ReflectDiffu: Reflect between Emotion-intent Contagion and Mimicry for Empathetic Response Generation via a RL-Diffusion Framework

Anonymous ACL submission

Abstract

Empathetic response generation necessitates the integration of emotional and intentional dynamics to foster meaningful interactions. Existing research either neglects the intricate interplay between emotion and intent, leading to suboptimal controllability of empathy, or resorts to large language models (LLMs), which incur significant computational overhead. In this paper, we introduce ReflectDiffu, a lightweight and comprehensive framework for empathetic response generation. This framework incorporates emotion contagion to augment emotional expressiveness and employs an emotion-reasoning mask to pinpoint critical emotional elements. Additionally, it integrates intent mimicry within reinforcement learning for refinement during diffusion. By harnessing an *intent twice* reflect mechanism of *Exploring-Sampling-Correcting*, ReflectDiffu adeptly translates emotional decision-making into precise intent actions, thereby addressing empathetic response misalignments stemming from emotional misrecognition. Through reflection, the framework maps emotional states to intents, markedly enhancing both response empathy and flexibility. Comprehensive experiments reveal that ReflectDiffu outperforms existing models regarding relevance, controllability, and informativeness, achieving state-of-the-art results in both automatic and human evaluations. Our code and datasets are available in supplementary materials.

1 Introduction

Empathetic dialogue generation endows dialogue models with human-like emotional capabilities to recognize, understand, and express emotions (Davis, 1990; Cuff et al., 2016). In psychology, empathy mechanisms are empirically linked to sociological studies on emotional contagion (Hatfield et al., 1993) and empathetic mimicry (Carr et al., 2003). Recent research has delved into various aspects of empathetic mechanisms in chatbots, including dynamically tailoring responses based on perceived emotional triggers (Gao et al., 2021, 2023) or mimicking empathetic emotions (Majumder et al., 2020; Bi et al., 2023).

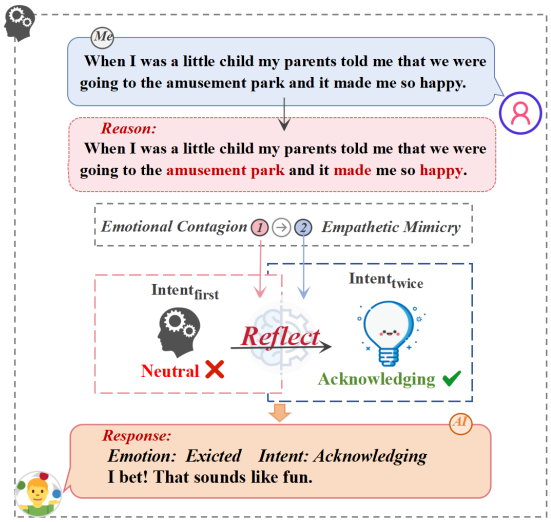

Figure 1: An example from the EMPATHETICDIALOGUES dataset involves incorporating Emotional Contagion and Mimicking, using *intent twice* mechanism to enhance empathy.

Existing models typically generate responses based on either mimicking emotional states (Lin et al., 2019; Majumder et al., 2020) or incorporating external knowledge including multi-resolution strategies (Li et al., 2020), commonsense reasoning through predefined sources (Li et al., 2022a) or extracted via COMET (Hwang et al., 2021; Sabour et al., 2022), and multi-grained signals including causes (Bi et al., 2023; Hamad et al., 2024) to enhance contextual understanding.

Recent advances in large language models (LLMs) (Dubey et al., 2024; Yang et al., 2024) have promoted several empathetic dialogue models utilizing multiple-stage Chain-of-Thought (COT)

(Chen et al., 2024; Hu et al., 2024) with fine-tuning (Zhang et al., 2020; Cai et al., 2024). However, their unstable performance (Lu et al., 2022; Xie et al., 2023) and reliance on external knowledge and high training costs (Kaplan et al., 2020) complicate practical implementation. Consequently, current research focuses on enhancing small-scale empathetic models through empathy mechanisms (Majumder et al., 2020; Gao et al., 2023; Zhou et al., 2023; Wang et al., 2024) as a more lightweight, practical alternative to LLMs. In summary, lightweight empathetic models encounter three major limitations: (1) They primarily rely on supplementary knowledge signals (Gao et al., 2023) rather than underlying psychological mechanisms, which impedes controllability and empathetic capability. (2) They often overlook the internal mechanisms behind emotional causes, emotions, and intents, which rely heavily on external knowledge or pre-trained annotators (Bi et al., 2023; Chen et al., 2024), resulting in hard-coded enhancements rather than genuine understanding and iterative correction, thus impacting empathy, diversity and flexibility. (3) There is a shortage of multi-task datasets for emotion reason masking, intent prediction, and empathetic dialogue. Most models rely on supplementary datasets for auxiliary tasks (Li et al., 2024; Bi et al., 2023), which does not guarantee that the advantages of multi-task training are fully realized or effectively aligned.

To address above limitations, we propose ReflectDiffu, a lightweight and comprehensive framework for empathetic response that seamlessly blends emotional contagion with intent prediction through a reflect mechanism. In sociology, empathetic actions are caused by emotional contagion (Hatfield et al., 1993) and empathetic mimicking (De Waal and Preston, 2017), which indicate an imitation feedback mechanism between human emotions and intentional actions (Rizzolatti and Craighero, 2005; Iacoboni, 2009), as depicted in Figure 1. Our key contributions include:

- We introduce a novel empathetic framework, ReflectDiffu, guided by sociological theories on emotional contagion and empathetic mimicry to improve empathy.
- We propose an *intent twice* mechanism, termed *Exploring-Sampling-Correcting* guided by reflect mechanism to align emotion with intent and minimize empathetic response misalignment caused by emotional misrecognition.
- We utilize LLMs to expand annotation on `EmpatheticDialogue` (Rashkin et al., 2018) and create a multi-task dataset for emotional reason masking, emotion prediction, intent prediction, and empathetic dialogue generation.
- We conducted extensive experiments demonstrating that ReflectDiffu outperforms state-of-the-art models in both automatic and human evaluations.

2 Related Work

2.1 Empathetic Response Generation

Empathetic response generation entails recognizing emotional states and producing suitable emotional responses (Davis, 1983; Rashkin et al., 2018). Early studies primarily aimed at generating emotion-specific responses based solely on emotional states (Lin et al., 2019; Majumder et al., 2020), but faced challenges with the explainability and controllability of empathy. Additionally, reinforcement learning (RL) has been employed to refine dialogue policies, with works like (Li et al., 2024) leveraging policy-based RL to optimize empathetic response generation. Recent studies have integrated external commonsense reasoning (Li et al., 2022a; Zhou et al., 2023), predefined knowledge (Li et al., 2020; Gao et al., 2023; Wang et al., 2024) or pre-trained causal factors (Hwang et al., 2021; Sabour et al., 2022) to enhance emotional perception, but they overlook the established interconnections among factors (De Waal and Preston, 2017), which restricts deeper interpretability and empathy.

Unlike previous approaches, ReflectDiffu introduces reflection interconnection to systematically convert emotional dimensions into actionable intents, thereby improving empathy.

