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

Guanqun Bi[1,2], **Lei Shen**[3], **Yanan Cao**[1,2]*, **Meng Chen**[3]*,
Yuqiang Xie[1,2], **Zheng Lin**[1,2], **Xiaodong He**[3]

[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3]JD AI Research, Beijing, China
{biguanqun,caoyanan,xieyuqiang,linzheng}@iie.ac.cn
{shenlei20,chenmeng20,xiaodong.he}@jd.com

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 EMPATHETICDIALOGUE 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
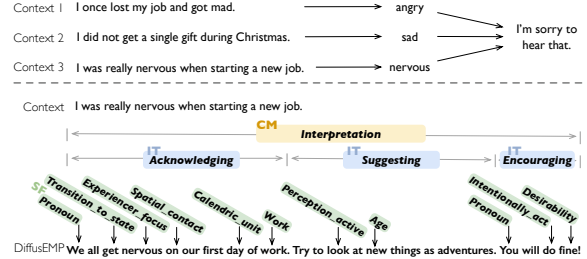
---
*Corresponding authors.

Figure 1: A monotonous empathetic response (upper) and an informative empathetic response (lower). "CM", "IT", and "SF" are abbreviations for "Communication Mechanism", "Intent", and "Semantic Frame", which represent control signals at the utterance, sentence, and token level respectively.

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 re-

sponses 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

---

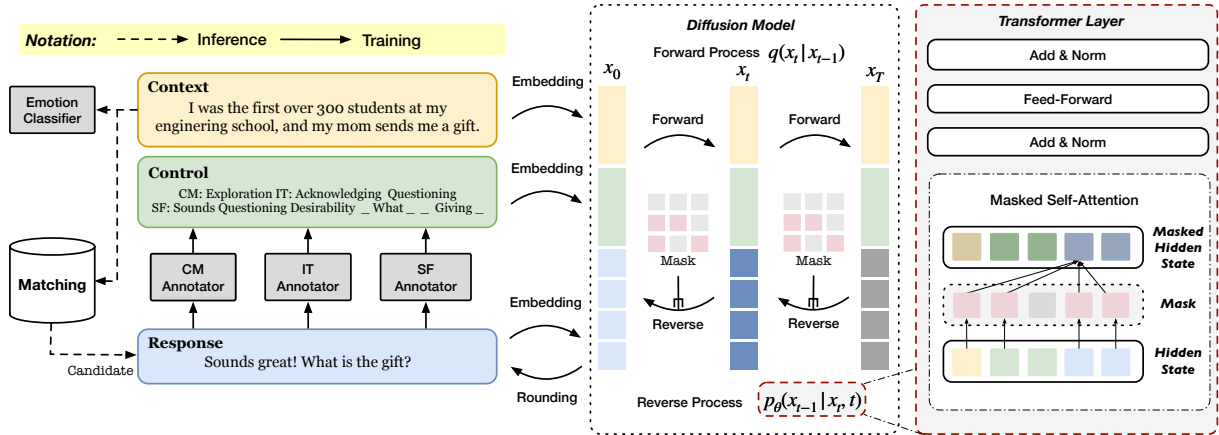[†]An utterance, the response here, may consist of more than one sentence.

Figure 2: The overview of DIFFUSEMP. The left part describes the training and inference stages, the middle part shows the forward process and reverse process in the diffusion model, and the right part illustrates details in a Transformer (Vaswani et al., 2017) block with control-range masking for the reverse process.

the forward process. DiffusionBERT (He et al., 2022) combines pretrained language models with absorbing-state discrete diffusion models for text. To the best of our knowledge, we are the first to achieve controllable empathetic response generation using a diffusion model.

