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

Jiaqiang Wu, Xuandong Huang, Zhouan Zhu, Shangfei Wang[*]
University of Science and Technology of China
{jqwu,xuandong,zza2021}@mail.ustc.edu.cn
sfwang@ustc.edu.cn

Abstract

Empathetic dialogue systems improve user experience across various domains. Existing approaches mainly focus on acquiring affective and cognitive knowledge from text, but neglect the unique personality traits of individuals and the inherently multimodal nature of human face-to-face conversation. To this end, we enhance the dialogue system[1] with the ability to generate empathetic responses from a multimodal perspective, and consider the diverse personality traits of users. We incorporate multimodal data, such as images and texts, to understand the user's emotional state and situation. Concretely, we first identify the user's personality trait. Then, the dialogue system comprehends the user's emotions and situation by the analysis of multimodal inputs. Finally, the response generator models the correlations among the personality, emotion, and multimodal data, to generate empathetic responses. Experiments on the MELD dataset and the MEDIC dataset validate the effectiveness of the proposed approach.

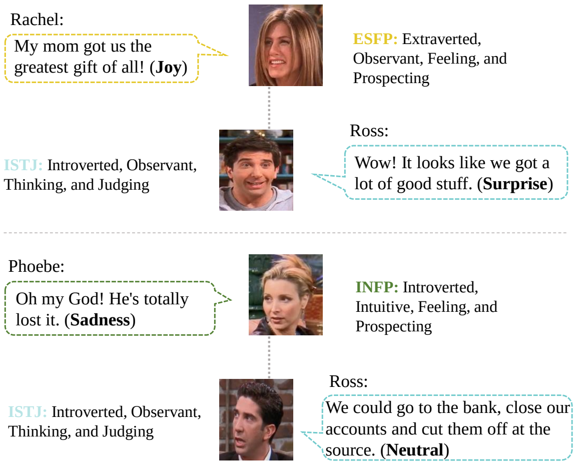

Rachel:
My mom got us the greatest gift of all! (Joy)
ESFP: Extraverted, Observant, Feeling, and Prospecting
Ross:
ISTJ: Introverted, Observant, Thinking, and Judging
Wow! It looks like we got a lot of good stuff. (Surprise)
Phoebe:
Oh my God! He's totally lost it. (Sadness)
INFP: Introverted, Intuitive, Feeling, and Prospecting
Ross:
ISTJ: Introverted, Observant, Thinking, and Judging
We could go to the bank, close our accounts and cut them off at the source. (Neutral)

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, Ross considers the distinct personality traits of each speaker in his interactions, which facilitates his ability to achieve empathy with them.

1 Introduction

Empathy is 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; Liu and Picard, 2005). The advent of the EmpatheticDialogues dataset (Rashkin et al., 2019) attracts much 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 works focus on discerning the user's 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 proposed to apprehend the user's emotions at utterance level, including mixture of empathetic listeners (Lin et al., 2019), emotion mimicry (Ghosal et al., 2020), while others examined strategies to model the user's 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 information directly from dialogues (Sabour et al., 2022; Wang et al., 2022; Zhou et al., 2023), which function as prior knowledge and guide dialogue systems in generating responses that are more relevant and consistent. Recently, the newly introduced large

[*] Corresponding author.

[1] Our code is available at: https://github.com/personalityempathy/Personality-Aware-MERG

Proceedings of the 31st International Conference on Computational Linguistics, pages 8925–8938
January 19–24, 2025. ©2025 Association for Computational Linguistics

language models (LLMs), such as GPT 4 (OpenAI, 2023) and Claude 3 (Anthropic, 2023), which demonstrate 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, the aforementioned empathetic studies ignore the influence of the user's 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). In conversations, individuals not only resort to their habitual modes of expressing empathy but also tailor their responses to match the personality traits of their interlocutors (Chae, 2016).

Therefore, we propose a multimodal dialogue system that is attentive to personality intricacies and can produce targeted empathetic responses. To achieve this, we utilize a pre-trained MBTI classifier (Ryan et al., 2023) to infer the user's personality from the 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 features (Li et al., 2022). A cross-modal feature fusion module integrates the multimodal features, which emphasizes relevant image aspects in the context of the dialogue and ensures that the features are well-optimized for the response generation stage.

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

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

(2) We acquire the affective and cognitive knowledge in human face-to-face conversations from a multimodal perspective to achieve empathy.

2 Related Work

2.1 Personalized Response Generation

Personalization can enhance the user's engagement with dialogue systems (Kwon et al., 2023). Researchers have explored various ways to represent persona information, such as unstructured persona descriptions (Zhang et al., 2018; Zhong et al., 2020a; Ahn et al., 2023), structured key-value attributes (Qian et al., 2018; Wu et al., 2021; Gao et al., 2023), specific personalities (e.g., MBTI, Big Five personalities) (Mairesse and Walker, 2007; Wen et al., 2021; Fernau et al., 2022), and dialogue histories (Qian et al., 2021; Zeng and Nie, 2021). These works concentrate on generating responses through the dialogue history and personalities to increase the personalization of the dialogue, rather than focusing on empathy.

