DIFFUSEMP: A Diffusion Model-Based Framework with Multi-Grained Control for Empathetic Response Generation

Guanqun Bi1,2, Lei Shen3, Yanan Cao1,2,*, Meng Chen3,*, Yuqiang Xie1,2, Zheng Lin1,2, Xiaodong He3

1Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
2School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
3JD AI Research, Beijing, China

Abstract

Empathy is a crucial factor in open-domain conversations, which naturally shows one’s caring and understanding to others. Though several methods have been proposed to generate empathetic responses, existing works often lead to monotonous empathy that refers to generic and safe expressions. In this paper, we propose to use explicit control to guide the empathy expression and design a framework DIFFUSEMP based on conditional diffusion language model to unify the utilization of dialogue context and attribute-oriented control signals. Specifically, communication mechanism, intent, and semantic frame are imported as multi-grained signals that control the empathy realization from coarse to fine levels. We then design a specific masking strategy to reflect the relationship between multi-grained signals and response tokens, and integrate it into the diffusion model to influence the generative process. Experimental results on a benchmark dataset EMPATHETIC:DIALOGUE show that our framework outperforms competitive baselines in terms of controllability, informativeness, and diversity without the loss of context-relatedness.

1 Introduction

Empathetic response generation, as a conditional text generation task, aims to endow agents with the ability to understand interlocutors and accurately express empathy in their communication (Rashkin et al., 2019; Lin et al., 2019; Li et al., 2020; Shen et al., 2021). However, the generated responses tend to be generic and monotonous (Chen et al., 2022), i.e., showing shallow empathy and few connections to the context. As shown in the upper part of Figure 1, “I’m sorry to hear that.” is used as a reaction to different contexts with negative feelings. To alleviate the problem, existing works mainly incorporate emotion or knowledge modules into the encoder-decoder framework and train their models with the maximum likelihood estimation (MLE) (Rashkin et al., 2019; Lin et al., 2019; Majumder et al., 2020; Li et al., 2020; Sahand Sabour, 2021; Li et al., 2022a).

Recently, diffusion models (Ho et al., 2020; Dhariwal and Nichol, 2021) have emerged as a brand-new and promising paradigm for generative models. A few prior works that explored using diffusion models on text data are mainly designed for unconditional text generation (Austin et al., 2021; Hoogeboom et al., 2021; He et al., 2022). For text generation with extra conditions (control signals or contexts), Diffusion-LM (Li et al., 2022b) applies extra-trained classifiers to make the generated text satisfy input signals like sentiment and syntactic structure. DiffuSeq (Gong et al., 2022) is proposed as a classifier-free diffusion model that uses “partial noising” in the forward process to distinguish the input and output text.

In this paper, we add control signals to empathetic response generation and propose a diffusion model-based framework, DIFFUSEMP, to solve the aforementioned monotonous empathy problem. First, since empathy is a multi-dimensional factor (Davis et al., 1980), i.e., several factors affect the realization of empathy, we use explicit control sign-
ers at different levels to guide response generation. At the utterance level, communication mechanism (CM) (Sharma et al., 2020) divides text-based empathy into emotional reaction, interpretation, and exploration to describe the high-level functionality. Then, we use intent (IT) (Welivita and Pu, 2020) to reflect the behaviors of an agent in each sentence, such as questioning (e.g., What happened to you?). Finally, the fine-grained signal semantic frame (SF) (Baker et al., 1998) is imposed on each token, which represents their universal categories of events, concepts, and relationships. An example of how multi-grained control signals work is illustrated in the lower part of Figure 1. To have exact guidance over responses, these signals are extracted from golden responses in the training process, while during inference, an emotion-enhanced matching method is used to obtain response candidates as the source of control signals.

We then design a diffusion model to make the generated responses not only relevant to dialogue contexts but also express specific empathy under the multi-grained control. The dialogue context, multi-grained control, and response are considered as the model input. For the forward diffusion process, we apply the partial noising (Gong et al., 2022) strategy so that both the context and control signals are unchanged, and only the response is noised. To fulfill the reverse diffusion process, we use the transformer architecture (Vaswani et al., 2017) and introduce a masking strategy to indicate the control range of each signal on response tokens. Specifically, each CM/IT controls all tokens in an utterance/sentence, while an SF term corresponds to exactly one token. Tokens out of the control range are masked in the self-attention layer. Finally, we conduct experiments on a benchmark dataset EMPATHETICDIALOGUE to demonstrate the effectiveness of DIFFUSEMP.

The main contribution of this paper is threefold: (1) We introduce explicit multi-grained control signals to solve the monotonous empathy problem, and convert the empathetic response generation into a controllable setting. (2) We propose DIFFUSEMP, a novel diffusion model-based framework, to unify the utilization of dialogue context and control signals, achieve elaborate control with a specific masking strategy, and integrate an emotion-enhanced matching method to produce diverse responses for a given context. (3) Experimental results show that our method outperforms competitive baselines in generating informative and empathetic responses.

2 Related Work

2.1 Empathetic Response Generation

Rashkin et al. (2019) firstly formulate the empathetic response generation task and construct the EMPATHETICDIALOGUE dataset. Existing works that focus on this task can be divided into two lines. The first is to detect and utilize the user’s emotion with diverse structures (Lin et al., 2019; Majumder et al., 2020; Shen et al., 2021). The second is to consider cognition-based factors other than emotions (EM), such as dialogue act (DA) (Welivita and Pu, 2020), communication mechanism (CM) (Sharma et al., 2020), emotion cause (Jiang et al., 2019), psychological skill (Kim et al., 2021), and commonsense (Sabour et al., 2021; Li et al., 2022a).