3 DIFFUSEMP

In this paper, we perform empathetic response generation in a controllable setting. The dialogue context is an alternating sequence of utterances from a speaker and a listener, i.e. $\mathbf{w}^u = \{u_1, u_2, \ldots, u_n\}$. Here, we aim to generate an empathetic and context-related response $\mathbf{w}^y = \{y_1, y_2, \ldots, y_n\}$ conditioned on the given context $\mathbf{w}^u$ and a set of control signals $\mathbf{w}^c$ obtained in advance (Section 3.1). Then, the context, control signals, and response are concatenated and fed into a diffusion model with control-range masking (Section 3.2). In the training process, golden responses are used to extract control signals, while during inference, we integrate an emotion-enhanced matching method to get proper response candidates (Section 3.3). The framework of DIFFUSEMP is illustrated in Figure 2.

3.1 Acquisition of Control Signals

To better model and express multi-dimensional empathy, we use control signals at different levels. However, the benchmark dataset EMPATHETICDI-ALOGUE does not contain such annotations. Here, we introduce three types of signals used in this paper and the way to collect them for each golden response or response candidate using pre-trained tagging models. The definition and components of 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

Signal Type	Accuracy	F1	#Classes
CM-ER	79.43	74.46	2
CM-IP	84.04	62.60	2
CM-EX	92.61	72.58	2
IT	87.75	87.71	9
SF	-	86.55	1222

Table 1: The performance of tagging tools used to get control signals. Since SF is from a frame semantic parsing task, we only report the F1 score following the original task setting.

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., $\mathbf{w} = \mathbf{w}^u \oplus \mathbf{w}^c \oplus \mathbf{w}^y$. Then we use an *embedding* function (Li et al., 2022b) EMB($\cdot$) to map the discrete text $\mathbf{w}$ into a continuous representation $\mathbf{x}_0 = \mathbf{u}_0 \oplus \mathbf{c}_0 \oplus \mathbf{y}_0$, where $\mathbf{u}_0$, $\mathbf{c}_0$, and $\mathbf{y}_0$ represent parts of $\mathbf{x}_0$ that belong to $\mathbf{w}^u$, $\mathbf{w}^c$, and $\mathbf{w}^y$, respectively.

Forward Process. In forward process q, the model adds noise to the original sample $\mathbf{x}_0$ step by step:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \quad (1)$$

where $\mathbf{x}_1, ..., \mathbf{x}_T$ make up a chain of Markov variants and $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$. $\beta_t \in (0, 1)$ is a noise schedule that controls the noise scale added in each step. Note that the conventional diffusion models corrupt the entire $\mathbf{x}_0$. 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 $\mathbf{x}_t$ that belong to $\mathbf{w}^y$, i.e., $\mathbf{y}_t$.

Reverse process. Once the forward process is completed, the reverse process aims to gradually recover $\mathbf{x}_0$ by denoising $\mathbf{x}_T$ according to:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_\theta(\mathbf{x}_t, t)), \quad (2)$$
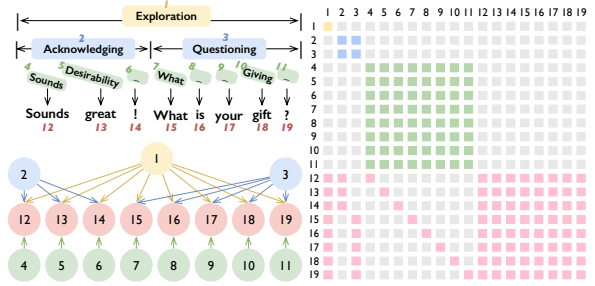

Figure 3: An example of control signals and control-range masking. The upper left part shows a response with labeled signals, the lower left part illustrates the control range of each signal on response tokens, and the right part is the corresponding mask matrix. "-" means the SF signal is empty.

where $\mu_\theta(\cdot)$ and $\sigma_\theta(\cdot)$ are predicted mean and standard variation of $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ (derived using Bayes' rule) in forward process and can be implemented by a Transformer (Vaswani et al., 2017) model f_θ. In the reverse process, we add a *rounding* step (Li et al., 2022b), parameterized by $p_\theta(\mathbf{w}|\mathbf{x}_0) = \prod_{i=1}^{n} p_\theta(w_i|x_i)$, 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:

$$Q^{i+1}, K^{i+1}, V^{i+1} = h^i W_q, h^i W_k, h^i W_v, \quad (3)$$

$$S^{i+1} = softmax(\frac{Q^{i+1}K^{i+1\mathsf{T}} + M}{\sqrt{d_k}}), \quad (4)$$

$$h^{i+1} = S^{i+1}V^{i+1}, \quad (5)$$

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_θ. Specifically, for a mask matrix, the value on position (i, j) is 0 if $token_j$ is controlled by $token_i$; otherwise is negative infinity:

$$M(i,j) = \begin{cases} 0, & i \Rightarrow j \\ -\inf, & i \not\Rightarrow j \end{cases} \quad (6)$$