2.2 Generative Model for Dialogue Generation

Generative models have exhibited outstanding performance, facilitating text generation. Majumder et al. (2020) pioneered introducing Variational Autoencoders (VAEs) (Park et al., 2018) to imbue text with empathetic expressions. Further research by Gao et al. (2023) introduced latent variables (Sohn et al., 2015) accounting for cognition, affection, and behavior to better model emotional dependencies in dialogues.

Subsequently, Denoising Diffusion Probabilistic Models (DDPMs) (Ho et al., 2020) stands out for generating high-quality samples via iterative

denoising. Hoogeboom et al. (2021) and Austin et al. (2021) paved the way for character-level text generation with diffusion. Li et al. (2022b) uses an embedding and rounding strategy and additional classifiers for controllable text generation. Gong et al. (2022) introduces a classifier-free diffusion model for dialogue generation. Bi et al. (2023) incorporated multi-grained control signals but their multi-stage pre-training approaches increase computational costs and the difficulty of practical implementation.

As far as we are aware, we are among the first to achieve multi-task empathetic response generation using reinforcement learning within diffusion guided by psychological knowledge.

3 Methodology

Our model, ReflectDiffu, is inspired by sociological studies on emotional contagion (Hatfield et al., 1993) and empathetic mirroring (De Waal and Preston, 2017), which suggest that empathy involves aligning emotional states and mimicking empathetic behavior in interpersonal interactions. Positive emotions are met with positivity, while in situations involving negative emotions, the empathetic response strategy incorporates a congruent emotional tone infused with positivity and a precise empathy intent to resonate deeply with speakers (Majumder et al., 2020; Chen et al., 2022a).

ReflectDiffu comprises two essential components: an *Emotion-Contagion Encoder*, enhanced with an emotional reasoning annotator for improved semantic comprehension, and a *Rational Response Generation Decoder* guided by an *Intent Exploring-Sampling-Correcting* mechanism, which mirrors human reflective dialogue behavior to enhance empathy with robustness. The architecture of our model is depicted in Figure 2.

3.1 Task Definition

The conversation history consists of multiple interactions between a user and a chatbot, represented as $C = [c_0, c_1, \ldots, c_{n-1}]$, where n denotes the number of conversation rounds. Each utterance, c_i, is tokenized into a sequence of words:$C = [w_0^0, w_1^0, \ldots, w_0^1, w_1^1, \ldots, w_{m-1}^{n-1}]$. The primary aim is twofold: accurately discerning users' emotional state, denoted as emo, and formulating an empathetic response, c_n. Additionally, We introduce two auxiliary tasks: emotion reasoning annotation and intent prediction. Within c_i, emotional keywords are marked with the tag <em>, while non-emotional words are labeled <$noem$>. The chatbot also predicts the underlying conversation intent, $intent$, based on the entire dialogue sequence C.

3.2 Multi-task Emotion-Contagion Encoder

Emotion Reason Annotator. The Emotion Reason Annotator (*ERA*) identifies emotional cues and generates reasoning masks within conversational turns. To efficiently handle labeled data for downstream NER tasks, we follow Bogdanov et al. (2024) to utilize a LLM for data annotation and conduct distillation training with other models. *ERA* builds upon BERT (Devlin et al., 2019), an attention-based semantic composition network, and conditional random fields (CRF) to effectively annotate emotional phrases with tags in the sequence r consisting of <em> or <$noem$> and reasoning representations $\tilde{h}$ described in Appendix C.1.

Emotion-Contagion Encoder. The Emotion-Contagion Encoder incorporates the reasoning masks learned by *ERA* into a transformer encoder to emulate emotional contagion.

Given that the reasoning masks r for <em> or <$noem$> are only applied to users' emotional state, the chatbot's r is always set as <$noem$> because empathy is user-oriented, the distinction between users' states and system states has been made. Therefore, unlike previous methods(Sabour et al., 2022; Gao et al., 2023; Bao et al., 2024), we enhanced the context embedding by defining it as the sum of three embeddings (E^C): semantic word embedding (E^W), positional embedding(E^P), and reason embedding (E^R), incorporating the <$em/noem$> into the final embedding, E^C, formally:

$$E^C = E^W(C) + E^P(C) + E^R(C), \quad (1)$$

where $E^W(C), E^P(C), E^R(C) \in \mathbb{R}^{D_{emb}}$.

Then, each token w_i^j in E^C is transformed into its vector representation utilizing the context embedding E_C. Following the existing methods (Gao et al., 2023; Wang et al., 2024; Hamad et al., 2024), we use a transformer encoder with one additional token, CTX, prepended to gain the speaker context. The transformer encoder, denoted as TRS_{Enc}, encodes the flattened E^C into a context representation H:

$$H = \mathsf{TRS}_{\mathsf{Enc}}(E^C(C)). \quad (2)$$

Finally, given the token-level context representation H and the reasoning representations $\tilde{h}$ ob-

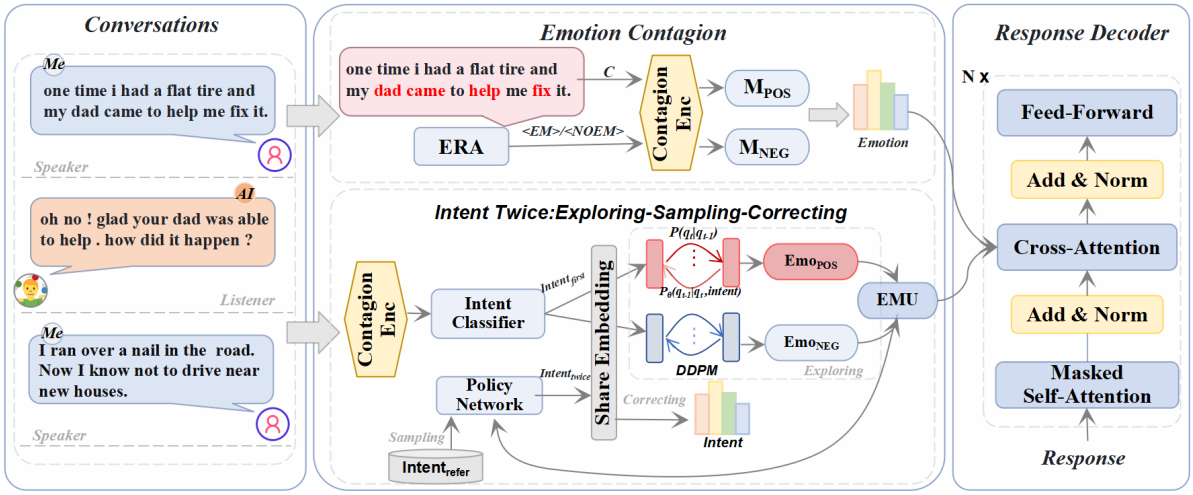

Figure 2: Architecture of our model (ReflectDiffu), which comprises three primary components: Empathetic Imitation influenced by Emotional Contagion, *Intent Twice: Exploring-Sampling-Correcting* Mechanisms and a Response Decoder.