Zheng et al. (2021) propose a framework CoMAE to model the relationship among CM, DA, and EM at the utterance level. The differences between CoMAE and DIFFUSEMP are: (1) Instead of predicting each factor based on the context representation, DIFFUSEMP explicitly uses control signals that are highly related to a response as task input. (2) We achieve the elaborate control with multi-grained signals, i.e., tokens in response are influenced by different signals, while CoMAE applies the same combined factor to all decoding positions.

2.2 Diffusion Models

Diffusion models are a class of generative models with promising performance and have been used in a variety of real-world applications. Most existing works of diffusion models focus on continuous data, such as vision (Nichol et al., 2021; Radford et al., 2021; Rombach et al., 2021b) and audio (Popov et al., 2021; Yang et al., 2022; Tae et al., 2021). Due to the discrete nature of text data, the utilization of diffusion models for NLP is challenging. Hoogeboom et al. (2021) and Austin et al. (2021) extend diffusion models to discrete state spaces for character-level text generation. Diffusion-LM (Li et al., 2022b) uses embedding and rounding strategy to bridge the continuous and discrete domain, and trains extra classifiers for controllable text generation. DiffuSeq (Gong et al., 2022) leverages partial noising for sequence-to-sequence text generation to keep the text input unchanged in
To better model and express multi-dimensional empathy in psychology are complex (Davis et al., 1980; de Waal, 2008; Decety and Meyer, 2008), and we choose the control signals that intersect with computational linguistics. Note that the design of DIFFUSEMP is not limited to the following control signals, other factors of empathy can also be used.

**Communication Mechanism (CM).** We employ the taxonomy in Sharma et al. (2020): Emotional Reaction (ER), Interpretation (IP), and Exploration (EX). ER expresses emotions such as warmth, compassion, and concern, IP represents an understanding of feelings and experiences inferred from the speaker, and EX stands for exploring the feelings and experiences not stated in previous utterances. Following Sharma et al. (2020), we use three RoBERTa-based (Liu et al., 2019) classifiers to individually identify whether a response implies a certain mechanism.

**Intent (IT).** A previous analysis (Welivita and Pu, 2020) argues that humans demonstrate a wide range of intents when regulating empathy and proposes a dataset EMPATHETICINTENT. Besides, many works (Xie et al., 2022; Zheng et al., 2021) insist that intents and emotions have a strong relationship. Specifically, listeners are much more likely to respond to positive or negative emotions with specific empathetic intents such as acknowledgment, consolation, and encouragement, rather than only expressing similar or opposite emotions. We train a BERT-based (Devlin et al., 2019) classifier on EMPATHETICINTENT to label responses.

**Semantic Frame (SF).** Semantic frames are based on FrameNet (Baker et al., 1998), a linguistic knowledge graph containing information about lexical and predicate-argument semantics. The frame
of a token represents its universal categories of events, concepts, and relationships, and can be regarded as a high-level abstraction of meaning. For example, tokens like bird, cat, dog, horse, sheep share the same frame label Animals. Here, we utilize the open-SESAME model (Swayamdipta et al., 2017) to extract semantic frames from responses.

The performance of tagging tools is listed in Table 1. Note that control signal tokens are concatenated into a flat sequence from coarse to fine.

### 3.2 Diffusion Model with Control-Range Masking

A diffusion model contains a forward process and a reverse process. We first concatenate a context with the control signals and corresponding response, i.e., \( w = w^a \oplus w^c \oplus w^y \). Then we use an embedding function (Li et al., 2022b) \( \text{EMB}(\cdot) \) to map the discrete text \( w \) into a continuous representation \( x_0 = u_0 \oplus c_0 \oplus y_0 \), where \( u_0, c_0 \), and \( y_0 \) represent parts of \( x_0 \) that belong to \( w^a \), \( w^c \), and \( w^y \), respectively.

**Forward Process.** In forward process \( q \), the model adds noise to the original sample \( x_0 \) step by step:

\[
q(x_t| x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t I), \tag{1}
\]

where \( x_1, \ldots, x_T \) make up a chain of Markov variants and \( x_T \sim \mathcal{N}(0, I) \). \( \beta_t \in (0, 1) \) is a noise schedule that controls the noise scale added in each step. Note that the conventional diffusion models corrupt the entire \( x_0 \). However, empathetic response generation is a conditional text generation (Seq2Seq) task and we only concern with the generative effect on response. Therefore, we use partial noising (Gong et al., 2022) to only impose noise on the parts of \( x_t \) that belong to \( w^y \), i.e., \( y_t \).

**Reverse process.** Once the forward process is completed, the reverse process aims to gradually recover \( x_0 \) by denoising \( x_T \) according to:

\[
p_{\theta}(x_{t-1}|x_t,t) = \mathcal{N}(x_{t-1}; \mu_{\theta}(x_t, t), \sigma_{\theta}(x_t, t)), \tag{2}
\]

where \( \mu_{\theta}(\cdot) \) and \( \sigma_{\theta}(\cdot) \) are predicted mean and standard variation of \( q(x_{t-1}|x_t) \) (derived using Bayes’ rule) in forward process and can be implemented by a Transformer (Vaswani et al., 2017) model \( f_{\theta} \). In the reverse process, we add a *rounding* step (Li et al., 2022b), parameterized by \( p_{\theta}(w|x_t) = \prod_{i=1}^{n} p_{\theta}(w_i|x_i) \), where \( p_{\theta}(w_i|x_i) \) is a softmax distribution.