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_θ in the reverse process, we minimize the variational lower bound following Gong et al. (2022):

$$\begin{aligned} \mathcal{L}_{\text{vlb}} = \sum_{t=2}^{T} & ||\mathbf{y}_0 - \tilde{f}_\theta(\mathbf{x}_t, t)||^2 \\ & + ||\text{EMB}(\mathbf{w}^y) - \tilde{f}_\theta(\mathbf{x}_1, 1)||^2 \\ & + \mathcal{R}(||\mathbf{x}_0||^2), \end{aligned} \quad (7)$$

where $\tilde{f}_\theta(\mathbf{x}_t, t)$ denotes the fractions of recovered $\mathbf{x_0}$ corresponding to $\mathbf{y}_0$, and $\mathcal{R}(\cdot)$ 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:

$$Score = \text{SIM}_{\text{semantic}} + \gamma \text{SIM}_{\text{emotional}}, \quad (8)$$

where γ 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 trans-

Method	#Params	Relevance		Controllability			Informativeness				Length
		BERTScore ↑	MIScore ↓	ACC-CM ↑	ACC-IT ↑	F1-SF ↑	D1 ↑	D2 ↑	D4 ↑	sBL ↓	AvgLen ↑
Transformer-Based Methods											
TRS	15M	0.5717	4598.26	60.98	22.07	15.74	0.42	1.55	4.26	13.63	10.53
MTRS	15M	0.5735	7156.26	60.48	25.77	15.62	0.50	1.89	5.56	11.26	9.92
MoEL	21M	0.5758	14595.61	59.29	26.20	16.51	0.40	1.65	4.62	12.83	11.47
MIME	17M	0.5800	4878.71	61.16	22.00	16.54	0.26	0.87	2.15	14.21	11.12
EmpDG	29M	0.5745	9088.11	61.94	20.06	17.36	0.60	2.54	7.75	11.78	10.11
CEM	17M	0.5713	7635.05	62.28	30.09	14.20	0.54	2.00	4.98	9.13	8.25
Pre-Trained Language Model-Based Methods											
TransferTransfo	117M	0.5634	2138.39	59.70	25.08	18.39	2.81	17.22	36.54	2.68	11.40
BART	140M	**0.5977**	706.31	60.39	30.69	18.98	**2.88**	14.12	38.82	2.79	11.09
Diffusion Model-Based Methods											
DiffuSeq	91M	0.5101	715.95	59.23	28.58	17.26	1.79	26.97	**88.17**	1.29	10.30
DIFFUSEMP	91M	0.5205	**626.92**	**92.36**	**84.24**	**52.79**	2.84	**29.25**	73.45	**1.09**	**14.12**
References											
DIFFUSEMP (Oracle)	91M	0.7458	615.13	92.38	83.66	51.95	2.84	30.46	89.35	1.11	14.01
Human	-	1.0000	507.97	100.00	100.00	98.40	19.49	43.55	49.02	0.85	13.04

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.

fer learning-based GPT-2 (Radford et al., 2019) model fine-tuned on EMPATHETICDIALOGUE. (2) BART (Lewis et al., 2020) is a pre-trained encoder-decoder Transformer with great success in many seq2seq tasks.

Diffusion Model-Based Method. DiffuSeq (Gong et al., 2022) is proposed as a conditional diffusion language model for seq2seq tasks.

Two more results are provided as references. Under the Oracle setting, control signals are obtained from golden responses in the test set, which can be regarded as the upper bound of DIFFUSEMP. Golden responses themselves are also evaluated, which reflects human performance on the task. More details are listed in Appendix A.1.