tained from *ERA*, we enhance r by integrating it with $\tilde{h}$ through an attention layer and meaning aggregation, yielding overall representation Q, formally:

$$Q = \texttt{mean-pooling}(\texttt{Attention}(H, \tilde{h})). \quad (3)$$

Contrastive-Experts Emotion Classification. Inspired by (Lin et al., 2019; Majumder et al., 2020; Chen et al., 2022b), we put forward two-expert models *C-Experts*: $\mathbb{M}pos$ for positive emotions and $\mathbb{M}neg$ for negative emotions to enhance emotion recognization by exploiting each model's proficiency. Neutral emotions are addressed via a voting mechanism between experts, yielding the candidate emotion probability distribution p as follows::

$$p = \begin{cases} \texttt{softmax}(W_{neg}\mathbf{E}_{emo}Q) & if\ v = n_{neg} \\ \text{softmax}(W_{pos}\mathbf{E}_{emo}Q) & if\ v = n_{pos} \\ Voting(W_{\text{neg}}\mathbf{E}_{\text{emo}}Q), W_{\text{pos}}\mathbf{E}_{\text{emo}}Q))) & \text{if } v = \text{n}_{\text{neu}} \end{cases} \quad (4)$$

where v represents the maximum count among $\text{n}_{\text{neu}}, \text{n}_{\text{pos}}, \text{n}_{\text{neg}}$ within a batch, based on preliminary real-time sentiment analysis via VADER(Hutto and Gilbert, 2014). W_{pos} and W_{neg} are trainable weight matrices. `Voting(*)` is a soft voting mechanism based on experts $\mathbb{M}pos$ and $\mathbb{M}neg$.

Additionally, we customize the n_{emo} NT-Xent loss (n_{emo} = 32), denoted as L_{NTX}, using pseudo labels to enhance context representation learning(Chen et al., 2020; Zheng et al., 2023) on Q, while utilizing cross-entropy loss, L_{ce}, for classification. Consequently, the overall loss for emotion classification, L_{em}, is detailed in Appendix C.2.

3.3 Multi-task Rational Response Generation Decoder

Building on psychological works(Hatfield et al., 1993; De Waal and Preston, 2017), we propose that empathy involves mirroring users' emotions, responding positively to positive emotions and combining support with optimism for negative states. To enhance emotional encoding and empathetic expression with controllability, we conceptualize response intentions as actions(Chen et al., 2022a; Gao et al., 2023). Unlike existing methods (Bi et al., 2023; Chen et al., 2022a)that rely on signals from externally fine-tuned classifiers, our multi-task response decoder integrates reinforcement learning into a diffusion framework(Ho et al., 2020; Gong et al., 2022) to refine intent and enhance empathy, integrating Intent Twice, Emotion-Intent Mimicry, and Response Decoding to ensure coherent and empathetic interactions.

3.3.1 Intent Twice: Exploring-Sampling-Correction.

Exploring: First Intent Initialization. To enrich the contextual representation Q with extra precise intent information, we consider both internal and external factors to score each candidate intent. In particular, we fine-tune a `BERT` classifier on the `EMPATHETICINTENT` dataset (Chen et al., 2022a) offline to obtain the intent distribution p_{intent}. Following a similar procedure as in Section 3.2, we compute $p_{semantic}$ online using similarity metrics and combine the two distributions to re-rank the in-

tents so that we can get a more accurate first intent prediction $Intent_{first}$:

$$Intent_{first} = p_{semantic} + \alpha p_{intent}. \quad (5)$$

Here, α is a hyperparameter that balances internal and external factors.

Sampling: RL-Diffusion for Intent Twice. Inspired by (Majumder et al., 2020; De Waal and Preston, 2017), we hypothesize that empathetic behavior requires mimicking user emotions and integrating references to common emotion-corresponding actions with one's own cognitive process when learning empathic expression(Majumder et al., 2020). Hence, the alignment between current emotions and inferred intents with universal intents, denoted as $Intent_{refer}$, is crucial for refining intention predictions, especially when errors arise in expert emotion recognition. To enhance the accuracy of action predictions and improve both controllability and effectiveness of empathetic responses, we integrate policy-based reinforcement learning (RL) within Denoising Diffusion Probabilistic Models (DDPMs)(Ho et al., 2020; Gong et al., 2022), to sample more accurate and universal intents. Our framework leverages the exploration-exploitation trade-off $\mathbb{M}_p$ to balance the learned intent actions and the sampling of new empathetic actions. When previous emotion recognition errors occur, *intent twice* mechanism can alleviate emotional misrecognition and correct wrong intents by sampling universal intents.

To define $Intent_{refer}$ for reference, we perform a statistical survey for each emotion to find the top-n empathetic intention actions. The optimal value of n is 3 as shown in Table 1. We provide an experimental analysis for $n = 3$ in Appendix B.1.

Emotion	$\text{Intent}_{\text{refer}}$
surprised, proud, impressed, nostalgic, trusting, faithful, prepared	acknowledging, encouraging, neutral
excited, confident, joyful, grateful, content, caring, faithful	encouraging, sympathizing, acknowledging
angry, disappointed	consoling, suggesting, encouraging
hopeful, sentimental	encouraging, wishing, consoling
anticipating, lonely, afraid, anxious, guilty, embarrassed, sad, apprehensive, terrified, jealous	consoling, encouraging, neutral
hopeful, sentimental	encouraging, wishing, consoling

Table 1: Mapping of Emotion-Group to Top-3 Universal Intents for Reference

State Representation: Emotion Mimicry Unit. Emotion Mimicry Unit(*EMU*) initially splits the emotion-contagion encoding Q into positive and negative polarity representations following emotion grouping(Majumder et al., 2020), but with $Intent_{first}$ guidance in diffusion. We train two distinct DDPMs for positive-polarity representation Emo_{pos} with $L_{kl_{pos}}$ and negative-polarity representation Emo_{neg} with $L_{kl_{neg}}$ following emotion-group(Majumder et al., 2020), we integrate the captured nuances of each emotional polarity with the content encoding H to obtain state Emo_{fused}.

Given the emotion-contagion encoding Q and a fixed step t, the diffusion process iteratively adds Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ to Q over t steps:

$$Q_t = \sqrt{1-\beta_t} Q_{t-1} + \sqrt{\beta_t}\epsilon. \quad (6)$$

Here, Q_t denotes emotion-contagion encoding at time step t, and $\beta_t \in [1\text{e}-5, 5\text{e}-2]$ is the noise level at time step t.To recover the corrupted encodings q_t to their original context representation, we propose an intent-aware Conditional Variational Auto Encoder(CVAE) $\mathcal{M}_\theta$ that predicts the noise ϵ at each step, motivated by Park et al. (2018); Chung et al. (2022):

$$\tilde{Q}_{t-1} = \frac{1}{\sqrt{1-\beta_t}} \left(Q_t - \frac{\beta_t \mathcal{M}_\theta(Q_t, t, Intent_{first})}{\sqrt{1-\sum_{s=1}^{t}\beta_s}} \right). \quad (7)$$

Here, $\tilde{Q}_{t-1}$ represents the reconstructed encoding, and θ denotes the parameters of $\mathcal{M}_\theta$. Finally, we integrate with context encoding H via cross-attention to get state Emo_{fused}, expressed as:

$$Emo_{fused} = \texttt{CrossAttention}([Emo_{pos}, H], [Emo_{neg}, H]). \quad (8)$$

Action Definition: $\text{Intent}_{\text{Twice}}$. The action $Intent_{Twice}$ involves selecting an empathetic intent from a predetermined set of $intent_{refer}$ (Table 1), as determined by the policy network $\mathbb{M}_p$. This network comprises two linear layers and returns a probability distribution p_{act} over $intent_{refer}$. Consequently, an action is sampled in accordance with this distribution. The importance sampling ratio $\frac{\pi(Intent_{Twice}|Emo_{fused})}{\mu(Intent_{Twice}|Emo_{fused})}$ is employed to rectify any discrepancies within the policy.

