DialoGPS: Dialogue Path Sampling in Continuous Semantic Space for Data Augmentation in Multi-Turn Conversations

Ang Lv[1]*, **Jinpeng Li**[2]*, **Yuhan Chen**[1], **Xing Gao**[3], **Ji Zhang**[3], **Rui Yan**[1,4]†

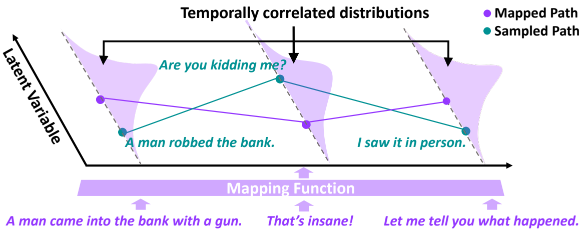
[1]Gaoling School of Artifical Intelligence, Renmin University of China
[2]Wangxuan Institute of Computer Technology, Peking University
[3]Alibaba DAMO Academy
[4]Engineering Research Center of Next-Generation Intelligent
Search and Recommendation, Ministry of Education
`{anglv, yhchen, ruiyan}@ruc.edu.cn, lijinpeng@stu.pku.edu.cn,`
`{gaoxing.gx,zj122146}@alibaba-inc.com`

Abstract

In open-domain dialogue generation tasks, contexts and responses in most datasets are one-to-one mapped, violating an important many-to-many characteristic: a context leads to various responses, and a response answers multiple contexts. Without such patterns, models poorly generalize and prefer responding safely. Many attempts have been made in either multi-turn settings from a one-to-many perspective or in a many-to-many perspective but limited to single-turn settings. The major challenge to many-to-many augment multi-turn dialogues is that discretely replacing each turn with semantic similarity breaks fragile context coherence. In this paper, we propose DialoGue Path Sampling (DialoGPS) method in continuous semantic space, the first many-to-many augmentation method for multi-turn dialogues. Specifically, we map a dialogue to our extended Brownian Bridge, a special Gaussian process. We sample latent variables to form coherent dialogue paths in the continuous space. A dialogue path corresponds to a new multi-turn dialogue and is used as augmented training data. We show the effect of DialoGPS with both automatic and human evaluation.

The original version	
A:	A man came into the bank with a gun.
B:	That's insane!
A:	Let me tell you what happened.
The modified version	
A:	I am also afraid... Have a hope.
B:	Wow! What a great news!!
A:	Ha ha.. I knew mom. Bye bye.

(a) Discrete replacement causes incoherence.



(b) Sampled dialogue paths in the continuous semantic space correspond to coherent discrete dialogues.

Figure 1: (a) When replacing each utterance in the original conversation by semantic similarity, the modified dialogue is incoherent. (b) We map dialogues into a continuous semantic space where latent distributions of utterances correlate with each other, and sample dialogue paths for training. Each path corresponds to a discrete multi-turn conversation.

1 Introduction

Open-domain dialogue generation has received significant attention and has made notable advancements (Zhang et al., 2020b; Shuster et al., 2022; OpenAI, 2022). However, it still faces challenges due to the nature of the data. One specific challenge is the many-to-many relationship between contexts and responses in open-domain conversations. A context can lead to various responses, and a response can be relevant to multiple contexts. Unfortunately, most datasets only provide one-to-one mappings between contexts and responses. This limitation results in models being poorly generalized when they rely on learned one-to-one patterns, making them prone to generating safe yet uninteresting responses (Jiang and de Rijke, 2018; Jiang et al., 2019).

To address this limitation, many attempts (Sai et al., 2020; Qiu et al., 2019; Xie et al., 2022) have been made from a one-to-many perspective which involves constructing multiple responses for a context. Furthermore, some works are proposed from a many-to-many perspective but are limited to single-turn settings. To construct new dialogue sentence pairs, they either replace sentences based on se-

*Equal contribution.

†Corresponding author: Rui Yan (`ruiyan@ruc.edu.cn`).

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 1267–1280
July 9-14, 2023 ©2023 Association for Computational Linguistics

mantic similarity (Zhang et al., 2020a) or sample new sentences from probabilistic models (Li et al., 2019). Next, they adopt BERT (Devlin et al., 2019) or GAN (Goodfellow et al., 2014) discriminators to filter incoherent sentence pairs.