**Control-Range Masking.** The non-autoregressive nature of conventional diffusion models make one input token can attend to all other tokens with the full self-attention mechanism to update its representation. Instead, we need to distinguish between tokens of control signals and responses, and further model the relationship between them with a mask matrix \( M \) and integrate it into the self-attention layer in Transformer:

\[
Q^{i+1}, K^{i+1}, V^{i+1} = h^i W_q, h^i W_k, h^i W_v, \tag{3}
\]

\[
S^{i+1} = \text{softmax}(\frac{Q^{i+1}K^{i+1}M}{\sqrt{d_k}}), \tag{4}
\]

\[
h^{i+1} = S^{i+1}V^{i+1}, \tag{5}
\]

where \( W_q, W_k \) and \( W_v \) are trainable parameters, \( h^i \) is the hidden state of the \( i \)-th transformer layer. \( d_k \) is the dimension of \( K \), which is used for scaling.

Basically, if token \( i \) controls \( j \), then the calculation of \( j \) is influenced by \( i \). In terms of implementation, we do not mask \( i \) when updating the representation of \( j \). Particularly, tokens at the same level, including IT signal tokens, SF signal tokens, and response tokens, are also designed to control each other, thus ensuring the overall logic and fluency of the generated responses. For example, it is reasonable that *Sympathizing* is followed by *Questioning* at the intent level, i.e., expressing more
concerns by questioning after showing sympathy for a negative situation or feeling. Therefore, to model the control relationship among tokens, we design the control-range masking and utilize it in the self-attention layer of $f_\theta$. Specifically, for a mask matrix, the value on position $(i, j)$ is 0 if token $i$ is controlled by token $j$; otherwise is negative infinity:

$$M(i, j) = \begin{cases} 0, & i \Rightarrow j \\ -\infty, & i \not\Rightarrow j \end{cases}$$

Figure 3 gives an example of control-range masking. For the intent signal Acknowledging (index 2), it is visible to Questioning (line 3) and corresponding response tokens Sounds great! in the first sentence (line 12-14). Meanwhile, since the response token great (line 13) is controlled by Exploration (index 1), Acknowledge (index 2), Desirability (index 5), and the rest of response tokens (index 12-19), it attends to them in the mask matrix.

With the existence of control-range masking, we can elaborately guide the generation of each response token with signals from different levels that reflect diverse factors for empathy expression.

3.3 Training and Inference

**Training.** In the training process, we label control signals based on golden responses as described in 3.1. To train model $f_\theta$ in the reverse process, we minimize the variational lower bound following Gong et al. (2022):

$$L_{vlb} = \sum_{t=2}^{T} \left| \left| y_0 - \tilde{f}_\theta(x_t, t) \right| \right|^2 + \left| \left| EMB(w^y) - \tilde{f}_\theta(x_1, 1) \right| \right|^2 + R(\left| \left| x_0 \right| \right|^2),$$

where $\tilde{f}_\theta(x_t, t)$ denotes the fractions of recovered $x_0$ corresponding to $y_0$, and $R(\cdot)$ is a mathematically equivalent regularization term to regularize the embedding learning.

**Inference.** During inference, since golden responses are unavailable, we design an emotion-enhanced matching method to obtain response candidates and use them to extract control signals. We treat dialogue contexts in the training set as the candidate pool and use each context in the test set as a query to perform context-context matching. Then the response corresponding to a returned context with the highest similarity is used as the candidate. Regarding the importance of emotions in empathetic response generation, we consider two aspects to score each candidate, semantic similarity and emotional consistency, in context-context matching. Specifically, we first train a BERT model (Devlin et al., 2019) on the training set to classify emotions for contexts. Then, we use this model to get emotional distribution for contexts in both the candidate pool and queries. Finally, we compute the cosine similarity of both sentence embeddings and predicted emotional distributions for each query-context pair. The contexts are re-ranked according to a weighted sum of two similarity scores:

$$Score = \text{SIM}_{\text{emotional}} + \gamma \text{SIM}_{\text{semantic}},$$

where $\gamma$ is a hyperparameter to balance the semantic and emotional similarity.

4 Experimental Setup

4.1 Dataset

EMPATHETICDIALOGUE (Rashkin et al., 2019) dataset comprises 24,850 open-domain multi-turn conversations between two interlocutors. Each conversation contains one emotion label, a situation where the speaker feels the exact emotion, and utterances about the speaker’s descriptions of the situation or the listener’s empathetic replies. There are 32 evenly-distributed emotion labels in the dataset. We apply the data provided by the original paper with the split ratio of 8:1:1 for training/validation/test set and use the script released by Lin et al. (2019) to preprocess the data.

4.2 Comparable Methods

We compare our method with three groups of representative methods.

**Transformer-Based Methods.** (1) TRS (Rashkin et al., 2019) is a vanilla Transformer with MLE loss. (2) MTRS (Rashkin et al., 2019) uses multi-task learning with emotion classification in addition to MLE loss. (3) MoEL (Lin et al., 2019) utilizes different decoders to combine different outputs for each emotion category. (4) MIME (Majumder et al., 2020) applies emotion grouping, emotion mimicry, and stochasticity strategies. (5) EmpDG (Li et al., 2020) learns emotions and responses based on adversarial learning. (6) CEM (Sahand Sabour, 2021) leverages commonsense to enhance empathetic response generation.

**Pre-Trained Language Model-Based Methods.** (1) TransferTransfo (Wolf et al., 2019) is a tran-
Table 2: Automatic evaluation results. The best results of standard settings are reported in the **bold** format. “ACC”, “D”, and “sBL” are abbreviations of Accuracy, Dist, and Self-BLEU, respectively. “ACC-CM” is the average Accuracy of ER, IP, and EX, which are three mechanisms of CM.