Reward Calculation: The reward r is calculated based on how well the selected $\text{Intent}_{\text{Twice}}$ aligns with the user's emotional state e, involving two key components: a reward for positive

alignment and a penalty for negative alignment, formally:

$$R(e) = \begin{cases} \texttt{sigmoid}(\mathit{Emo}_{pos}[i] \cdot \mathit{intent}_{refer}) & \mathit{if\ is\_pos}(e) \\ \texttt{sigmoid}(\mathit{Emo}_{neg}[i] \cdot \mathit{intent}_{refer}) & \mathit{otherwise} \end{cases}. \tag{9}$$

Here, intent_{refer} is the selected intent's embedding.

Correction: Intent Adjustment Finally, the intent embeddings are updated through a shared-weight layer during *intent twice* to obtain the final intent *intent* with optimizing cross-entropy loss, denoted as L_{intent}, ensuring consistency and effectiveness in learning and mimicking empathetic intents.

Overal, the loss for intent_{twice} mechanism is represented as L_{twice} :

$$L_{twice} = L_{kl_{pos}} + L_{kl_{neg}} + L_{intent}. \tag{10}$$

3.3.2 Response Decoding:

Consequently, We generate the final response using the integrated response-emotion context, $\text{Emo}_{\text{fused}}$. Following Lin et al. (2019); Majumder et al. (2020); Sabour et al. (2022); Zhou et al. (2023), We apply a transformer decoder, TRS_{dec} with pointer generator network $P_{\text{Gen}}(*)$, where Emo_{fused} serves as both the key and value to predict the word distribution P_v, as detailed below:

$$\begin{aligned} P_w &= P(R_t \mid E_{R<t}, \mathit{Emo}_{fused}, C) \\ &= P_{\text{Gen}}(\texttt{TRS}_{\texttt{dec}}(E^C(T_{R<t}), \mathit{Emo}_{fused}), \end{aligned} \tag{11}$$

where $E_{R<t}$ denotes the embeddings of all prior responses up to time t-1, $E^C(T_{R<t})$ indicates the embedding of the target response, and $P_{\text{Gen}}(*)$ represents the pointer generator network (See et al., 2017). TRS_{dec} refers to the transformer decoder function.

3.4 Training

Lastly, all parameters of ReflectDiffu are trained jointly in an end-to-end manner to optimize the model by integrating all losses L, employing loss weight averaging with hyperparameters δ, ζ, η as follows:

$$L = \delta \text{L}_{\text{em}} + \zeta L_{twice} + \eta L_{res}. \tag{12}$$

4 Experiments Settings

4.1 Dataset

We evaluate our approach, ReflectDiffu, using the EMPATHETICDIALOGUES dataset (Rashkin et al., 2018), which consists of 24,850 open-domain, multi-turn conversations between two interlocutors where the chatbot provides empathetic responses to the user. 32 emotion categories evenly distributed across all dialogues. Utilizing the ChatGLM4 [1] (GLM et al., 2024; Kojima et al., 2022; Zhong et al., 2024) to annotate emotional reasoning within the EMPATHETICDIALOGUES dataset. Additionally, we utilize a fine-tuned Hugging Face model, Commonsense-QA-LLM[2] , to reason and annotate the intents. Ultimately, we extend the original dataset with annotations for emotion reasoning, intent prediction and empathetic dialogue, adhering to the predefined 8:1:1 train/validation/test split.

4.2 Baselines

In our experiments, we compare ReflectDiffu with both classic and recent state-of-the-art (SOTA) benchmarks including MTRS (Rashkin et al., 2018) , MOEL (Lin et al., 2019) , MIME (Majumder et al., 2020) , EmpDG (Li et al., 2020) , KEMP (Li et al., 2022a) , CASE (Zhou et al., 2023) , and CAB (Gao et al., 2023). Additionally, we incorporate a comparative analysis with QWen2-7B (Yang et al., 2024) and Llama-3.1-8B(Dubey et al., 2024), two prominent generative language models(Laskar et al., 2024). More details about baselines are shown in Appendix A.1.

4.3 Implement Details

ReflectDiffu employs 300-dimensional pre-trained GloVe vectors (Pennington et al., 2014) and follows baselines(Rashkin et al., 2018; Lin et al., 2019; Majumder et al., 2020; Li et al., 2020; Gao et al., 2023; Zhou et al., 2023) for a fair comparison. It is implemented in `PyTorch 2.1.2` and trained on two `NVIDIA GeForce RTX 4090 GPUs` with a batch size of 32 using NoamOpt as the optimizer with learning rate warmup steps of 6000 and a learning rate decay factor of 0.01. The diffusion step is set to 1000 and the model converges after about 16000 iterations with early stopping.

4.4 Evaluation Metrics

Automatic Evaluations. To assess ReflectDiffu's performance, we use automatic evaluation metrics for relevance, controllability, and informativeness, including *BLEU-n*, BART_{Score}, Emo-

[1] https://huggingface.co/THUDM/glm-4-9b-chat-1m

[2] https://huggingface.co/rvv-karma/Commonsense-QA-Mistral-7B

Models	Relevance					Controllability		Informativeness		
	BLEU-1 ↑	BLEU-2 ↑	BLEU-3 ↑	BLEU-4 ↑	$BART_{Score}$ ↑	Acc_{emo} ↑	Acc_{intent} ↑	PPL ↓	Dist-1 ↑	Dist-2 ↑
MTRS (Rashkin et al., 2018)	17.87	8.51	4.36	2.61	0.5173	32.96	-	37.98	0.40	1.57
MOEL (Lin et al., 2019)	18.02	8.67	4.35	2.73	0.5166	31.02	-	36.81	0.43	1.76
MIME (Majumder et al., 2020)	19.82	8.86	4.43	2.77	0.5182	30.26	-	36.93	0.51	1.92
EmpDG (Li et al., 2020)	19.12	8.91	4.89	2.85	0.5171	32.90	-	37.55	0.49	1.65
KEMP (Li et al., 2022a)	17.92	8.54	4.38	2.71	0.5232	36.40	-	36.59	0.66	2.43
CASE (Zhou et al., 2023)	19.66	8.95	4.92	2.90	0.5336	38.96	-	35.97	0.70	2.66
CAB (Gao et al., 2023)	20.23	9.39	4.96	3.01	0.5392	40.52	-	35.06	0.89	2.95
Qwen2-7B + COT(Yang et al., 2024)	23.31	11.21	5.20	3.45	0.5447	23.10	41.61	25.45	0.87	3.87
llama-3.1-8B + COT(Dubey et al., 2024)	23.38	**11.29**	5.25	3.47	0.5480	21.15	32.02	24.92	0.92	4.13
ReflectDiffu	**23.59**	11.25	**5.35**	**3.62**	**0.5630**	**48.76**	**80.32**	**24.56**	**0.98**	**4.35**
ReflectDiffu w/o *ERA*	22.59	10.66	5.02	3.28	0.5520	42.37	78.68	24.78	0.95	4.27
ReflectDiffu w/o *C-Experts*	23.13	11.05	5.06	3.31	0.5619	39.44	77.44	24.82	0.91	4.03
ReflectDiffu w/o *Intent twice*	20.91	9.86	4.87	3.16	0.5436	44.56	66.44	29.25	0.85	3.97
ReflectDiffu w/o *EMU*	21.95	10.05	4.96	3.22	0.5490	48.35	79.24	27.45	0.69	2.96

Table 2: Results of automatic evaluations and ablation study.

tion Accuracy Acc_{emo}, Intent Accuracy Acc_{intent}, *Distinct-1*, *Distinct-2*, and Perplexity *PPL* (see Appendix A.2 for details).