These methods cannot be trivially extended to multi-turn settings. Considering T utterances in a dialogue and K candidates for each utterance, they need to (1) prepare a large sentence set as candidates for replacement or a strong generative model, and (2) check the coherence of the modified conversation at least K^{T-1} times, which is impractical. Figure 1(a) shows a case in which we replace each utterance in a conversation following Zhang et al. (2020a). The modified conversation is still incoherent across turns. Therefore, to enhance multi-turn dialogue generation from a many-to-many perspective, we resort to a continuous semantic space that satisfies two requirements. First, it describes semantic distributions of utterances, allowing for sampling semantic neighbors of each utterance. Second, latent variables sampled from any two distributions should be temporally correlated, contributing to a new coherent dialogue path in the latent space without requiring post-checks. This path can be utilized as a new training sample to augment the model. Our motivation is illustrated in Figure 1(b).

Driven by this motivation, we propose a novel method for augmenting open-domain dialogues from a many-to-many perspective, called DialoGue **P**ath **S**ampling (DialoGPS), aiming to enhance generalization and improve the quality of generated responses. Specifically, our approach involves the following steps: (1) We map each utterance in a multi-turn dialogue to a special Gaussian process in a continuous semantic space known as the Brownian Bridge (Revuz and Yor, 2013). (2) For each utterance x_i, we sample K latent variables z_i^j, $j \in [1, K]$, establishing K different dialogue paths in the bridge. Each path corresponds to a new multi-turn conversation in the discrete space. (3) DialoGPS utilizes an encoder-decoder architecture. To construct augmented data, we mix the latent variable z_i with representations of x_i in the encoder if x_i is part of the context, and in the decoder if it is the response. (4) Finally, we train the model using the augmented data.

To ensure the effectiveness of DialoGPS, we address several key issues. First, traditional Brownian Bridges have deterministic endpoints, which prevent response sampling and lead our method degenerating into a many-to-one paradigm, further impairing generalization. To overcome this limitation, we derive the formula of endpoint distributions. Second, since augmented data that lacks discrete utterance labels makes the optimization challenging, we propose a self-distillation framework where the model first learns from the ground truth and then distills its knowledge to guide itself in utilizing augmented data.

We evaluate DialoGPS on two multi-turn open-domain datasets. Both automatic and human evaluation show that DialoGPS performs better than strong baselines and even outperforms the model trained on manually denoted multi-reference data, which demonstrates the benefit of the many-to-many augmentation paradigm. Because DialoGPS is plug-and-play, we add it to BART (Lewis et al., 2020) and achieve competitive results with the state-of-the-art model, DialoFlow (Li et al., 2021). Our contributions are as follows:

• DialoGPS is the first work to augment multi-turn dialogues from a many-to-many perspective.

• To ensure the effectiveness of DialoGPS, we have introduced dialogue-specific designs, including endpoint sampling of Brownian Bridges and self-distillation for model optimization.

• Experiments conducted on both non-pretrained and pre-trained models show that our DialoGPS method outperforms all baselines.

2 Related Work: Dialogue Generation Augmentation

In general, dialogue generation can be categorized into two groups: task-oriented and open-domain. Open-domain generation is a context-aware process that lasts for turns. The model learns to generate a proper but open response from the preceding utterances (i.e., contexts). Task-oriented dialogues progress for specific purposes and are limited to specific domains, such as obtaining knowledge (Zhao et al., 2020; Tao et al., 2021). However, due to the specific domains in task-oriented dialogues, the many-to-many relationship is not as apparent compared to open-domain dialogues.