4.3 Metrics

Automatic Evaluation. We evaluate the generated responses from four aspects: (1) Relevance: *BERTScore* (Zhang et al., 2020a) computes a semantic similarity between generated responses and golden references. *MIScore* is the likelihood of generating a context with the given response, which applies the idea of Maximum Mutual Information (MMI) (Li et al., 2016; Zhang et al., 2018) and indicates whether the generated response is context-related. (2) Controllability: We calculate the success rate of empathy expression with multi-grained control signals to validate the controllability of DIFFUSEMP. For utterance-level CM and sentence-level IT, we report Accuracy, while for token-level SF, we report F1. (3) Informativeness: *Dist-n* (Li et al., 2016) calculates the number of distinct n-grams in generated responses. *Self-BLEU* (Zhu

et al., 2018) reflects the difference of all generated responses to a large extent. We calculate the average BLEU-5 overlap between each two generated responses. (4) Response Length: *AvgLen* represents the average number of tokens for generated responses. Intuitively, too short text often fails to convey good content. More details about automatic metrics are shown in Appendix A.2.

Human Evaluation. We evaluate the response quality based on the following aspects: (1) *Empathy* reflects whether a response understands the speaker's feeling or situation and responds appropriately. (2) *Relevance* considers whether a response is relevant to the topic mentioned by the speaker. (3) *Informativeness* evaluates whether a response provides rich and meaningful information. More details about the human evaluation guidance are given in Appendix A.3.

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 random initialization. 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. γ 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.

Method	Empathy	Relevance	Informativeness
TRS	2.96	2.49	2.31
CEM	2.84	2.69	2.75
BART	3.04	2.94	3.92
DiffuSeq	2.77	2.66	3.74
DIFFUSEMP	**3.68**	**3.39**	**4.63**

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

Method	CM		IT		SF
	ACC ↑	F1 ↑	ACC ↑	F1 ↑	F1 ↑
DIFFUSEMP	92.36	90.26	84.24	77.15	52.79
w/o Mask	90.76	87.99	73.80	66.58	49.43
w/o CM	89.34	85.55	83.80	76.38	52.89
w/o IT	92.24	90.21	47.92	41.77	52.63
w/o SF	89.70	86.96	83.12	74.90	22.48

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 justify 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

		DIFFUSEMP	w/o SF
Relevance	BERTScore ↑	52.05	51.47
	MIScore ↓	626.92	993.44
Informativeness	Dist-1 ↑	2.84	1.69
	Dist-2 ↑	29.26	22.83
	self-BLEU ↓	1.09	1.31
Length	AvgLen ↑	14.13	13.23

Table 5: Importance of the fine-grained signal SF.
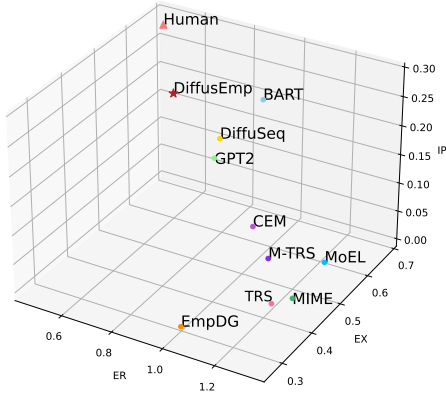
Figure 4: Visualization for CM of different methods.

Context	I caught my boyfriend texting his ex.
Golden	Wow. Dump him and beat him up!
MTRS	Oh no! What happened?
MIME	Oh no, did he get hurt?
CEM	What did he do?
TransferTransfo	That is terrible! Was he able to get back to you?
BART	Oh no! Did you confront him about it?
DiffuSeq	Were you hurt?
Candidate A	Ok do[1] not[2] feel[3] bad[4] be happy[5] and search[6] for bad future[7] behalf
Control A	EMOTIONAL_REACTION SUGGESTING _ INTENTIONALLY_act[1] No[2] PERCEPTION_EXPERIENCE[3] DESIRABILITY[4] _ EMOTION_DIRECTED[5] _ SCRUTINY[6] _ _ ALTERNATIVES[7] _
Response A	Just do[1] not[2] feel[3] bad[4], happy[5] to study[6] in your future[7].
Candidate B	That could[1] be embarrassing, do[2] you[3] have[4] a new[5] partner ?[6]
Control B	EXPLORATION QUESTIONING _ POSSIBILITY[1] _ _ _ INTENTIONALLY_act[2] PRONOUN[3] POSSESSION[4] _ AGE[5] _ ?[6]
Response B	That could[1] be disgusting, do[2] you[3] have[4] a new[5] relationship ?[6]