Human Evaluation. For human evaluation, we conduct A/B testing on empathy, relevance and fluency with three recruited annotators and a supervisory LLM to resolve disagreements (see Appendix A.2 for details).

5 Results and Discussion

Automatic Evaluation Results As shown in table 2, our model, ReflectDiffu, outperforms all baseline models and significantly enhances all metrics. Compared with empathy-specific models(Rashkin et al., 2018; Lin et al., 2019; Majumder et al., 2020; Li et al., 2020, 2022a; Zhou et al., 2023; Gao et al., 2023), which mainly explore the connection between emotion states and empathetic contexts but ignore the internal mechanisms of emotional causes, emotions, and intents and only rely on inferred external knowledge, resulting in suboptimal empathetic controllability (low emotion accuracy Acc_{emo}), weak similarity and coherence with the empathetic ground truth (indicated by low *BLEU-n* and $BART_{Score}$), and a lack of diversity (implied by low *Distinct-1* and *Distinct-2*). In contrast, ReflectDiffu exhibits remarkable superiority, exceeding the best baseline, CAB, approximately in *BLEU-1*, *BLEU-2*, *BLEU-3*, *BLEU-4*, $BART_{Score}$, Acc_{emo} by **16.6%**, **20%**, **8.1%**, **20.3%**, **4.6%** and **20.3%** respectively for Emotion-Contagion Encoder to enhance semantic understanding and achieve **80.32%** intent accuracy for its *intent twice* mechanism. Moreover, ReflectDiffu shows improvements of approximately 30.1% in PPL, **10.1%** in *Distinct-2*, and **47.4%** in *Distinct-2* compared with CAB for Diffusion within intent guidance.

Moreover, compared with llama-3.1-8B(Dubey et al., 2024) with Chain-of-Thought(COT) via fewshots (a SOTA LLM-based empathetic dialogue model in our experiments), ReflectDiffu outperforms llama-3.1-8B obviously by 1.90%,4.32%,2.73%,**130.54%**,**150.94%**,6.52% and 5.32% in *BLEU-3*, *BLEU-4*, $BART_{Score}$, Acc_{emo}, Acc_{intent}, *Distinct-1* and *Distinct-2*.Higher $BART_{Score}$, Acc_{emo} and Acc_{intent} robustly underscore the efficacy of ReflectDiffu in fostering empathy. Lower PPL and higher *Distinct-1* and *Distinct-2* further corroborate the empathetic diversity that ReflectDiffu can engender.

Human Evaluation Results Table 3 presents the results of the human A/B testing, comparing ReflectDiffu with various baseline models across three criteria: empathy (Emp.), relevance (Rel.), and fluency (Flu.). The evaluations reveal that ReflectDiffu consistently outperforms the baseline models across all criteria.

Ablation Study. As shown in Table 2, we conducted four ablation studies to evaluate the key components of our model: (1) **w/o *ERA***: Removing the Emotion Reason Annotator (*ERA*) that improves emotion understanding with reasoning masks; (2) **w/o *C-Experts***: Excluding the Contrastive-Experts for emotion classification; (3) **w/o *Intent twice***: Eliminating the Intent *Exploring-Sampling-Correcting* mechanism; and (4) **w/o *EMU***: Lacking the Emotion Mimicry Unit (*EMU*) with DDPMs for state representation.

Effect of *ERA*. Excluding Emotion Reason Annotator(*ERA*) designed to improve emotion un-

Comparison	Aspects	Win	Lose	Tie
ReflectDiffu vs. MultiTRS	Emp.	**51.1**	18.0	30.9
	Rel.	**48.1**	17.5	34.4
	Flu.	**40.1**	11.7	48.2
ReflectDiffu vs. MOEL	Emp.	**45.4**	21.2	33.4
	Rel.	**37.3**	22.5	40.2
	Flu.	**31.4**	13.7	54.9
ReflectDiffu vs. MIME	Emp.	**50.3**	20.8	28.9
	Rel.	**43.7**	19.2	37.1
	Flu.	**38.4**	9.1	52.5
ReflectDiffu vs. EmpDG	Emp.	**52.2**	19.8	27.9
	Rel.	**50.8**	16.5	32.7
	Flu.	**36.4**	10.3	53.3
ReflectDiffu vs. KEMP	Emp.	**55.2**	23.1	21.7
	Rel.	**62.4**	29.8	7.8
	Flu.	**35.7**	13.3	51.0
ReflectDiffu vs. CAB	Emp.	**53.6**	22.4	24.0
	Rel.	**56.1**	24.6	19.3
	Flu.	**32.3**	10.2	57.5
ReflectDiffu vs. CASE	Emp.	**52.0**	15.0	33.0
	Rel.	**45.5**	25.0	29.5
	Flu.	**49.0**	27.1	23.9
ReflectDiffu vs. Qwen2-7B + COT	Emp.	**52.5**	22.2	25.3
	Rel.	**53.1**	25.3	21.6
	Flu.	**41.2**	12.5	46.3
ReflectDiffu vs. llama-3.1-8B + COT	Emp.	**51.2**	21.8	27.0
	Rel.	**54.4**	24.5	21.1
	Flu.	**33.8**	18.5	47.7

Table 3: Human A/B evaluation results between ReflectDiffu and baselines.

derstanding by reasoning masks leads to a significant decrease in *BLEU-n*, $BART_{Score}$ and Acc_{emo}, indicating **w/o *ERA*** compromises emotion perception and thereby results in inferior empathetic responses' relevance and quality.

Effect of *C-Experts*. Removing Contrastive-Experts *C-Experts* leads to a notable decline in Acc_{emo} from 48.76 to 39.44, indicating that **w/o *C-Experts*** deteriorates the ability to classify emotions, consequently negatively affecting the controllability of empathy, making it harder to precisely match responses with desired emotional states.

Effect of *Intent twice*. Eliminating the *Intent Exploring-Sampling-Correcting* mechanism significantly reduced Acc_{intent} from 80.32 to 66.44, along with poor *BLEU-n* and $BART_{Score}$, higher *PPL*, **w/o *Intent twice*** impairs the model's ability to accurately capture and fulfill response intent, weakening empathetic responses' relevance and quality.