In this paper, we focus on open-domain dialogue generation augmentation from an X-to-many perspective. From a one-to-many perspective, Sai et al. (2020) manually denoted multiple responses for a dialogue context. Based on such multi-reference datasets, Qiu et al. (2019) proposed to capture the common feature in feasible responses and then add the specific feature to obtain the final output, which augments the utility of the data and improves the generalization. Xie et al. (2022) proposed that with only one-to-one data, models can construct pseudo-target data in the decoder and improve the model by bootstrapping. From a many-to-many perspective, existing methods work in single-turn settings. Li et al. (2019) generated multiple context or responses with CVAE (Zhao et al., 2017) and introduced a GAN (Goodfellow et al., 2014) discriminator to filter incoherent sentence pairs. Zhang et al. (2020a) augmented a one-to-one dialogue dataset D_p with an unpaired sentence set D_u. They sample sentences from D_u and replace the most similar sentences in D_p. They use BERT (Devlin et al., 2019) and knowledge distillation to filter noise in incoherent sentence pairs. Until now, many-to-many augmentation in multi-turn settings are understudied.

3 Method

We first present some preliminaries (§ 3.1). Then, we introduce mapping dialogue texts to the desired latent space (§ 3.2), augmented data construction (§ 3.3), augmented data utilization (§ 3.4), and inference details (§ 3.5). Figure 2 shows the overview of DialoGPS.

3.1 Preliminary

In open-domain dialogue generation, given a multi-turn dialogue $X = [x_0, x_1, ..., x_T]$, the goal is to predict the response x_T based on the context $X_{0:T-1}$. The number of tokens in x_t is denoted as $|x_t|$, $t \in \{0, 1, \ldots, T\}$. The i-th token in the x_t is denoted as x_t^i. A Brownian Bridge $\mathcal{B}$ defined on time range $[0, T]$ is a special Gaussian process established on deterministic endpoints μ_0 and μ_T. At time t, the latent variable z_t follows a Gaussian distribution $\mathcal{B}(t|\mu_0, \mu_T)$:

$$z_t \sim \mathcal{B}(t|\mu_0, \mu_T) = \mathcal{N}(\mu_0 + \frac{t}{T}(\mu_T - \mu_0), \frac{t(T-t)}{T}), \quad (1)$$

3.2 Extended Brownian Bridge

In DialoGPS, given X, a non-linear function f_θ maps each x_t to μ_t, the expectations of the corresponding semantic distribution. Based on μ_0 and μ_T, we can establish a Brownian Bridge, and from which we sample the latent variable z_t as the semantic neighbor of x_t. Meanwhile, $z_0, z_1, ..., z_T$ compose a coherent dialogue path because in a Brownian Bridge, the covariance between t_1 and t_2, with $0 < t_1 < t_2 < T$ is $\frac{t_1(T-t_2)}{T}$, where the constant positive covariance guarantees that $\mathcal{B}(t_1|\mu_0, \mu_T)$ and $\mathcal{B}(t_2|\mu_0, \mu_T)$ are temporally correlated.

However, as defined in Eq. 1, a conventional Brownian Bridge $\mathcal{B}$ has deterministic endpoints, which prevents us from sampling for x_T, the response, and x_0, the first utterance in the context. To avoid degenerating to a many-to-one mode that impairs the generalization, we derive an extended Brownian Bridge β with samplable endpoints. Take the derivation of $\beta(T|\mu_0, \mu_T)$ as example: given a $\mathcal{B}$, both the distance d_δ between μ_T and $z_{T-\delta}$ and the summation of d_δ and $z_{T-\delta}$ follow the Gaussian distribution, we can derive the distribution of z_T as follows:

$$\left.\begin{aligned} z_{T-\delta} &\sim \mathcal{N}(\frac{T-\delta}{T}\mu_T + \frac{\delta}{T}\mu_0, \frac{\delta(T-\delta)}{T}) \\ d_\delta = \mu_T - z_{T-\delta} &\sim \mathcal{N}(\frac{\delta}{T}\mu_T - \frac{\delta}{T}\mu_0, \frac{\delta(T-\delta)}{T}) \end{aligned}\right\} \Rightarrow$$
$$z_T = d_\delta + z_{T-\delta} \sim \mathcal{N}(\mu_T, \frac{2\delta(T-\delta)}{T}). \quad (2)$$