### 4.4 Implementation Details

**DIFFUSEMP** is based on the architecture of BERT-base (Devlin et al., 2019). For diffusion model settings, we adopt the square-root noise schedule (Li et al., 2022b) and set 2000 diffusion steps in the training and inference process. The maximum input length is 128 with WordPiece tokenizer and word embeddings are in the size of 128 with randomly initialized. For training settings, we use AdamW optimizer and set the learning rate as $1e^{-4}$. The batch size and dropout value are set as 128 and 0.1, respectively. $\gamma$ in Equation 8 equals to 0.2. For all comparable methods, we use their official codes with settings that follow the original papers. For more details, please refer to Appendix A.4.
Table 3: Human evaluation results. The Fleiss’ kappa (Fleiss and Cohen, 1973) of the results is 0.47, indicating a moderate level of agreement.

5 Results and Discussions

5.1 Main Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Empathy</th>
<th>Relevance</th>
<th>Informativeness</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRS</td>
<td>2.96</td>
<td>2.49</td>
<td>2.31</td>
</tr>
<tr>
<td>CEM</td>
<td>2.84</td>
<td>2.69</td>
<td>2.75</td>
</tr>
<tr>
<td>BART</td>
<td>3.04</td>
<td>2.94</td>
<td>3.92</td>
</tr>
<tr>
<td>DiffuSeq</td>
<td>2.77</td>
<td>2.66</td>
<td>3.74</td>
</tr>
<tr>
<td>DIFFUSEMP</td>
<td>3.68</td>
<td>3.39</td>
<td>4.63</td>
</tr>
</tbody>
</table>

Table 4: Ablation study on control-range masking and control signals.

Automatic Evaluation Results. The overall results are shown in Table 2. DIFFUSEMP substantially exceeds transformer-based and pre-trained model-based methods on almost all metrics. First, the improvement in controllability is significant. The high success rate indicates the effectiveness of control-range masking for elaborate token generation and demonstrates the ability of DIFFUSEMP to customize responses with desired factors. For informativeness, diffusion model-based methods perform the best, and DIFFUSEMP is even better than DiffuSeq. It has been proven that the diffusion model is a powerful backbone for generating diverse texts. With the integration of control signals, especially fine-grained signal SF, the meaning of each to-be-generated response token is more specific, thus the final response is more informative.

When considering informativeness values along with MIScore and AvgLen, we can find that those informative responses generated by DIFFUSEMP are also context-related and long, which satisfies the demand for proper responses to speakers. The BERTScore of DIFFUSEMP is not the highest, and we think this is reasonable since BERTScore indicates the similarity of generated and golden responses, while DIFFUSEMP encourages creativity instead of similarity. Besides, the difference between BERTScore and MIScore can justifiably mean that the generated responses are both creative and coherent.

Human Evaluation Results. Human evaluation results are listed in Table 3. Our method achieves the highest scores in all aspects, and the greatest improvement is achieved in informativeness, which shows that responses generated by DIFFUSEMP are preferred by annotators. Meanwhile, results of the Oracle setting show that the performance will be further improved when accurate control signals are given, which indicates that obtaining better control signals can be a feasible research topic.

5.2 Ablation Study

Ablation on Control-Range Masking. To verify the effectiveness of control-range masking, we remove the mask matrix and conduct full self-attention on all input tokens, i.e., input tokens can control or influence the representation of each other. As shown in Table 4, the controllability of three signals decreases when the mask is removed (“w/o Mask”), which justifies that our masking strategy is useful for multi-grained control. Besides, the most significant declines appear at the sentence level, which illustrates that IT has the strongest dependency on the masking strategy. We suppose it is because sentence-level signals are not that explicit like token-level signals with word-by-word alignments or utterance-level signals with global modeling in a dialogue session.

Ablation on Control Signals. Another question is whether each control signal plays the corresponding role. We keep the structure of the control-range mask untouched and remove each signal to validate. In detail, we remove the control signal from both the input text and the corresponding row(s) and column(s) in the original mask matrix. Table 4 shows that a success rate decreases when the corresponding control is removed (“w/o CM”, “w/o IT”, and “w/o SF”), and the finer the granularity of the control signal, the more the performance declines. We can come to the conclusion that each control signal and its control range defined in the mask matrix play an important role in response controllability.

5.3 Discussions

Analysis on Fine-Grained Signal SF. Compared with CoMAE (Zheng et al., 2021) which utilizes...
coarse control signals at the utterance level, we claim that a fine-grained signal is more useful for better empathy expression. To validate this claim, we remove the fine-grained labels, i.e., token-level SF, to see the performance change. Results are shown in Table 5. Without the token-level control, almost all evaluation metrics decrease in varying degrees. We conjecture that the token-level guidance gives a direct prompt on the content this token should entail, which greatly narrows the space of acceptable output generation.

**Analysis on Coarse-Grained Signal CM.** Emotional Reaction (ER), Interpretation (IP), and Exploration (EX) are three different high-level mechanisms for empathy expression. To explore the ways in which different mechanisms express empathy, we score generated responses in these three aspects with RoBERTa-based annotators as mentioned in Section 3.1. Results are visualized in Figure 4. For each method, the average ER, IP, and EX of generated responses on the test set are represented as the coordinate value of a point. DIFFUSEMP is the closest to human responses in distance, indicating that the way our method expresses empathy is the most similar to human beings.