Effect of *EMU*. Lacking the Emotion Mimicry Unit (*EMU*) for state representation results in a considerable decrease in *BLEU-n*, *Distinct-1* and *Distinct-2*, along with *PPL*, indicating that **w/o *EMU*** negatively affects the quality and distinctivenes of empathetic responses.

Emotion	Terrified
Context	Yeah about 10 years ago I had a horrifying experience. It was 100% their fault but they hit the water barrels and survived. They had no injuries but they almost ran me off the road.
MTRS	that is pretty scary ! i am glad you are ok .
MOEL	that is so terrible! i am so sorry.
MIME	oh no ! i am so sorry to hear that .
EmpDG	oh no , i am so sorry to hear that .
KEMP	oh no ! i hope you are okay .
CASE	i hope you can get it fixed. Are you okay now?
CAB	I hope you are able to get it fixed,and hope you are ok!
Qwen2-7B+COT	I'm sorry to hear about your experience. It sounds stressful and dangerous.
llama3.1-8B+COT	I'm sorry to hear that. If you want, you can talk more about it.
ReflectDiffu	$Intent_{first}$: *encouraging*× $Intent_{twice}$:*consoling*✓ oh no! That sounds absolutely terrifying . I hope you were not hurt, Were you injured ?
Golden	Did you suffer any injuries?

Table 4: Case study comparison between ReflectDiffu and baselines.

Case Study. In this case (Table 4), ReflectDiffu shows improved empathetic response generation by identifying and mimicking the user's emotional state. Using the *Intent Exploring-Sampling-Correcting* mechanism, the model refines its initial intent from *encouraging* to *consoling*, resulting in a more supportive reply. Compared to baselines, ReflectDiffu better aligns with users' emotions, offers a clear and empathetic follow-up, enhancing interaction quality. (Details on mitigating emotion recognition errors are provided in Appendix B.1.)

6 Conclusion

In this paper, we propose ReflectDiffu, a novel psychological multi-task framework for empathetic dialogue that integrates Emotion-Contagion Encoder and Response Generation Decoder guided by an *Intent Twice* mechanism to better understand users' emotional states, predict intents accurately, and generate highly intent-aligned empathetic responses. Both automated and human evaluations demonstrate that ReflectDiffu excels in relevance, controllability, and informativeness of empathetic dialogue. Our research may inspire future studies on modeling emotion-intent interaction in human discourse and other linguistic behaviors.

Limitations

Our ReflectDiffu framework, integrating emotion contagion and intent prediction mechanisms with the *Intent Twice* mechanism, has performed exceptionally in both automatic and human evaluations, significantly enhancing the relevance, controllability, and informativeness of empathetic responses.

We discuss the primary limitation of this work as follows: The integration of Denoising Diffusion Probabilistic Models (DDPMs) and reinforcement learning mechanisms has augmented the computational requirements for training, presenting challenges for deployment in resource-constrained settings or on devices with limited capabilities. To alleviate this limitation, we have adopted reparameterization and multi-task techniques for optimization. As a result, the overall training time is notably shorter than that of multi-stage LLM (Chen et al., 2024; Yang et al., 2024; Dubey et al., 2024) while achieving state-of-the-art outcomes.

In conclusion, despite the existing limitations, ReflectDiffu is relatively lightweight compared to LLM. Moreover, our ongoing research efforts aim to achieve lightweight quantization to accelerate the model's implementation and collaboration.

Ethical Considerations

Our research utilizes the EMPATHETICDIALOGUES dataset Rashkin et al. (2018), an open-source resource devoid of any personal privacy information. To annotate the data for emotion reasoning and intent prediction, we leverage prompts teqhniques (Kojima et al., 2022) and LLM contrastive voting mechanisms (Zhong et al., 2024) to label intent and emotional reason, thereby minimizing human bias and reducing the risk of model hallucination. Our human evaluations are conducted by three professional annotators, who operate anonymously to protect privacy and ensure objective assessments following our instructions (refer to Appendix D). Annotators are compensated fairly for their contributions.

References

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993.

Yinan Bao, Dou Hu, Lingwei Wei, Shuchong Wei, Wei Zhou, and Songlin Hu. 2024. Multi-stream information fusion framework for emotional support conversation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11981–11992.

Guanqun Bi, Lei Shen, Yanan Cao, Meng Chen, Yuqiang Xie, Zheng Lin, and Xiaodong He. 2023. Diffusemp: A diffusion model-based framework with multi-grained control for empathetic response generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2812–2831.

Sergei Bogdanov, Alexandre Constantin, Timothée Bernard, Benoit Crabbé, and Etienne Bernard. 2024. Nuner: Entity recognition encoder pre-training via llm-annotated data. *arXiv preprint arXiv:2402.15343*.

Mingxiu Cai, Daling Wang, Shi Feng, and Yifei Zhang. 2024. Empcrl: Controllable empathetic response generation via in-context commonsense reasoning and reinforcement learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5734–5746.

Laurie Carr, Marco Iacoboni, Marie-Charlotte Dubeau, John C Mazziotta, and Gian Luigi Lenzi. 2003. Neural mechanisms of empathy in humans: a relay from neural systems for imitation to limbic areas. *Proceedings of the national Academy of Sciences*, 100(9):5497–5502.

Mao Yan Chen, Siheng Li, and Yujiu Yang. 2022a. Emphi: Generating empathetic responses with human-like intents. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1063–1074.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Wei Chen, Jinglong Du, Zhao Zhang, Fuzhen Zhuang, and Zhongshi He. 2022b. A hierarchical interactive network for joint span-based aspect-sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7013–7019.

Xinhao Chen, Chong Yang, Man Lan, Li Cai, Yang Chen, Tu Hu, Xinlin Zhuang, and Aimin Zhou. 2024. Cause-aware empathetic response generation via chain-of-thought fine-tuning. *arXiv preprint arXiv:2408.11599*.

Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. 2022. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*.

Benjamin MP Cuff, Sarah J Brown, Laura Taylor, and Douglas J Howat. 2016. Empathy: A review of the concept. *Emotion review*, 8(2):144–153.

Carol M Davis. 1990. What is empathy, and can empathy be taught? *Physical therapy*, 70(11):707–711.

Mark H Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology*, 44(1):113.

Frans BM De Waal and Stephanie D Preston. 2017. Mammalian empathy: behavioural manifestations and neural basis. *Nature Reviews Neuroscience*, 18(8):498–509.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jun Gao, Yuhan Liu, Haolin Deng, Wei Wang, Yu Cao, Jiachen Du, and Ruifeng Xu. 2021. Improving empathetic response generation by recognizing emotion cause in conversations. In *Findings of the association for computational linguistics: EMNLP 2021*, pages 807–819.

Pan Gao, Donghong Han, Rui Zhou, Xuejiao Zhang, and Zikun Wang. 2023. Cab: Empathetic dialogue generation with cognition, affection and behavior. In *International Conference on Database Systems for Advanced Applications*, pages 597–606.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. 2022. Diffuseq: Sequence to sequence text generation with diffusion models. In *arXiv preprint arXiv:2210.08933*.

Omama Hamad, Ali Hamdi, and Khaled Shaban. 2024. Asem: Enhancing empathy in chatbot through attention-based sentiment and emotion modeling. *arXiv preprint arXiv:2402.16194*.