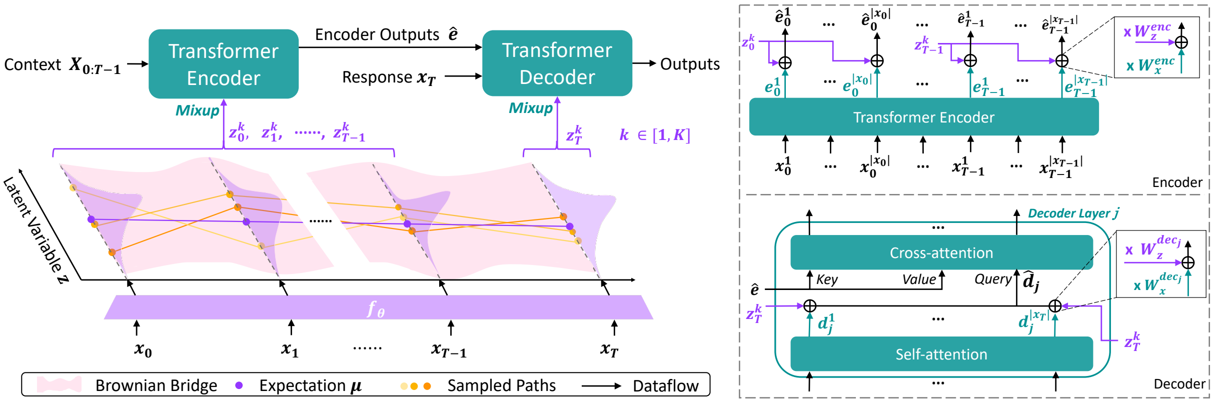
Due to the symmetry, z_0 follows $\mathcal{N}(\mu_0, \frac{2\delta(T-\delta)}{T})$. Here, δ serves as a hyper-parameter. To sum up, we define the extended Brownian Bridge β as:

$$\beta(t|\mu_0, \mu_T) = \begin{cases} \mathcal{N}(\mu_t, \frac{2\delta(T-\delta)}{T}), \text{ t = 0 or T}, \\ \mathcal{N}(\mu_0 + \frac{t}{T}(\mu_T - \mu_0), \frac{t(T-t)}{T}), \text{ otherwise.} \end{cases} \quad (3)$$

To optimize the mapping function f_θ, we follow (Wang et al., 2022) to adopt a contrastive learning framework where positive samples are ordered sentence triplets from the same conversation (x_{t_0}, x_{t_1}, x_{t_2}, $t_0 < t_1 < t_2$) and negative samples are constructed by randomly replacing the middle point x_{t_1} with other sentences $x_{t_1'}$ from the mini-batch $\mathbb{B}$. The objective is as below:

$$\mathcal{L}_\beta = \mathbb{E}_X \left[\log \left(1 + \frac{\sum_{(x_{t_0}, x_{t_1'}, x_{t_2}) \in \mathbb{B}} \exp(d(x_{t_0}, x_{t_1'}, x_{t_2}; f_\theta))}{\exp(d(x_{t_0}, x_{t_1}, x_{t_2}; f_\theta))} \right) \right], \quad (4)$$

where $d(x_{t_0}, x_{t_1}, x_{t_2}; f_\theta) = -\frac{1}{2\sigma_{t_1}^2} \| f_\theta(x_{t_1}) - (1 - \frac{t_1}{t_2}) f_\theta(x_{t_0}) - \frac{t_1}{t_2} f_\theta(x_{t_2}) \|_2^2$. The essence of Eq. 4 is to optimize the outputs of f_θ, i.e., μ_{t_0}, μ_{t_1}, and μ_{t_2} to the linear relationship as defined in Eq. 1. In DialoGPS, a 4-layer MLP serves as f_θ. To embed utterance as inputs of f_θ, there are many choices such as averaging token embeddings or encoding



(a) Method overview.

(b) Mixup details on encoder and decoder.

Figure 2: (a) The overview of DialoGPS. Teacher forcing is applied during training. Each utterance in the dialogue is mapped into a semantic distribution on a Brownian Bridge. We sample K paths and conduct mixup operations in the encoder and decoder, respectively. (b) Mixup details.

by a language model. We leave the embedding details in §5.3.