Despite considerable efforts dedicated to the development of persona-based dialogue models (Zhong et al., 2020b; 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 personality traits (Zhong et al., 2020a; Ahn et al., 2023). However, our work not only considers personalities but also emphasizes empathetic responses to the user within a multimodal context.

2.2 Empathetic Response Generation

Empathetic response generation necessitates that dialogue systems understand the user's emotions and situation (Li et al., 2021), so that generate pertinent responses and achieve empathy with the user. The seminal work of Rashkin et al. (2019) introduces the task and establishes the benchmark dataset, which has catalyzed heightened interest in this area. Some works endeavor to endow dialogue systems with the capability to comprehend affective knowledge via emotion perception. Lin et al. (2019) employed n encoders to identifying emotions with a specific category (**MoEL**). Ghosal et al. (2020) divided emotions into two groups according to their polarity and integrate emotions with stochasticity (**MIME**). Li et al. (2020) identi-

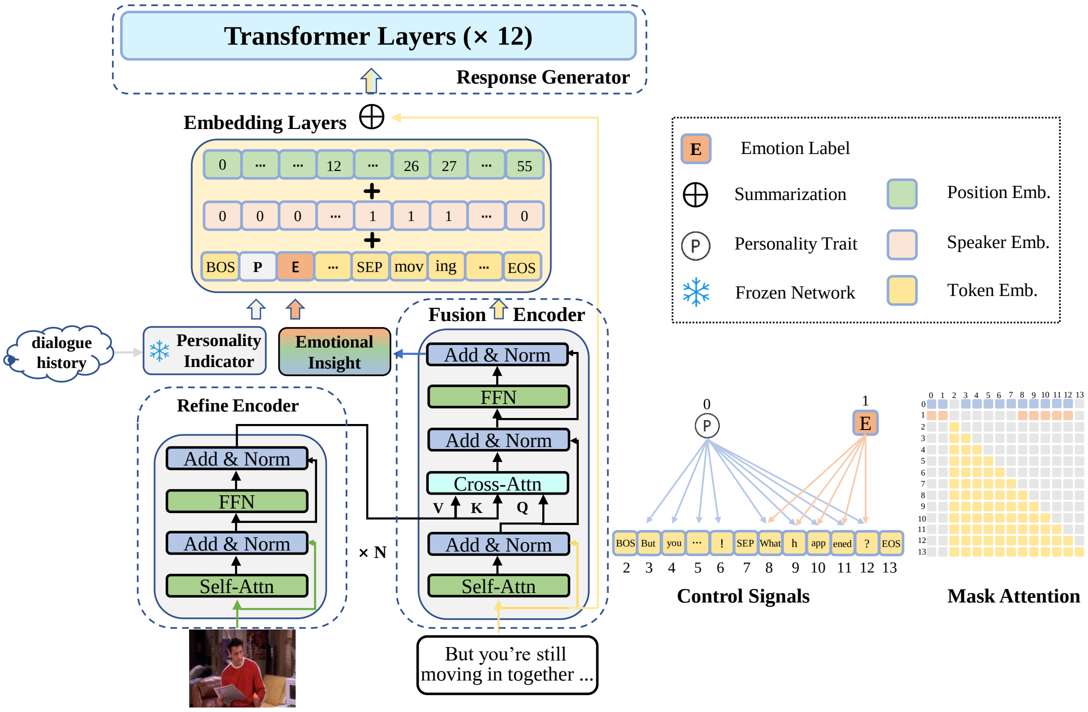


Figure 2: Overview of the proposed framework. The visual features refined by a specialized encoder, are integrated with textual features in a cross-modal fusion encoder for multimodal emotion recognition. The incorporation of personalities, emotional labels, and multimodal representations augments the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

fied 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 the user's situation. Sabour et al. (2022) fed the dialogue history to Comet (Bosselut et al., 2019), and obtain inferences from five distinct aspects (**CEM**). Wang et al. (2022) addressed the challenge of capturing dynamic emotional shifts in conversations, as well as the potential discrepancies between knowledge graph inferences and the emotions (**SEEK**). Zhao et al. (2023) proposed a framework, consisting of self-other differentiation and modulation mechanism, and a response generator (**EmpSOA**). (Zhou et al., 2023) constructed the cognition graph utilizing inferred knowledge and the emotional concept graph to align the user's cognitive and affective knowledge (**CASE**). Yang et al. (2024) proposed an iterative interaction attention mechanism to identify semantically related words within dialogue utterances, thereby facilitating a deeper understanding of underlying emotional and cognitive states.