### 5.4 Case Study

Table 6 shows the syntactically acceptable examples generated by DIFFUSEMP and other comparable methods. Transformer-based methods tend to generate plain and safe words, lacking a deep understanding of the context. In contrast, responses generated by TransferTransfo and BART have more rich information and details. All comparable methods tend to respond in general expressions, and even the way to ask questions is also monotonous, which may be due to the large number of such samples in the dataset. DIFFUSEMP responses entail features from both context and guidance. Feelings (*disgusting, don’t feel bad*), questions (*new relationship*), and advice (*study for future*) fit the situation of the speaker. Our framework is also helpful for generating different responses for a given context. With the support of an emotion-enhanced matching method, multiple response candidates can be returned to further guide response generation with diverse control signals. Control A and B contain intent *Suggesting* and *Questioning*, respectively. Thus, DIFFUSEMP A aims to give advice while B focuses on asking questions. More cases are shown in Appendix C.

### 6 Conclusion and Future Work

We propose DIFFUSEMP, a diffusion model-based framework, for empathetic response generation. To better model multi-dimensional empathy and improve its expression, we utilize multi-grained control signals at utterance, sentence, and token levels. These control signals are directly extracted from golden responses in the training process, while response candidates obtained from an emotion-enhanced matching method are used as the signal source. Then we also design a control-range masking strategy and integrate it into the diffusion language model to fulfill elaborate control on the generation of response tokens. Experimental results on a benchmark dataset EMPEHETICDIALOGUE show that our method outperforms compet-

---

**Table 6: Case study of DIFFUSEMP.**

<table>
<thead>
<tr>
<th>Context</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Golden</td>
<td>Wow, Dump him and beat him up!</td>
</tr>
<tr>
<td>MTRS</td>
<td>Oh no! What happened?</td>
</tr>
<tr>
<td>MIME</td>
<td>Oh no, did he get hurt?</td>
</tr>
<tr>
<td>CEM</td>
<td>What did he do?</td>
</tr>
<tr>
<td>TransferTransfo</td>
<td>That is terrible! Was he able to get back to you?</td>
</tr>
<tr>
<td>BART</td>
<td>Oh no! Did you confront him about it?</td>
</tr>
<tr>
<td>DiffuseSeq</td>
<td>Were you hurt?</td>
</tr>
</tbody>
</table>

| Candidate A |  |  |
|-------------|  |  |
| Control A   | **EMOTIONAL\_REACTION**  |
|             | SUGGESTING      |
|             | _ INTENTIONALLY\_ACT\(_IV\) No² PERCEPTION\_EXPERIENCE³ DESIRABILITY² _ EMOTION\_DIRECTED³ _ SCRUTINY\(_IV\) _ ALTERNATIVES\(_V\) |
| Response A  | Just do\(_V\) not\(_I\) feel\(_I\) bad\(_V\), happy\(_V\) to study\(_V\) in your future\(_I\), |

| Candidate B |  |  |
|-------------|  |  |
| Control B   | **EXPLORATION** QUESTIONING |
|             | _ POSSIBILITY\(_V\) _ INTENTIONALLY\_ACT\(_IV\) |
|             | PRONOUN\(_IV\) POSSESSION\(_IV\) AGE\(_IV\) |
| Response B  | That could\(_IV\) be disgusting, do\(_IV\) you have\(_IV\) a new\(_V\) relationship\(_V\) |

---

**Figure 4:** Visualization for CM of different methods.
itive baselines in generating more context-related, informative, and empathetic responses. Our framework is scalable for more control signal types and can also be extended to other controllable conditional text generation tasks.

In future work, we will extend DIFFUSEMP to more empathetic control signals, and improve the performance of annotators and retrieval tools. Besides, it is interesting to explore DIFFUSEMP on various controllable text generation tasks.

Acknowledgement

We thank the reviewers for their detailed and insightful advice. This work is supported by the National Key Research and Development Program of China (NO.2022YFB3102200) and Strategic Priority Research Program of the Chinese Academy of Sciences with No. XDC02030400.

Limitations

The difficulty of obtaining accurately-labeled control signals constrains our results. As we report in Table 1, the performance of tagging tools can be further improved. However, when the original dataset lacks multi-grained annotations, relying on pre-trained tools is the most feasible solution. Considering that control signals come from response candidates in the inference stage, the performance of the context-context matching method is another constraint. Finally, the drawback of diffusion models also has an impact on our approach. Despite its high-quality generative performance, the diffusion model has a high requirement for GPU resources and still suffers from slow sampling. We discuss some attempts to address these limitations in Appendix B.

Ethics Statement

The EMPATHETICDIALOGUE dataset (Rashkin et al., 2019) used to train and evaluate in the paper is collected by crowd-sourcing using the ParlAI platform to interact with Amazon Mechanical Turk. Besides, we use EMPATHETICINTENT (Welivita and Pu, 2020), REDDIT (Sharma et al., 2020) and FRAMENET (Baker et al., 1998) to train tagging tools for control signals. All the above datasets are well-established and publicly available. Sensitive and personal privacy information have been removed during the dataset construction. In our human evaluation, participants were fully informed of the purpose of our study and were appropriately compensated. It is important to clarify that our work is only a study of open-domain dialogue with empathy. We claim that our system does not provide professional psychological counseling. In other words, it does not make any treatment recommendations or diagnostic claims.

References


Mark H. Davis, Miles P. Davis, M Davis, Matthew Davis, Mark Davis, Mn Davis, M Davis, F. Caroline Davis, Heather A Davis, and Ilus W. Davis. 1980. A multidimensional approach to individual differences in empathy.


Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022a. Knowledge bridging for empathetic dialogue generation. In *AAAI*.


Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Saintry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language...


A Additional Experiment Details

A.1 Comparable Methods

The following models are chosen as comparable methods and divided into three groups according to their architecture.

Transformer-Based Methods.

- **TRS** (Rashkin et al., 2019): A vanilla Transformer with maximum likelihood estimation (MLE) loss.
- **MTRS** (Rashkin et al., 2019): A multi-task model trained with emotion classification loss in addition to MLE loss.
- **MoEL** (Lin et al., 2019): A model using different decoders to generate and combine different outputs for each emotion category.
- **MIME** (Majumder et al., 2020): A model utilizing emotion grouping, emotion mimicry, and stochasticity strategies to generate responses.
- **EmpDG** (Li et al., 2020): An adversarial model applying two discriminators for interacting with user feedback.
- **CEM** (Sahand Sabour, 2021): A model leverages commonsense as additional information to further enhance empathetic response generation.

Pre-Trained Language Model-Based Methods.

- **TransferTransfo** (Radford et al., 2019; Wolf et al., 2019): A combination of a transfer learning-based training scheme and a high-capacity GPT-2 model which shows strong improvements over end-to-end conversational models.
- **BART** (Lewis et al., 2020): A pre-trained encoder-decoder Transformer with great success in many seq2seq tasks.

Diffusion Model-Based Methods.

- **DiffuSeq** (Gong et al., 2022): A diffusion model proposed as a conditional language model and trained end-to-end in a classifier-free manner. It is designed for sequence-to-sequence text generation tasks.

Noted that we did not use Diffusion-LM (Li et al., 2022b) as a baseline because it is incompatible with the sequence-to-sequence task setting. We provide the result of oracle setting as a reference. Under the standard setting, the attributes are not given and need to be predicted from the retrieve-based methods, and we focus on evaluating the response quality. Under the oracle setting, the true attributes from the ground truth response are provided, so it can be considered as the theoretical upper limit performance of DIFFUSEMP.

A.2 Automatic Evaluation

We evaluate the generated empathetic responses from the following four aspects: relevance, controllability, informativeness, and response length.

**Relevance.** We use BertScore and the MIScore of response to evaluate relevance.

- **BertScore** (Zhang et al., 2020a): BertScore computes a similarity score using contextual embeddings for each token in the candidate sentence with each token in the reference sentence. We use deberta-large-mnli to calculate the BertScore.
- **MIScore**: A good response should be informative and relevant to the context. When given the response, it should have the ability to infer its context, while a safe response is generic and can be used in any context, so it is hard to infer the context. From this perspective, we use the idea of Maximum Mutual Information (MMI) (Li et al., 2016; Zhang et al., 2018). The idea of MIScore is employing a pre-trained backward model to predict context sentences from given responses, i.e., \( P(\text{Context}|\text{Response}) \). Intuitively, MIScore encourages the model to generate responses that are more specific to the context, while generic responses are largely less preferred, since they can be used in any case. We calculate MIScore according to the following equation:

\[
\exp\left(-\frac{1}{m} \sum_{t=1}^{m} \log P(x_t|y_1, \ldots, y_n, x_{<t})\right),
\]

where \( m \) and \( n \) are the numbers of tokens in the context and response respectively. It is implemented with a reverse 345M Di-aloGPT (Zhang et al., 2020b), which is a fine-tuned GPT-2 (Radford et al., 2019) with the
training objective to predict the context from the response.

**Controllability.** We calculate the attribute control accuracy success rate to validate the controllability of models. For session-level CM and sentence-level IT, we report accuracy. For token-level SF, we report F1.

**Informativeness.** We use Distinct n-gram (Li et al., 2016) and self-BLEU (Zhu et al., 2018) to evaluate informativeness.

- **Distinct n-gram** (Li et al., 2016): Distinct n-gram calculates the number of distinct n-grams in generated responses. The value is scaled by the total number of generated tokens to avoid favoring long sentences.

- **Self-BLEU** (Zhu et al., 2018): Self-BLEU regards one sentence as a hypothesis and the others as a reference, we can calculate the BLEU score for every generated sentence, and define the average BLEU score to be the Self-BLEU of the document.

**Response Length.**

- **Average Length** (Singh and Jin, 2016): The length of the response text is also used as a quality indicator when comparing different model generations since shorter texts usually contain less information.

It is noteworthy that open-domain dialogue and controllable text generation contain a great deal of creativity. When a sentence is forced to remain identical to a fixed standard sentence, such evaluation metrics may unfairly penalize creative texts, notwithstanding they are capable of responding to the given context. As a result, instead of comparing the word overlap between generated responses and standard responses, we give the metric values of standard responses as a reference.

A.3 Human Evaluation

Quantitative automatic metrics are straightforward to compare, but they may be less effective at reflecting overall levels of empathy. Human judgment is necessary for an open-domain dialogue system (Liu et al., 2016).

We recruit three third-party graduate researchers (average age 23.3) to analyze the results of various models. We acquired permission for their participation and paid them in accordance with local hourly wages. The response quality of all models is evaluated in terms of the following three aspects: Empathy, Relevance, and Informativeness. We randomly sample 100 dialogues and corresponding generated responses for different models and then ask three professional annotators to give each response a rating score from the following aspects.

- **Empathy** reflects whether the listener understands the feeling of the speaker and responds appropriately.

- **Relevance** considers how the content of the reply is relevant to the topic mentioned by the speaker.

- **Informativeness** evaluates grammar correctness and readability.

The specific instruction given to them for the evaluation is shown in Figure 5. Each aspect is on a scale of 1 to 5, in which 1 is “unacceptable” and 5 is “excellent performance”.