Elaine Hatfield, John T Cacioppo, and Richard L Rapson. 1993. Emotional contagion. *Current directions in psychological science*, 2(3):96–100.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. 2021. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*, 34:12454–12465.

Yuxuan Hu, Minghuan Tan, Chenwei Zhang, Zixuan Li, Xiaodan Liang, Min Yang, Chengming Li, and Xiping Hu. 2024. Aptness: Incorporating appraisal theory and emotion support strategies for empathetic response generation. *arXiv preprint arXiv:2407.21048*.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6384–6392.

Marco Iacoboni. 2009. Imitation, empathy, and mirror neurons. *Annual review of psychology*, 60:653–670.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213.

Md Tahmid Rahman Laskar, Sawsan Alqahtani, M Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Amran Bhuiyan, Chee Wei Tan, Md Rizwan Parvez, et al. 2024. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. *arXiv preprint arXiv:2407.04069*.

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

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 Inter-*

national Conference on Computational Linguistics, pages 4454–4466.

Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022a. Knowledge bridging for empathetic dialogue generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 10993–11001.

Wendi Li, Wei Wei, Kaihe Xu, Wenfeng Xie, Dangyang Chen, and Yu Cheng. 2024. Reinforcement learning with token-level feedback for controllable text generation. *arXiv preprint arXiv:2403.11558*.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori Hashimoto. 2022b. Diffusion-lm improves controllable text generation. In *arXiv preprint arXiv:2205.14217*.

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.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Yookoon Park, Jaemin Cho, and Gunhee Kim. 2018. A hierarchical latent structure for variational conversation modeling. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1792–1801.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Pengnian Qi and Biao Qin. 2023. Ssmi: Semantic similarity and mutual information maximization based enhancement for chinese ner. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13474–13482.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: a new benchmark and dataset. In *Proceedings of the Association for Computational Linguistics*, pages 467–478.

Giacomo Rizzolatti and Laila Craighero. 2005. Mirror neuron: a neurological approach to empathy. In *Neurobiology of human values*, pages 107–123. Springer.

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.

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

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2015. Hierarchical neural network generative models for movie dialogues. *arXiv preprint arXiv:1507.04808*, 7(8):434–441.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2019. Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv:1909.10649*.

Yufeng Wang, Chao Chen, Zhou Yang, Shuhui Wang, and Xiangwen Liao. 2024. Ctsm: Combining trait and state emotions for empathetic response model. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4214–4225.

Qiming Xie, Zengzhi Wang, Yi Feng, and Rui Xia. 2023. Ask again, then fail: Large language models' vacillations in judgement. *arXiv preprint arXiv:2310.02174*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Xiaolin Zheng, Mengling Hu, Weiming Liu, Chaochao Chen, and Xinting Liao. 2023. Robust representation learning with reliable pseudo-labels generation via self-adaptive optimal transport for short text clustering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10493–10507.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2024. Rose doesn't do that: Boosting the safety of instruction-tuned large language models with reverse prompt contrastive decoding. *arXiv preprint arXiv:2402.11889*.

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.

A Experimental Details

A.1 Baselines

In our experiments, we compare ReflectDiffu with both classic and recent state-of-the-art (SOTA) benchmarks.

- **Multitask-Transformer(MTRS):**Rashkin et al. (2018) introduced a Transformer model trained for both sentiment detection and empathetic response generation.
- **MOEL:** Lin et al. (2019) proposed a Transformer model with 32 emotion-specific decoders and a meta-listener to generate contextually appropriate responses.
- **MIME:** Majumder et al. (2020) combined a Transformer with a VAE to generate empathetic responses by mimicking user emotions through polarity-based clustering and stochastic emotion mixtures.
- **EmpDG:** Li et al. (2020) used a Transformer with a WGAN to capture emotional nuances via a token-level perception mechanism.
- **KEMP:** Li et al. (2022a) proposed leveraging external knowledge, including commonsense and emotional lexical knowledge, to enhance empathetic dialogue generation.
- **CASE:** Zhou et al. (2023) integrated a commonsense cognition graph and an emotional concept graph to align user cognition and affection for empathetic responses.
- **CAB:** Gao et al. (2023) integrated cognition, affection, and behavior to enhance empathetic dialogue generation.
- **QWen2-7B + COT:** We fine-tune QWen2-7BYang et al. (2024), and then employ Chain-of-Thought (COT) to infer emotion, intent, and generate empathetic responses for improved empathy.
- **llama3.1-8B + COT:** We fine-tune llama3.1-8BDubey et al. (2024), and then employ Chain-of-Thought (COT) to infer emotion, intent, and generate empathetic responses for improved empathy.

A.2 Evalutions Metrics

Automatic Evaluation. To assess ReflectDiffu's performance, we use automatic evaluation metrics in three categories: relevance, controllability, and informativeness.

- **Relevance**: We use *BLEU* (Papineni et al., 2002) and $BART_{Score}$(Yuan et al., 2021) to measure similarity between generated and reference texts. Higher scores indicate more relevant outputs.
- **Controllability**: This is measured by Emotion Accuracy (Acc_{emo}) and Intent Accuracy (Acc_{intent}), which check the model's ability to detect emotions and recognize user intent.
- **Informativeness**: Evaluated using Distinct-1, Distinct-2 (Li et al., 2016), and Perplexity (PPL) (Serban et al., 2015).
  - **Distinct-N**: Measures the proportion of unique unigrams and bigrams, indicating diversity. Higher scores show more varied responses.
  - **Perplexity (PPL)**: Lower PPL scores indicate better performance, as the model predicts the next word more accurately, resulting in more fluent and coherent text.

Human Evaluation. Following Zhou et al. (2023); Gao et al. (2023); Wang et al. (2024), we conduct Human A/B testing between response pairs based on the following criteria evaluated by three annotators: **(1) Empathy (Emp.):** Assessing the model's ability to generate empathetic content, including understanding the user's emotional state and responding appropriately. **(2) Relevance (Rel.):** Determining how well the model's responses relate to the dialog history, ensuring coherence and logical progression. **(3) Fluency (Flu.):** Evaluating the naturalness and readability of the replies, including grammatical correctness and ease of understanding. To ensure fair scoring, we introduced a supervisory LLM, ChatGPT[3] inspired by Zheng et al. (2024). In cases of significant disagreement among annotators, ChatGPT provided the final rating.

B Additional Experiments

B.1 Explanation of the hyperparameter n of $\text{intent}_{\text{infer}}$

After annotating the dataset with empathetic intentions, we conducted a statistical analysis to determine the frequency of each intention for every emotion. Figure 3 illustrates this data, where rows represent distinct emotions and columns represent specific empathetic intentions. The color intensity in each cell indicates the relative frequency of a particular intention corresponding to an emotion, with darker shades signifying higher frequencies. Figure 3 aids in understanding the predominant empathetic actions associated with each emotional state, thereby providing insights into the alignment of universal intents ($\text{Intent}_{\text{refer}}$) with user emotions. We observed that setting $n = 3$ effectively avoids non-universal intentions while ensuring that, besides the neutral intent, a more meaningful intent is sampled within the top-2 $\text{Intent}_{\text{refer}}$.