In summary, previous studies extract the user's emotional states and situations from a solely textual perspective, but they restrict the deeper understanding that dialogue systems can reach regarding speakers. Differently, our work explores affective and cognitive knowledge related to the user from the perspectives of both personality and multimodality, to achieve empathy.

3 Problem Statement

We denote a dialogue context as a sequence of n paired utterances-images, denoted as $U = \{u_1, \ldots, u_n\}$, where $u_i = \{u_i^t, u_i^v\}, i \in [1, n]$. u_i^t and u_i^v denoting textual and visual data of each paired utterance-image. $\hat{P} = \{p_1, p_2\}$ represents the set of personality traits associated with the two speakers engaged in a conversation. Besides, $u_i^t = \{w_i^1, \ldots, w_i^k\}$ denotes that the utterance u_i^t consists of k words. k vary from various utterances. Each utterance is provided with an emotion label. The task is to train a model to generate the next utterance Y that is coherent to the dialogue context U and empathetic to the other speaker's personality, emotions and situation.

4 Methodology

Our proposed personality-aware framework is present in Figure 2, which mainly incorporates a cross-modal fusion encoder for multimodal emotional insights, a pre-trained MBTI personality classifier as the personality indicator, and an empathetic response generator. Different special tokens are shown in token samples. For example, the BOS token and EOS token indicate the beginning and the end of an utterance, and the SEP token separates the utterances.

4.1 Multimodal Emotional Insights

To understand the speaker's emotional states from the dialogue history, we employ multimodal emotion recognition techniques. Specifically, for each multimodal paired data $\{u^t, u^v\}$, we utilize the pre-trained BLIP model and the pre-trained GPT-2 model 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 $W_q, W_k, W_v \in \mathbb{R}^{d \times d_k}$ represents learnable parameter matrices, $Q_{r^v}, K_{r^v}, V_{r^v}$ are the query, key and value matrices, and d_k is the dimension of attention layers. Then, the self-attention mechanism encodes the visual features by matching their query and key matrices, which is calculated by Equation 2:

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

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, 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 attention mechanism 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:

$$A_{tv} = \sigma\left(\frac{A_{r^t}\, A_{r^v}^T}{\sqrt{d_k}}\right) A_{r^v} \tag{3}$$

where $A_{r^v}^T$ is the transposed key matrix of A_{r^v}. 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 $\tilde{e}$, formalized by Equation 4:

$$\tilde{e} = \sigma\left(LN\left(W_h H\right)\right) \tag{4}$$

where LN represents the linear layers within the classifier, $W_h \in \mathbb{R}^{C \times d}$ is learnable parameters, C is the number of emotion categories, and $\tilde{e}$ indicates the predicted emotion label. Therefore, we calculate the loss of the multimodal emotion recognition by Equation 5:

$$\mathcal{L}_{\tilde{e}} = -\frac{1}{\sum_{h=1}^{m} f(h)} \sum_{j=1}^{m} \sum_{i=1}^{f(j)} e_{ji} \log(\tilde{e}) \tag{5}$$

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, e_{ij} represents the ground truth emotion label.

4.2 Personality Indicator

We employ a pre-trained MBTI personality classifier $\mathcal{C}$, which achieves an average classification accuracy of 84.34% on Kaggle's MBTI dataset[2] (Ryan et al., 2023), to infer personality traits for each speaker in conversations. We begin by grouping the utterances in the conversation by speakers. For a given speaker s, we concatenate the utterances to form a set $U_s = \{u_{s1}, u_{s2}, \cdots\}$, which serves as the input to the personality classifier $\mathcal{C}$. The classifier then predict 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 section A.1. In the experiments, we prepend a CLS token to each description, creating $\widetilde{\mathcal{R}} = [[CLS]; \mathcal{R}]$. We then input $\widetilde{\mathcal{R}}$ into the pre-trained 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 p_s collaboratively control the generation process. We differentiate between tokens that serve as control signals and those that form dialogues. As shown in the right part of Figure 2, we model their relationship with a mask matrix W_m during

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

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:

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

This mechanism allows us to use the mask matrix to guide the generation of each response token using signals from various perspectives, representing diverse 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:

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

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 pre-trained model, which can be formulated as:

$$I = W^G h^G + W^H H \tag{8}$$

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 pre-trained 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:

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

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:

$$\mathcal{L} = \lambda \mathcal{L}_Y + \gamma \mathcal{L}_E \tag{10}$$

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 MEDIC dataset (Zhu et al., 2023). Both of them include multiple conversations and multimodal data, such as video, and text data. We use the original partitioning of the two datasets for training, validation and testing.

MELD is a widely-used multimodal dataset. It contains over 1,400 dialogues and 13,000 utterances that are sampled from the TV series *Friends*. Each utterance is annotated with 8 emotions and 3 sentiment categories.