Besides, We conduct an A/B test to directly compare our method with other baselines. Another 100 dialogues are randomly sampled from each model. Three annotators are given generated responses from either our method or baselines in random order and are asked to choose a better one. They can either choose one of the responses or select “Tie” when the quality of provided options is hard to access.

A.4 Implementation Details

Our DIFFUSEMP calculates diffusion model parameters with a BERT-base (Devlin et al., 2019) architecture with 12 layers and 80M parameters. For diffusion settings, we set 2000 diffusion steps in both the training stage and the inference stage. We adopt the square root noise schedule. The max input length is 128, the dimensions of word embedding and time embedding are all 128, and the embedding is randomly initialized\(^*\). For training settings, we use AdamW optimizer and set the learning rate as 1e-4, dropout as 0.1. We set gradient clipping to \(-1.0\). \(\gamma\) equals to 0.2. We use WordPiece tokenizer\(^\dagger\). The batch size is 128 and the micro-batch size is 64. For all baseline models, we use their official codes to implement and keep the settings in the original paper.

\(^*\)We also attempt the initialization with pre-trained bert-base-uncased vocabulary but the result is poor.

\(^\dagger\)Firstly we try to build vocabulary for our own dataset but find it heavily suffers from the out-of-vocabulary problem.
B Future Work

The limitations of our work have been mentioned in Section 6. Here, we propose some attempts to overcome these limitations.

**Control Signals.** In the acquisition of control signals, there are two main constraints for performance, including (1) the accuracy of control signals and (2) the suitability of retrieval results in the testing step.

With regard to (1), the results of the oracle setting demonstrate that our framework has a high ceiling when ground-true control signals are given. Therefore, we have tried to enhance robustness by noising the control factors. Noising methods contain adding, removing, and replacing random control tokens. However, experimental results show that noising methods compromise the success rate of control, which is contrary to the motivation of this work. In the future, this approach can be tried to further improve language quality in scenarios where the demand for controllability is weak.

With respect to (2), we focus on the performance of the retrieval model in the inference stage. The control signals straightforwardly come from the retrieved responses. In this paper, we have proposed a task-specific design that combines semantic and emotional similarity to retrieve but it is still simple compared to those SOTA dialogue response selection models. In future work, it is meaningful to replace our retrieval model with more powerful response selection methods.

As an advantage of DIFFUSEMP, both the annotating taggers and the retrieval model are orthogonal to empathetic response generation. It is easy for followers to employ higher-performance response selection models and attribute annotating taggers to empower the DIFFUSEMP.

**Diffusion Models.** Finally, the diffusion model requires a lot of GPU computational resources and is slow when inference, which limits its application. There are many attempts to reduce the computational resources (Rombach et al., 2021a) required by the diffusion model as well as to speed up the process (Vahdat et al., 2021) and inference (Song et al., 2021; Bao et al., 2022). Theoretically, the relevant improvements would also have an enhancing effect on our framework and would be helpful for spreading the diffusion model to the NLP community.

C Case Study

We give more responses generated by DIFFUSEMP in Table 7 and Table 8.
Case 1

<table>
<thead>
<tr>
<th>Context</th>
<th>I feel lonely sometimes cause I live in a house with 5 girls and all of them have boyfriends and sometimes they are all hanging out with their significant others and I feel lonely.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold</td>
<td>Awww, That sucks man. Hopefully they don’t say mean things to you about being lonely.</td>
</tr>
<tr>
<td>TRS</td>
<td>I am sorry to hear that. I am sorry to hear that.</td>
</tr>
<tr>
<td>Multi-TRS</td>
<td>I am sorry to hear that. I am sorry to hear that.</td>
</tr>
<tr>
<td>MoEL</td>
<td>I am sorry to hear that. have you tried talking to them?</td>
</tr>
<tr>
<td>MIME</td>
<td>I am sorry to hear that. I am sorry to hear that. I hope you are able to do that.</td>
</tr>
<tr>
<td>EmpDG</td>
<td>That is a shame. I am sorry to hear that.</td>
</tr>
<tr>
<td>CEM</td>
<td>I am so sorry to hear that. I am sure you will be fine.</td>
</tr>
<tr>
<td>GPT2-FT</td>
<td>Why were they so angry?</td>
</tr>
<tr>
<td>BART-FT</td>
<td>I’m sorry to hear that. Have you tried joining any clubs?</td>
</tr>
<tr>
<td>DiffuSeq</td>
<td>Why can not you too</td>
</tr>
<tr>
<td>Candidate A</td>
<td>I see. I’m sorry you’re home alone. I get lonely my self when you’re used to someone around. are you going to advertise for a new roommate?</td>
</tr>
<tr>
<td>Control A</td>
<td>EMOTIONAL_REACTION EXPLORATION INTERPRETATION ACKNOWLEDGING SYMPATHIZING NEUTRAL QUESTIONING PRONOUN PERCEPTION_experience _ PRONOUN _ PRONOUN _ BUILDINGS _ PRONOUN TRANSITION_to_state _ WHEN PRONOUN _ USING _ PEOPLE _ _ _ PRONOUN MOTION _ _ _ AGE _ ?</td>
</tr>
<tr>
<td>Response A</td>
<td>I see. I’m sorry you’re home alone. I got to lonely when I’ve used to noone with. are you going to find a brand new friend?</td>
</tr>
<tr>
<td>Candidate B</td>
<td>Maybe you could invite a friend over for a movie?</td>
</tr>
<tr>
<td>Control B</td>
<td>EXPLORATION SUGGESTING _ PRONOUN POSSIBILITY REQUEST _ PERSONAL _ RELATIONSHIP _ _ _ ?</td>
</tr>
<tr>
<td>Response B</td>
<td>maybe you could call a friend to talk with her?</td>
</tr>
<tr>
<td>Candidate C</td>
<td>Oh no really, that really sucks to hear. Why were you not able to go?</td>
</tr>
<tr>
<td>Control C</td>
<td>EMOTIONAL_REACTION EXPLORATION ACKNOWLEDGING QUESTIONING _ NO _ _ _ _ _ PERCEPTION_EXPERIENCE _ WHY _ PRONOUN NO CAPABILITY _ MOTION ?</td>
</tr>
<tr>
<td>Response C</td>
<td>Oh no, that is really amazing to hear. Why were you not able to move?</td>
</tr>
</tbody>
</table>