B.2 Case Study in Misclassification

We deliberately selected a ReflectDiffu emotion recognition error case to validate the effectiveness of our reflection mechanism. Table 5 compares responses from various models, including MOEL, MIME, EmpDG, KEMP, CASE, CAB, Qwen2-7B+COT,llama3.1-8B+COT and ReflectDiffu, to a user's context of feeling hopeful after applying for graduate school. Initially, ReflectDiffu misidentified the emotion as "joyful" and the intent as "acknowledging." However, after employing the reflection mechanism, it correctly identified the intent as "encouraging." This demonstrates the model's capability to correct errors and generate more empathetic responses through its reflection mechanism.

[3] https://chat.openai.com

Figure 3: Heatmap of Relative Frequencies of Empathetic Intentions for Each Emotion.

Emotion	hopeful
Context	i just applied for graduate school ! i feel good about my chances !
MTRS	i hope you have a great time !
MOEL	i am sure you will do great .
MIME	i am sure you will do great !
EmpDG	that is great ! i hope you are going to school for a new one ?
KEMP	that is great ! i hope you get it .
CASE	That is good, I am glad you are able to get it
CAB	That is awesome! glad you are better !
Qwen2-7B+COT	That's wonderful! Applying for graduate school is a significant step
llama3.1-8B+COT	Congratulations on taking this important step! That's fantastic!
ReflectDiffu	***emotion:*** joyful × $\textbf{\textit{Intent}}_{\textbf{\textit{first}}}$: acknowledging× $\textbf{\textit{Intent}}_{\textbf{\textit{twice}}}$:consoling✓ i am proud and sure you'll do just fine in school.
Golden	I'm so proud of you! I'll pray for your success!

Table 5: Case study in misclassification comparison between ReflectDiffu and baseline models.

C Implement Details

C.1 Emotion Reason Annotator

Our approach leverages BERT (Devlin et al., 2019), an attention-based semantic composition network, and conditional random fields (CRF) to effectively annotate emotional phrases with tags such as <*em*> or <*noem*>. Specifically, given the token-level dialogue history C, where w_i^j represents the i-th token in the j-th utterance, we use a pretrained BERT model to obtain contextualized token representations h_j^i:

$$h_j^i = \text{BERT}(w_j^i). \tag{13}$$

where h_j^i is the hidden state output by BERT corresponding to the token w_i^j.

Unlike traditional Named Entity Recognition (NER) models (Souza et al., 2019; Qi and Qin, 2023), we introduce an attention-based semantic composition network that progressively distinguishes between binary sets of words.

Each token representation h_j^i is initially treated as a word-level feature representation. The attention network computes the correlation between pairs of word vectors h_j^i and h_m^k. The relevance score α_{ik}^{jm} and reasoning representation $\tilde{h}$ are defined as:

$$\alpha_{ik}^{jm} = \frac{\exp\left(\text{Attention}(h_j^i, h_m^k)\right)}{\sum_{k,m}\exp\left(\text{Attention}(h_j^i, h_m^k)\right)}, \tag{14}$$

$$\tilde{h}_j^i = \sum_{k,m} \alpha_{ik}^{jm} h_m^k. \tag{15}$$

where $\tilde{h}_j^i$ is the attention-weighted representation for the token w_i^j, enriched with contextual information from related tokens within the conversational turn.

Finally, the enriched representations are passed through a CRF layer to obtain the final predictions:

$$P(\mathbf{r} \mid \tilde{\mathbf{h}}) = \frac{\exp\left(\sum_{j=1}^{n}\left(A_{r_{j-1},r_j} + \mathbf{W}r_j\tilde{h}_j^i\right)\right)}{\sum_{\mathbf{y}'\in\mathcal{R}(\tilde{\mathbf{h}})}\exp\left(\sum_{j=1}^{n}\left(A_{r'_{j-1},r'_j} + \mathbf{W}r'_j\tilde{h}_j^i\right)\right)}. \tag{16}$$

where $\mathbf{r} = (r_1, r_2, \ldots, r_n)$ represents the sequence of reasoning labels, each $y_i \in \{$<*em*>, <*noem*>$\}$, $\tilde{h}_j^i$ is the reasoning representation, A is the transition matrix, and $\mathbf{W}$ represents the weights for the CRF layer.

C.2 Definition of L_{em}

Inspired by (Chen et al., 2020; Zheng et al., 2023), we use the NT-Xent loss ($n_{\text{emo}} = 32$) L_{NTX} and cross-entropy loss (L_{ce}) for contrastive emotion classification (L_{em}) , formally:

$$L_{\text{NTX}}^{(n_{\text{emo}})^i} = -\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} \mathbb{1}_{[y_i=y_j]} \log \frac{\exp(\text{s}(i,j))}{\sum_{k=1}^{n}\exp(\text{s}(i,k))}, \tag{17}$$

$$L_{\text{NTX}} = \sum_{i=1}^{32} L_{\text{NTX}}^{(n_{\text{emo}})^i}, \tag{18}$$

$$L_{\text{cls}} = -\log \mathcal{P}[e], \tag{19}$$

$$L_{\text{em}} = L_{\text{NTX}} + L_{\text{cls}}. \tag{20}$$

Here, n represents the number of samples, y_i is the pseudo-label of the i-th sample, $\mathbb{1}_{[y_i=y_j]}$ is an indicator function that equals 1 if $y_i = y_j$ and 0 otherwise, $\text{s}(i,j)$ denotes the similarity between samples i and j, and $\mathcal{P}[e]$ is the predicted probability for the true emotion class e.

D Annotators Instructions for Human Evaluation.

Professional annotators received our detailed guidelines to guarantee high-quality and unbiased evaluations.

- **Evaluation Criteria**: Annotators assessed responses based on three key criteria::
  - **Empathy (Emp.)**: Evaluators were instructed to assess how well the response understood and mirrored the user's emotional state. Examples of high empathy included responses that acknowledged the user's feelings and provided appropriate support or encouragement. Low empathy responses were those that failed to recognize or appropriately respond to the user's emotions.
  - **Relevance (Rel.)**: This criterion focused on how well the response related to the previous conversation context. High relevance responses directly addressed the user's statements or questions, maintaining coherence. Low relevance responses were off-topic or did not logically follow the conversation flow.
  - **Fluency (Flu.)**: Evaluators assessed the grammatical correctness and naturalness of the responses. Fluent responses were

well-structured, easy to read, and free of grammatical errors. Non-fluent responses contained grammatical mistakes, awkward phrasing, or were difficult to understand.

- **Conflict Resolution**: Procedures were established to handle disagreements among annotators:
  - When annotators disagreed on the evaluation of a response, a discussion was initiated to reach a consensus.
  - If consensus could not be achieved, a supervisory Large Language Model (LLM) provided the final rating to ensure objective and consistent evaluations across different cases.
- **Anonymity and Privacy**: Annotators were assured that their evaluations would be anonymized to protect their identities. They were informed that their personal information would not be shared or disclosed in any part of the study, ensuring their privacy and confidentiality.
- **Compensation and Acknowledgment**: Annotators were informed about their compensation:
  - They were fairly compensated for their time and effort in evaluating the responses.
  - Their contributions would be acknowledged in the final publication of the study to recognize their important role in the research process.