MEDIC is an empathetic dataset designed to advance computational empathy understanding in the context of face-to-face psychological counseling sessions. The dataset comprises 771 video clips collected from the counseling sessions.

Datasets	Methods	Automotic Evaluation							Human Evaluation		
		PPL ↓	**Dist-1**	**Dist-2**	**Acc (%)**	P_{BERT}	R_{BERT}	F_{BERT}	**Emp.**	**Coh.**	**Flu.**
MELD	MoEL(Lin et al., 2019)	50.41	0.71	3.22	57.93	0.8098	0.8175	0.8136	2.91	3.09	3.37
	MIME(Ghosal et al., 2020)	48.50	0.64	2.88	56.90	0.7950	0.8029	0.7989	2.88	3.14	3.34
	EmpDG(Li et al., 2020)	50.51	0.89	4.05	57.62	0.8020	0.8159	0.8088	2.95	3.22	3.42
	CEM(Sabour et al., 2022)	54.00	0.97	4.36	57.55	0.7877	0.8040	0.7957	3.02	3.27	3.65
	SEEK(Wang et al., 2022)	54.72	1.01	4.54	58.95	0.8133	0.8297	0.8214	3.11	3.24	3.58
	EmpSOA(Zhao et al., 2023)	53.33	1.02	4.60	59.69	0.8271	0.8468	0.8368	3.13	3.28	3.61
	CASE(Zhou et al., 2023)	55.27	1.05	4.68	58.84	0.8159	0.8391	0.8273	3.12	3.25	3.63
	Ours	**35.38**	**2.12**	**8.38**	**67.05**	**0.8472**	**0.8516**	**0.8494**	**3.26**	**3.43**	**3.71**
	GPT-4-V Turbo (**LLM**)	-	5.92	36.58	60.35	0.8532	0.8305	0.8416	3.65	4.03	4.60
MEDIC	MoEL(Lin et al., 2019)	36.86	1.40	2.83	-	0.8109	0.8292	0.8199	3.01	3.02	3.34
	MIME(Ghosal et al., 2020)	36.48	1.17	4.14	-	0.8143	0.8293	0.8217	3.10	3.09	3.30
	EmpDG(Li et al., 2020)	35.80	1.08	4.10	-	0.8185	0.8375	0.8278	3.02	3.17	3.39
	CEM(Sabour et al., 2022)	36.17	1.58	5.77	-	0.8262	0.8310	0.8285	3.13	3.21	3.50
	SEEK(Wang et al., 2022)	36.91	1.89	6.81	-	0.8271	0.8424	0.8346	3.17	3.19	3.53
	EmpSOA(Zhao et al., 2023)	34.56	1.95	7.08	-	0.8211	0.8340	0.8274	3.20	3.29	3.58
	CASE(Zhou et al., 2023)	36.02	1.93	7.15	-	0.8309	0.8391	0.8349	3.18	3.26	3.57
	Ours	**29.47**	**2.93**	**10.46**	-	**0.8461**	**0.8548**	**0.8504**	**3.36**	**3.35**	**3.62**
	GPT-4-V Turbo (**LLM**)	-	5.90	35.91	-	0.7521	0.7355	0.7437	3.71	3.95	4.65

Table 1: Evaluations of our method and the baselines. Numbers in bold indicate that the improvement of the method is statistically significant (paired t-test with p-value $<$ 0.05).

5.2 Implementation Details

All codes are implemented with PyTorch. To build the framework, we incorporate the pre-trained BLIP model (Li et al., 2022) 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. For the training phase, we utilize two NVIDIA Geforce RTX 3090 GPU cards equipped with 24 GB RAM of each, and we maintain the training state until it becomes apparent that there is no additional decrease in loss achievable. 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 comparative analysis, we adhere to the original settings of official codes from all methods under consideration. All compared methods follow the same experimental procedure as ours. We also prompt GPT-4 Turbo (OpenAI, 2023) under zero-shot conditions with a temperature setting of 1.0 for comparisons. The detailed prompt is shown in section A.3.

5.3 Automatic Evaluation

Following the previous works (Sabour et al., 2022; Wang et al., 2022; Zhao et al., 2023), our evaluation employs automatic metrics: Lower values of **PPL** denote higher quality; higher scores of **Dist-1** and **Dist-2** indicates greater diversity. We leverage BERTScore ($P_{BERT}, R_{BERT}, F_{BERT},$) (Zhang et al., 2020) to evaluate the semantic similarity between generated responses and the ground truth. Additionally, we report the average prediction accuracy (**Acc.**) to provide a more holistic assessment of the models, considering that the compared methods involve emotion classification as part of their training.

Table 1 provides an experimental analysis, comparing the performance of our method with the contemporary state-of-the-art approaches. Due to the absence of prior work on multimodal empathetic response generation, for fairness, we select the GPT-4-V Turbo as the multimodal LLM (OpenAI, 2023) for comparison.