Table 6: Case study of DIFFUSEMP.

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 EMPATHETICDIALOGUE show that our method outperforms compet-

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

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 Tunk. 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

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. In *Neural Information Processing Systems*.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. 2022. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. *ArXiv*, abs/2201.06503.

Mao Yan Chen, Siheng Li, and Yujiu Yang. 2022. 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, Seattle, United States. Association for Computational Linguistics.

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

Frans B.M. de Waal. 2008. Putting the altruism back into altruism: The evolution of empathy. *Annual Review of Psychology*, 59:279–300.

Jean Decety and Meghan L. Meyer. 2008. From emotion resonance to empathic understanding: A social developmental neuroscience account. *Development and Psychopathology*, 20:1053 – 1080.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Prafulla Dhariwal and Alex Nichol. 2021. Diffusion models beat gans on image synthesis. *ArXiv*, abs/2105.05233.

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

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

Zhengfu He, Tianxiang Sun, Kuanning Wang, Xuanjing Huang, and Xipeng Qiu. 2022. Diffusionbert: Improving generative masked language models with diffusion models. *ArXiv preprint*, abs/2211.15029.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forr'e, and Max Welling. 2021. Argmax flows and multinomial diffusion: Learning categorical distributions. In *Neural Information Processing Systems*.

Shaojie Jiang, Pengjie Ren, Christof Monz, and Maarten de Rijke. 2019. Improving neural response diversity with frequency-aware cross-entropy loss. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 2879–2885. ACM.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020. EmpDG: Multi-resolution interactive empathetic dialogue generation. In *Proceedings of the 28th International Conference on Computational Linguistics*,

pages 4454–4466, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori Hashimoto. 2022b. Diffusion-lm improves controllable text generation. *ArXiv*, abs/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, Hong Kong, China. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692.

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

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail A. Kudinov. 2021. Grad-tts: A diffusion probabilistic model for text-to-speech. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8599–8608. PMLR.

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

supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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

Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021a. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021b. High-resolution image synthesis with latent diffusion models.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2021. Cem: Commonsense-aware empathetic response generation. In *AAAI Conference on Artificial Intelligence*.

Minlie Huang Sahand Sabour, Chujie Zheng. 2021. Cem: Commonsense-aware empathetic response generation. *ArXiv preprint*, abs/2109.05739.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.

Lei Shen, Jinchao Zhang, Jiao Ou, Xiaofang Zhao, and Jie Zhou. 2021. Constructing emotional consensus and utilizing unpaired data for empathetic dialogue generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3124–3134, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Abhishek Singh and Wei Jin. 2016. Ranking summaries for informativeness and coherence without reference summaries. In *FLAIRS*.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Swabha Swayamdipta, Sam Thomson, Chris Dyer, and Noah A. Smith. 2017. Frame-semantic parsing with softmax-margin segmental rnns and a syntactic scaffold. *ArXiv*, abs/1706.09528.

Jaesung Tae, Hyeongju Kim, and Taesu Kim. 2021. Editts: Score-based editing for controllable text-to-speech. In *Interspeech*.

Arash Vahdat, Karsten Kreis, and Jan Kautz. 2021. Score-based generative modeling in latent space. In *Neural Information Processing Systems*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4886–4899, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *ArXiv*, abs/1901.08149.

Yuqiang Xie, Yue Hu, Wei Peng, Guanqun Bi, and Luxi Xing. 2022. COMMA: Modeling relationship among motivations, emotions and actions in language-based human activities. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 163–177, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. 2022. Diffsound: Discrete diffusion model for text-to-sound generation. *ArXiv*, abs/2207.09983.

