Fandong Meng, Chenxu Yang, Weiping Wang, and Jie Zhou. 2022. Empathetic dialogue generation via sensitive emotion recognition and sensible knowledge selection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4634–4645, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhiyuan WEN, Jiannong CAO, Ruosong YANG, Shuaiqi LIU, and Jiaxing SHEN. 2021. Automatically select emotion for response via personality-affected emotion transition. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5010–5020, United States. Association for Computational Linguistics (ACL).

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022. Long time no see! open-domain conversation with long-term persona memory. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2639–2650, Dublin, Ireland. Association for Computational Linguistics.

Emmanuelle Zech and Bernard Rimé. 2005. Is talking about an emotional experience helpful? effects on emotional recovery and perceived benefits. *Clinical Psychology amp; Psychotherapy*, page 270–287.

Weixiang Zhao, Yanyan Zhao, Xin Lu, and Bing Qin. 2023. 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.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020a. Towards persona-based empathetic conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6556–6566, Online. Association for Computational Linguistics.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020b. Towards persona-based empathetic conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6556–6566, Online. Association for Computational Linguistics.

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.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of XiaoIce, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.

A Appendix

A.1 Personality Descriptions

We obtain the description of each personality type from this website [2], and the detailed descriptions are provided in Table 4.

A.2 Human Evaluation Details

We rigorously follow the human evaluation protocols and standards set by previous studies in this domain. To assess the responses generated by different models, we engage five independent graduate students (with an average age of 25.6 years, including two from Asia, two from North America and one from Europe) who have no conflict of interest with the authors. We obtain their consent to participate and provide compensation equivalent to the standard local hourly wages.

The quality of responses generated by all models is evaluated based on three aspects: empathy, relevance, and fluency. We randomly select 100 response pairs from various models and instruct the annotators to rate each response according to these criteria. The specific instructions provided to the annotators are presented in Figure 6, and the ratings are given on a scale from 1 to 5.

To perform aspect-based pairwise comparisons, the annotators are randomly presented with two distinct responses for a given dialogue context: one produced by our model and the other by an baseline model. During both the rating and aspect-based pairwise comparison stages, we ensure that the annotators remain blind to which response was generated by our model or any other model. Furthermore, in the aspect-based pairwise comparison stage, the presentation order of the two generated responses to the annotators is randomized.

Additionally, we incorporate attention checkers to enhance the quality of data collected during human evaluation. Specifically, we embed optional 'skip' choices at two random locations within each questionnaire. These points prompt the annotators to select the predefined 'skip' option on the questionnaire page.

A.3 GPT-4 Evaluation Details

We score the randomly sampled responses using the GPT-4 API by setting the temperature to 0 in the API parameters, consistent with the instructions illustrated in Figure 6.

And the specific prompt provided to GPT-4 is as follows: Please follow the instructions and accomplish your task. For each question, you should directly return the score. After all questions are scored, you should provide a list for each aspect that contains all the scores belonging to that aspect. Finally, you should calculate the average score for each aspect.

A.4 Ethics Considerations

The datasets cited in our paper are publicly available, and ethical considerations should have been taken into account when these datasets were published. Besides, we make sure the anonymization in the human evaluation process. We assert that our research adheres to the ethical guidelines.

[2] https://www.16personalities.com/

Table 4: The 16 personalities and their corresponding descriptions.