Compared with the best performance of non LLM-based methods, our method obtains apparent improvement in response quality, diversity and similarity. Concretely, we achieves 27.1%/14.7% in relative, 13.12/5.09 in absolute for **PPL**, and 79.1%/46.3% in relative, 3.70/3.31 in absolute for **Dist-2**, and 1.5%/1.9% in relative 0.013/0.015 in absolute for F_{BERT}. The improvements show that our method generates more relevant empathetic responses rich in diversity, as much affective and cognitive information is provided. Besides, the significant promotion of emotion recognition accuracy, achieving 12.3% in relative, 7.36 in absolute for **Acc**, indicates that by fusing visual and textual modalities, our model is better equipped to capture the nuanced nature of human emotions, which enables the model to establish more empathetic connections with users.

Compared with the performance of the LLM-based method (GPT-4-V Turbo), our method lags behind a lot, which is likely due to the huge training data of the LLM. Our model, with a mere 822M parameters, demonstrates highly competitive performance when benchmarked against the GPT-4-V Turbo with hundreds of billions of parameters. Additionally, while the LLM exhibit inferior performance on the MEDIC dataset in terms of BERTScore compared to their results on the MELD dataset, possibly owing to the absence of data from the MEDIC dataset in their training, and our model presents a more competitive performance on BERTScore.

5.4 Human 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 200 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 7 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. More details are in section A.2.

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. The outcomes, detailed in Table 3, reveal a clear preference for responses generated by our model, underscoring its empathetic response capabilities.

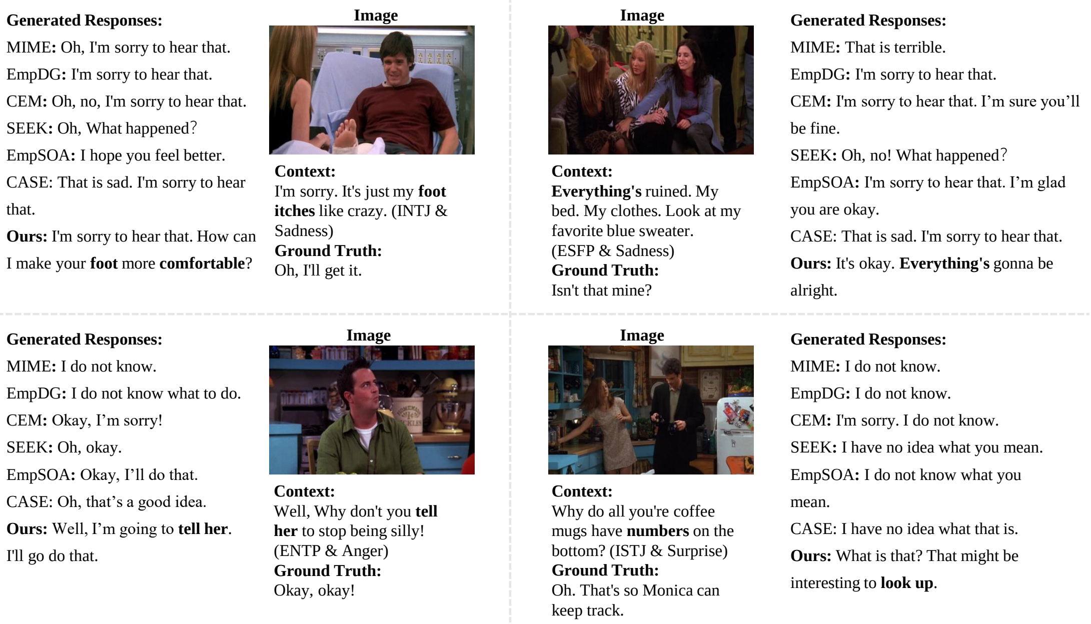
Datasets	Ablation	**PPL↓**	**Dist-1**	**Dist-2**	**Acc (%)**	$P.$	$R.$	$F.$
MELD	Ours	**35.38**	**2.12**	**9.83**	**67.05**	**0.8472**	**0.8516**	**0.8494**
	w/o P	36.92	1.54	6.38	65.94	0.8433	0.8502	0.8467
	w/o V	38.14	1.47	6.04	61.98	0.8186	0.8247	0.8216
	w/o P&V	40.25	1.04	4.75	61.20	0.8035	0.8190	0.8112
	w/o mask	40.09	1.61	6.24	65.45	0.8319	0.8403	0.8361
	w/o residual	46.58	1.72	6.46	64.21	0.8158	0.8210	0.8184
MEDIC	Ours	**30.64**	**5.63**	**20.46**	-	**0.8461**	**0.8548**	**0.8504**
	w/o P	30.91	3.95	14.52	-	0.8391	0.8472	0.8431
	w/o V	31.23	3.76	14.08	-	0.8144	0.8285	0.8213
	w/o P&V	33.95	3.48	12.83	-	0.8102	0.8258	0.8179
	w/o mask	33.46	3.99	14.55	-	0.8307	0.8394	0.8350
	w/o residual	38.28	4.21	15.26	-	0.8116	0.8213	0.8164

Table 2: P represents personalities, V is the visual input, mask and residual indicate the mask matrix and the residual connection. P., R., F., represents P_{BERT}, R_{BERT}, and R_{BERT} respectively.