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

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 1815–1825.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing

Liu, and Bill Dolan. 2020b. 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, Online. Association for Computational Linguistics.

Chujie Zheng, Yong Liu, Wei Chen, Yongcai Leng, and Minlie Huang. 2021. CoMAE: A multi-factor hierarchical framework for empathetic response generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 813–824, Online. Association for Computational Linguistics.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 1097–1100. ACM.

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.

Noticed 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:

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

where m and n are the numbers of tokens in the context and response respectively. It is implemented with a reverse 345M DialoGPT (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. γ equals to 0.2. We use WordPiece tokenizer[†]. 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.

[†]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
Context	**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.**
Gold	Awww, That sucks man. Hopefully they don't say mean things to you about being lonely.
TRS	I am sorry to hear that. I am sorry to hear that.
Multi-TRS	I am sorry to hear that. I am sorry to hear that.
MoEL	I am sorry to hear that. have you tried talking to them?
MIME	I am sorry to hear that. I am sorry to hear that. I hope you are able to do that.
EmpDG	That is a shame. I am sorry to hear that.
CEM	I am so sorry to hear that. I am sure you will be fine.
GPT2-FT	Why were they so angry?
BART-FT	I'm sorry to hear that. Have you tried joining any clubs?
DiffuSeq	Why can not you too
Candidate A	I see. I'm sorry you're home alone. I get lonely myself when you're used to someone around. are you going to advertise for a new roommate?
Control A	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 _ ?
Response A	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?
Candidate B	Maybe you could invite a friend over for a movie?
Control B	EXPLORATION SUGGESTING _ PRONOUN POSSIBILITY REQUEST _ PERSONAL _ RELATIONSHIP _ _ _ _ ?
Response B	maybe you could call a friend to talk with her?
Candidate C	Oh no really, that really sucks to hear. Why were you not able to go?
Control C	EMOTIONAL_REACTION EXPLORATION ACKNOWLEDGING QUESTIONING _ NO _ _ _ _ _ _ _ PERCEPTION_EXPERIENCE _ WHY _ PRONOUN NO CAPABILITY _ MOTION ?
Response C	Oh no, that is really amazing to hear. Why were you not able to move?

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

Case 2	
Context	**Today traffic was horrible and was so frustrating!**
Gold	I hate traffic too, it makes me angry
TRS	Oh no! What happened?
Multi-TRS	Oh no! I hate when that happens. Did you get hurt?
MoEL	I am sorry to hear that. What happened?
MIME	I am sorry to hear that. What happened?
EmpDG	I hate when that happens.
CEM	What happened?
GPT2-FT	Why were they so angry?
BART-FT	I hate traffic too. I hate it when it's so bad.
DiffuSeq	Was it hurt you ?
Candidate A	I completely understand. I get pretty aggravated myself driving home from work everynight.
Control A	EMOTIONAL_REACTION INTERPRETATION AGREEING AGREEING PRONOUN _ AWARENESS _ PRONOUN _ _ EXPERIENCER_OBJ _ SUBJECTIVE_INFLUENCE BUILDINGS _ WORK _ _
Response A	I completely understand. I have been tired to drive home from work everyday.
Candidate B	Yes! Whats even worse is when other people don't pay attention in bad traffic!
Control B	INTERPRETATION SUGGESTING QUESTIONING YES _ _ _ _ _ _ INCREMENT PEOPLE _ NO COMMERCE_PAY ATTENTION _ DESIRABILITY _ _
Response B	Yes! Traffics is the worst but other people don't pay attention to bad thing.
Candidate C	Yes, the cable company is infuriating. do they eventually help you though?
Control C	EXPLORATION NEUTRAL QUESTIONING YES _ _ _ BUSINESSES _ _ _ INTENTIONALLY_ACT PRONOUN TIME_VECTOR ASSISTANCE PRONOUN CONCESSIVE?
Response C	Yes, the bus company was annoying. Did they already help you out?

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

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