Personality	Description
INTJ	INTJ is a personality type with the Introverted, Intuitive, Thinking, and Judging traits. These thoughtful tacticians love perfecting the details of life, applying creativity and rationality to everything they do. Their inner world is often a private, complex one.
INTP	INTP is a personality type with the Introverted, Intuitive, Thinking, and Prospecting traits. These flexible thinkers enjoy taking an unconventional approach to many aspects of life. They often seek out unlikely paths, mixing willingness to experiment with personal creativity.
ENTJ	ENTJ is a personality type with the Extraverted, Intuitive, Thinking, and Judging traits. They are decisive people who love momentum and accomplishment. They gather information to construct their creative visions but rarely hesitate for long before acting on them.
ENFP	ENTP is a personality type with the Extraverted, Intuitive, Thinking, and Prospecting traits. They tend to be bold and creative, deconstructing and rebuilding ideas with great mental agility. They pursue their goals vigorously despite any resistance they might encounter.
INFJ	INFJ is a personality type with the Introverted, Intuitive, Feeling, and Judging traits. They tend to approach life with deep thoughtfulness and imagination. Their inner vision, personal values, and a quiet, principled version of humanism guide them in all things.
INFP	INFP is a personality type with the Introverted, Intuitive, Feeling, and Prospecting traits. These rare personality types tend to be quiet, open-minded, and imaginative, and they apply a caring and creative approach to everything they do.
ENFJ	ENFJ is a personality type with the Extraverted, Intuitive, Feeling, and Judging traits. These warm, forthright types love helping others, and they tend to have strong ideas and values. They back their perspective with the creative energy to achieve their goals.
ENFP	ENFP is a personality type with the Extraverted, Intuitive, Feeling, and Prospecting traits. These people tend to embrace big ideas and actions that reflect their sense of hope and goodwill toward others. Their vibrant energy can flow in many directions.
ISTJ	ISTJ is a personality type with the Introverted, Observant, Thinking, and Judging traits. These people tend to be reserved yet willful, with a rational outlook on life. They compose their actions carefully and carry them out with methodical purpose.
ISFJ	ISFJ is a personality type with the Introverted, Observant, Feeling, and Judging traits. These people tend to be warm and unassuming in their own steady way. They're efficient and responsible, giving careful attention to practical details in their daily lives.
ESTJ	ESTJ is a personality type with the Extraverted, Observant, Thinking, and Judging traits. They possess great fortitude, emphatically following their own sensible judgment. They often serve as a stabilizing force among others, able to offer solid direction amid adversity.
ESFJ	ESFJ is a personality type with the Extraverted, Observant, Feeling, and Judging traits. They are attentive and people-focused, and they enjoy taking part in their social community. Their achievements are guided by decisive values, and they willingly offer guidance to others.
ISTP	ISTP is a personality type with the Introverted, Observant, Thinking, and Prospecting traits. They tend to have an individualistic mindset, pursuing goals without needing much external connection. They engage in life with inquisitiveness and personal skill, varying their approach as needed.
ISFP	ISFP is a personality type with the Introverted, Observant, Feeling, and Prospecting traits. They tend to have open minds, approaching life, new experiences, and people with grounded warmth. Their ability to stay in the moment helps them uncover exciting potentials.
ESTP	ESTP is a personality type with the Extraverted, Observant, Thinking, and Prospecting traits. They tend to be energetic and action-oriented, deftly navigating whatever is in front of them. They love uncovering life's opportunities, whether socializing with others or in more personal pursuits.
ESFP	ESFP is a personality type with the Extraverted, Observant, Feeling, and Prospecting traits. These people love vibrant experiences, engaging in life eagerly and taking pleasure in discovering the unknown. They can be very social, often encouraging others into shared activities.

Empathetic Response Evaluation

We are a team of researchers specializing in natural language processing focused on generating empathetic responses. Below are several dialogue contexts and corresponding responses. Please assess each pair based on the following three principles present as blow.

Context: Why do all you're coffee mugs have numbers on the bottom?

Response: What is that? That might be interesting to look up.

* **Empathy:** whether the response empathizes, comprehends the emotions of others, and approaches and resolves issues from the perspective of the other party.

- 1: Completely not empathetic, potentially offensive, or likely to evoke negative emotions in the speaker.
- 2: Slightly empathetic, containing few words expressing understanding or offering help.
- 3: Empathetic, acknowledges the emotion and demonstrates understanding, but lacks depth in addressing it.
- 4: Moderately empathetic, acknowledging the speaker's emotions and interpreting their experience to some extent.
- 5: Highly empathetic, explicitly identifying the speaker's feelings or experiences, probing key questions about the situation, and providing substantial assistance.

* **Coherence:** whether the response aligns with the dialogue history and is consistent with the speaker's background situation.

- 1: Completely irrelevant to the context, or inconsistent with the dialogue history or background situation.
- 2: Slightly coherent to the context, but featuring numerous conflicts with the dialogue history and background situation.
- 3: Coherent to the context, but with some conflicts to the dialogue history or background situation.
- 4: Moderately coherent to the context, but with minor conflicts to the dialogue history or background situation.
- 5: Completely coherent and relevant to the context and background situation.

* **Fluency:** whether the response flows smoothly in a natural and linguistically correct manner, with proper use of grammar, vocabulary, and syntax.

- 1: Not fluent, and fails to communicate a coherent or understandable message.
- 2: Slightly fluent, featuring basic understandable communication, but hindered by unclear expressions.
- 3: Moderately fluent, with the response being understandable and somewhat natural, but marked by frequent awkward phrasing or inconsistencies that interfere with the clarity or logical progression of ideas.
- 4: Fluent, with a smooth and logical flow, but marred by occasional awkward or unclear expressions that disrupt communication.
- 5: Completely fluent, demonstrates seamless and natural communication that aligns perfectly with humans.

Figure 6: An example of our questionnaire for the human evaluation.