The results presented in Table 1 and Table 3 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.

5.5 Ablation Study

As illustrated in Table 2, to substantiate the essential roles of the components in our framework, we remove each newly introduced parts within the framework. Removing personality and visual input (**w/o** P, **w/o** V, and **w/o** P&V) significantly reduces the diversity of responses, especially the removal of visual data results in a substantial decrease in emotion recognition accuracy and a lower similarity to the ground truth, as both the variants provide much information about the dialogue context. Concretely, the personality provides personal information about the user, and the visual input conveys the user's situation and facial expression. Omitting the masking operation (**w/o** mask) results in declining automatic metrics, because the model treats personality and emotional tokens identically to standard tokens, which can disrupt the autoregressive language generation process due to these additional tokens. Removing the residual connection (**w/o** residual) leads to the decreased response quality and diversity, which indicates that the response



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

Generated Responses:
MIME: I do not know.
EmpDG: I do not know what to do.
CEM: Okay, I'm sorry!
SEEK: Oh, okay.
EmpSOA: Okay, I'll do that.
CASE: 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:
MIME: I do not know.
EmpDG: I do not know.
CEM: I'm sorry. I do not know.
SEEK: I have no idea what you mean.
EmpSOA: I do not know what you mean.
CASE: 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.

Comparisons	Aspects	Win	Lose	Tie
Ours vs. MoEL	Emp.	**56.2**	33.5	10.3
	Coh.	**52.8**	30.4	16.8
	Flu.	**46.4**	35.7	17.9
Ours vs. MIME	Emp.	**57.8**	33.2	9.0
	Coh.	**52.9**	31.4	15.7
	Flu.	**46.2**	34.6	19.2
Ours vs. EmpDG	Emp.	**51.7**	35.8	12.5
	Coh.	**49.1**	34.5	16.4
	Flu.	**48.3**	30.0	21.7
Ours vs. CEM	Emp.	**48.4**	32.3	19.3
	Coh.	**53.3**	37.4	9.3
	Flu.	**47.2**	40.2	12.6
Ours vs. SEEK	Emp.	**52.6**	30.9	16.5
	Coh.	**50.4**	38.7	10.9
	Flu.	**48.6**	41.6	9.8
Ours vs. EmpSOA	Emp.	**49.5**	31.1	19.4
	Coh.	**52.1**	36.5	11.4
	Flu.	**50.7**	39.7	9.6
Ours vs. CASE	Emp.	**49.5**	31.1	19.4
	Coh.	**52.1**	36.5	11.4
	Flu.	**50.7**	39.7	9.6

Table 3: Results of aspect-based pair comparisons (the statistical significance (t-test) with p-value < 0.05).

generator can leverage the exceptional skills of the pre-trained model in language generation through the residual connection. Notably, in the ablations conducted to evaluate the contributions of personality and visual data, the original positions of these inputs within the input sequences are replaced with randomly initialized embedding vectors without modifying the model architecture, which ensures that the observed effects could be attributed solely to the absence of the ablated features.

6 Case Study

Personality information serves to regulate the expression of empathy, which enables the dialogue system to adjust the style of empathetic responses based on the traits of the interlocutor. For instance, when interacting with an individual with a more rational personality, the system can generate responses that are concise and logically structured. For emotionally-oriented individuals, the system can produce responses with a greater degree of emotional resonance.

We exhibit cases across four scenarios in Figure 3, showing empathetic responses generated by our model and the baselines. Specifically, in the top-left, the speaker is characterized by the INTJ personality, marked by a reluctance to express sentiments. Our model empathizes towards the speaker's itchy condition and introverted nature, and proposes to alleviate the discomfort. In the top-right, the speaker is identified with the ESFP personality, demonstrating a willingness to share feelings. The baselines produce general comforting replies, but our model responds with more relevant information. In the bottom-left, the speaker is exemplified as the ENTP personality, characterized by tenacity to achieve goals. Among the generated responses, only SEEK and our model respond relevantly with the speaker's aspirations. In the bottom-right, the speaker is portrayed as the ISTJ personality, known for their thoughtful and inquisitive trait. The baselines' responses showcase a lack of engagement. But our model follows the cue of questioning by proposing to look up the number. These cases demonstrate that our model generates empathetic responses that align with the distinct personalities of the dialogue participants. Moreover, the responses generated by our model with randomly initialized personality embeddings, presented in order from top to bottom and left to right, are as follows: *I'm sorry to hear that.*, *I'm sorry for your loss.*, *Okay, I'm going to do that.*, and *I don't know about that.* Compared with the original responses generated by our model in Figure 3, these responses indicate that the model's ability of empathy expression deteriorates to a level comparable to previous methods. Specifically, it produces general and safe but insufficiently empathetic responses.