Table 7: Cases generated by DIFFUSEMP with different control signals.
### Case 2

<table>
<thead>
<tr>
<th>Context</th>
<th>Today traffic was horrible and was so frustrating!</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold</td>
<td>I hate traffic too, it makes me angry</td>
</tr>
<tr>
<td>TRS</td>
<td>Oh no! What happened?</td>
</tr>
<tr>
<td>Multi-TRS</td>
<td>Oh no! I hate when that happens. Did you get hurt?</td>
</tr>
<tr>
<td>MoEL</td>
<td>I am sorry to hear that. What happened?</td>
</tr>
<tr>
<td>MIME</td>
<td>I am sorry to hear that. What happened?</td>
</tr>
<tr>
<td>EmpDG</td>
<td>I hate when that happens.</td>
</tr>
<tr>
<td>CEM</td>
<td>What happened?</td>
</tr>
<tr>
<td>GPT2-FT</td>
<td>Why were they so angry?</td>
</tr>
<tr>
<td>BART-FT</td>
<td>I hate traffic too. I hate it when it’s so bad.</td>
</tr>
<tr>
<td>DiffuSeq</td>
<td>Was it hurt you?</td>
</tr>
<tr>
<td>Candidate A</td>
<td>I completely understand. I get pretty aggravated myself driving home from work everyday.</td>
</tr>
<tr>
<td>Control A</td>
<td>INTERPRETATION EMOTIONAL _REACTION AGREEMENT AGREEING PRONOUN _ AWARENESS _ PRONOUN _ EXPERIENCER OBJ _ SUBJECTIVE INFLUENCE BUILDINGS _ WORK _</td>
</tr>
<tr>
<td>Response A</td>
<td>I completely understand. I have been tired to drive home from work everyday.</td>
</tr>
<tr>
<td>Candidate B</td>
<td>Yes! What’s even worse is when other people don’t pay attention in bad traffic!</td>
</tr>
<tr>
<td>Control B</td>
<td>INTERPRETATION SUGGESTING QUESTIONING YES _ NO COMMERCES _ PAY ATTENTION _ DESIRABILITY _</td>
</tr>
<tr>
<td>Response B</td>
<td>Yes! Traffic is the worst but other people don’t pay attention to bad thing.</td>
</tr>
<tr>
<td>Candidate C</td>
<td>Yes, the cable company is infuriating. do they eventually help you though?</td>
</tr>
<tr>
<td>Control C</td>
<td>EXPLORATION NEUTRAL QUESTIONING YES _ BUSINESSNESS _ INTENTIONALLY _ ACT PRONOUN TIME VECTOR ASSISTANCE PRONOUN CONCESSIVE?</td>
</tr>
<tr>
<td>Response C</td>
<td>Yes, the bus company was annoying. Did they already help you out?</td>
</tr>
</tbody>
</table>

Table 8: Cases generated by DIFFUSEMP with different control signals.
Empathetic Response Evaluation

We are a group of natural language processing researchers working on empathetic dialogue generation. Here are some dialogue contexts and responses. For every context-response pair, you will be asked to rate it according to three measures described as below.

Conversation 1:

Context: I have been awarded a degree in marketing.
Response: Congratulations! Marketing is a very lucrative field.

- **Empathy:** whether the response puts itself in the other person’s shoes, understands the feelings of others, and thinks and deals with problems from the other person’s point of view.
  - 1: Completely not empathetic, can be offensive, or can lead to negative emotion in the speaker.
  - 2: A little bit empathetic, contains little words that express understanding or help
  - 3: Empathetic, mentioned the emotion or convey the understanding, but not in depth
  - 4: Somewhat empathetic, reaction to the speaker’s feeling or understand and interprets the experience.
  - 5: Very empathetic, specify the speaker’s feeling or experiences, explore some key question about the situation, give substance help

- **Relevance:** whether the response is relevant to the dialogue history and consistent with the speaker's background situation.
  - 1: Completely irrelevant with the contexts, or inconsistent with dialogue history or background situation.
  - 2: A little bit relevant to the context, but with many conflicts to the dialogue history and background situation.
  - 3: Relevant to the contexts, but with some conflicts to the dialogue history or background situation.
  - 4: Very relevant to the contexts, but with minor conflicts to the dialogue history or background situation.
  - 5: Completely relevant and coherent to the dialogue contexts and background situation.

- **Informativeness:** whether the response is interesting, has rich content and encourages you to continue the conversation.
  - 1: Very generic or boring. Do not want to continue this conversation.
  - 2: Contains little information, or information is not specific.
  - 3: Contains some information, but somewhat generic and ambiguous.
  - 4: Contains rich and substantive information.
  - 5: Very interesting and informative, you want to continue this conversation.

Figure 5: An example of the survey for our human evaluation.