7 Conclusion

In this paper, we identified a gap in current methods of empathetic response generation, especially their limitations in incorporating multimodality and personality dimensions. We capitalizes on the integration of multimodal data and personality traits to attain an understanding of the speaker's emotional state and situation. Our study not only advances the field but also underscores the significance of multimodal data and personality awareness in creating more empathetic interactions.

Acknowledgments

This work has been supported by the project from the National Natural Science Foundation of China (92048203, 62376255).

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.

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.

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.

Daniel Fernau, Stefan Hillmann, Nils Feldhus, and Tim Polzehl. 2022. Towards automated dialog personalization using mbti personality indicators. In *Interspeech*.

Jingsheng Gao, Yixin Lian, Ziyi Zhou, Yuzhuo Fu, and Baoyuan Wang. 2023. LiveChat: A large-scale personalized dialogue dataset automatically constructed from live streaming. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15387–15405, Toronto, Canada. Association for Computational Linguistics.

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.

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*.

Deuksin Kwon, Sunwoo Lee, Ki Hyun Kim, Seojin Lee, Taeyoon Kim, and Eric Davis. 2023. What, when, and how to ground: Designing user persona-aware conversational agents for engaging dialogue. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 707–719, Toronto, Canada. Association for Computational Linguistics.

Yoon Kyung Lee, Jina Suh, Hongli Zhan, Junyi Jessy Li, and Desmond C. Ong. 2024. Large language models produce responses perceived to be empathic. *ArXiv*, abs/2403.18148.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. 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.

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.

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.

François Mairesse and Marilyn Walker. 2007. PERSONAGE: Personality generation for dialogue. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 496–503, Prague, Czech Republic. Association for Computational Linguistics.

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

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.

Hongjin Qian, Xiaohe Li, Hanxun Zhong, Yu Guo, Yueyuan Ma, Yutao Zhu, Zhanliang Liu, Zhicheng Dou, and Ji-Rong Wen. 2021. Pchatbot: A large-scale dataset for personalized chatbot. In *Proceedings of the SIGIR 2021*. ACM.

Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Assigning personality/profile to a chatting machine for coherent conversation generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4279–4285. International Joint Conferences on Artificial Intelligence Organization.

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

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.

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).

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, Online. Association for Computational Linguistics.

Yuwei Wu, Xuezhe Ma, and Diyi Yang. 2021. Personalized response generation via generative split memory network. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1956–1970, Online. Association for Computational Linguistics.

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.

Zhou Yang, Zhaochun Ren, Wang Yufeng, Haizhou Sun, Chao Chen, Xiaofei Zhu, and Xiangwen Liao. 2024. An iterative associative memory model for empathetic response generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3081–3092, Bangkok, Thailand. Association for Computational Linguistics.

Yan Zeng and Jian-Yun Nie. 2021. A simple and efficient multi-task learning approach for conditioned dialogue generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4927–4939, Online. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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.

Zhouan Zhu, Chenguang Li, Jicai Pan, Xin Li, Yufei Xiao, Yanan Chang, Feiyi Zheng, and Shangfei Wang. 2023. Medic: A multimodal empathy dataset in counseling. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 6054–6062, New York, NY, USA. Association for Computing Machinery.

A Appendix

A.1 Personality Descriptions

We obtain the description of each personality type from this website [3], and the detailed descriptions are provided in Table 5.

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 7 independent graduate students with 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 200 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 4, 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 another 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 Prompt Details

We prompt the GPT-4-V Turbo for comparison. The prompt we used to recognize speakers' emotions is: What is the emotion of the person in the image? Choose one from the following options: anger, disgust, sadness, happiness, neutral, surprise.

Methods	PPL↓	Dist-1	Dist-2	Acc (%)
MoEL	38.35	0.44	2.10	32.2
MIME	37.33	0.41	1.62	29.6
EmpDG	37.77	0.53	2.26	31.4
CEM	36.86	0.64	2.84	37.3
SEEK	37.09	0.73	3.23	41.9
EmpSOA	35.02	0.71	3.96	**48.3**
CASE	35.37	0.74	4.01	40.2
Ours	**32.39**	**1.65**	**6.22**	45.4
w/o P	35.66	1.01	4.87	41.5

Table 4: Evaluations of the proposed method and the baselines on the text-only EmpatheticDialogues dataset.

Besides, the prompt we used to generate empathetic responses is as follows:

1. You are an empathetic conversational agent engaged in a dialogue. The following background information provides background information for the conversation:

2. Current Speaker's Emotion: {emotion}

3. Current Speaker's Personality: {MBTI_and_Description}

4. Dialogue History: {dialogue_history}

5. Current Speaker's Utterance: {current_utterance}

6. Your task: Given the context and considering the current speaker's emotional state and personality, generate a compassionate and empathetic response that addresses the speaker's needs and feelings.

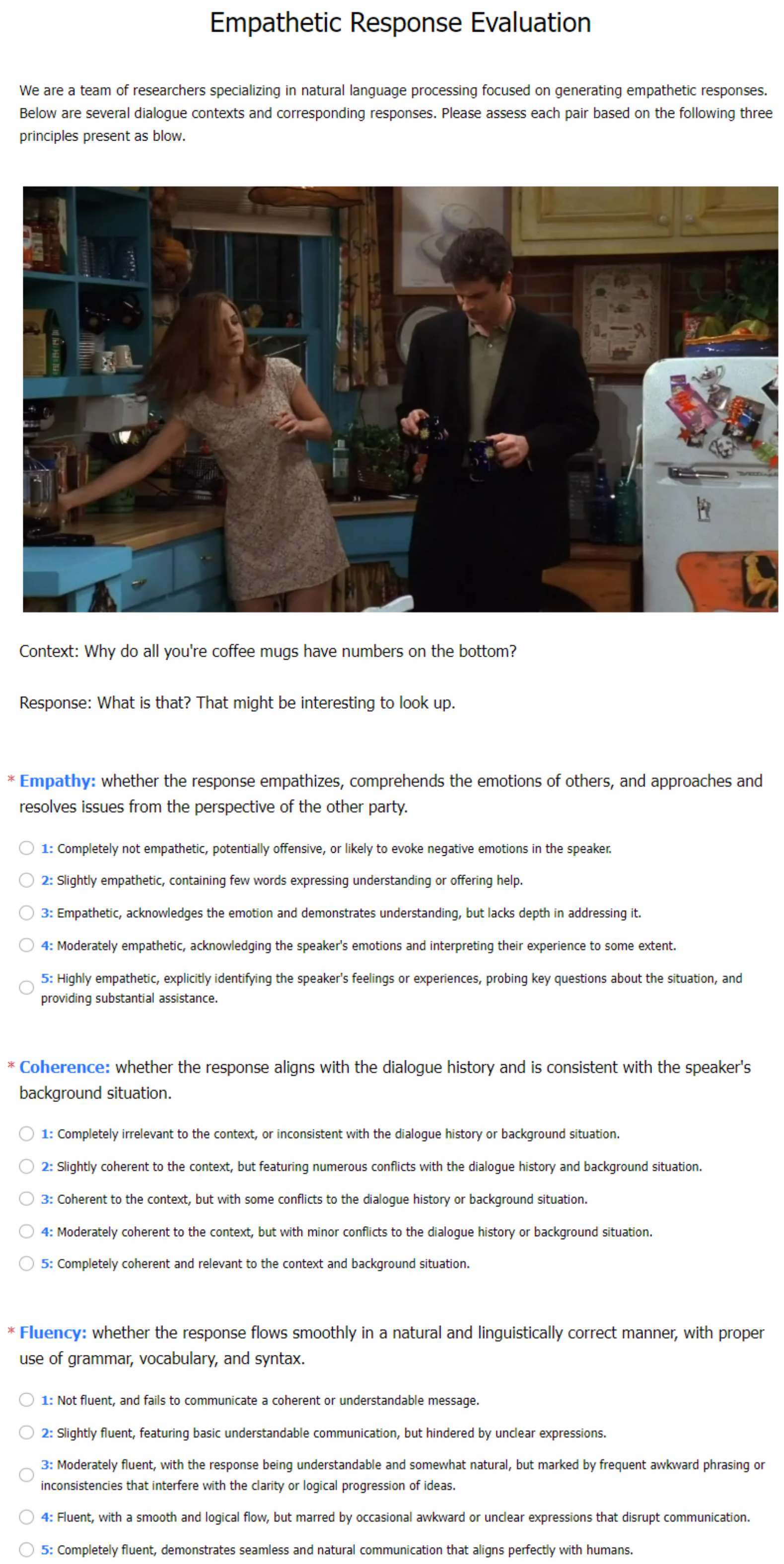
A.4 More Experimental Results

As shown in Table 4, to evaluate the performance of the proposed personality-aware method compared to existing approaches, we have conducted experiments on the text-only EmpatheticDialogues dataset (Rashkin et al., 2019), excluding the consideration of visual data. The results demonstrate that personality information can improve the model's performance on automatic evaluation metrics. However, there remains a 2.9% gap in emotion recognition accuracy between our method and EmpDG.

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

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.

Table 5: The 16 personalities and their corresponding descriptions.



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 4: An example of our questionnaire for the human evaluation.