3.3 Augmented Data Construction

As shown in Figure 2(a), we take Transformer (Vaswani et al., 2017) as the bone architecture. With f_θ, an extended Brownian Bridge β is established. We sample latent variables $z_t \sim \beta(t|\mu_0, \mu_T)$ and mix them with representations of corresponding x_t. In the encoder, for each utterance x_t in the context $X_{0:T-1}$, we conduct:

$$\begin{aligned} e_t^1, e_t^2, ...e_t^{|x_t|} &= \text{Encoder}(x_t), \\ \hat{e}_t^i &= W_x^{enc} \cdot e_t^i + W_z^{enc} \cdot z_t, \end{aligned} \tag{5}$$

where e_t^i is the output corresponding to the i-th token in x_t from the encoder, $i \in [1, |x_t|]$. W_z^{enc} and W_x^{enc} are trainable vectors of the same dimension as e and z. Finally, $\hat{e}$ is sent to the decoder for cross-attention. We conduct the mixup every decoder layer:

$$\begin{aligned} &\hat{d}_j^i = W_x^{dec_j} \cdot d_j^i + W_z^{dec_j} \cdot z_T, \\ &i \in [1, |x_T|], j \in [1, N], \end{aligned} \tag{6}$$

where N is the number of decoder layers, d_j^i is the self-attention output at position i in layer j. Also, $W_z^{dec_j}$ and $W_x^{dec_j}$ are trainable vectors. $\hat{d}_j$ is used as *Query*, and $\hat{e}$ are used as both *Key* and *Value* in the cross-attention. For a dialogue text X, we conduct sampling and mixup K times, which is equivalent to providing K extra discrete dialogues $\hat{X}^k = [\hat{x}_0^k, \hat{x}_1^k, ..., \hat{x}_T^k]$, $k \in [1, K]$ for training. Figure 2(b) shows mixup details.

3.4 Utilizing Augmented Data by Self-Distillation

In general, given X to a dialogue generation model, parameters ϕ of model are optimized by minimizing the negative log-likelihood:

$$\phi = \text{argmin}\left(\mathbb{E}_X \left[-\log(P_\phi(x_T|X_{0:T-1]}))\right]\right). \tag{7}$$

However, as aforementioned, what we obtain are continuous representations of $\hat{X}$ whereas the corresponding discrete sentences are inaccessible, which makes Eq. 7 intractable. Hence, to utilize the augmented data, we make an assumption that: There is an inaccessible many-to-many dialogue dataset D_{MtoM}. P_{MtoM} describes the conditional distribution of responses given contexts in this dataset. The accessible one-to-one dataset D_{1to1} is collected by sampling from D_{MtoM} uniformly, and thus P_{1to1} can be viewed as an approximation of P_{MtoM}.

Based on this assumption, we propose a self-distillation framework consisting of two steps: (1) It optimizes the model with the original discrete data following Eq. 7. (2) During training, as P_ϕ fits P_{1to1}, which is an approximation of P_{MtoM}, the model can use its output given X to teach itself when presented with augmented data, i.e., the representations of $\hat{X}$:

$$\phi = \text{argmin}\left(D_{KL}\left[P_\phi(x_T|X_{0:T-1})||P_\phi(\hat{x}_T|\hat{X}_{0:T-1})\right]\right), \tag{8}$$

where $D_{KL}[\cdot||\cdot]$ is the KL-divergence (Kullback and Leibler, 1951). In Eq. 8, to remove the gap between utilizing the original discrete data X and the augmented continuous data $\hat{X}$ in the same architecture, we mix each utterance in X with the expectations $\mu_{0:T}$. Formally, the overall training objective is to minimize:

$$\mathcal{L} = \underbrace{\mathcal{L}_\beta}_{\text{Mapping } X \text{ to } \beta} + \underbrace{\mathbb{E}_X\left[-\log(P_\phi(x_T|X_{0:T-1}, \mu_{0:T}))\right]}_{\text{Utilizing original discrete data}} + \underbrace{\frac{1}{K}\sum_{k}^{K} D_{KL}\left[P_\phi(x_T|X_{0:T-1}, \mu_{0:T})||P_\phi(\hat{x}_T^k|\hat{X}_{0:T-1}^k, z_{0:T}^k)\right]}_{\text{Utilizing augmented data}} \tag{9}$$

3.5 Inference

The inference goal is to predict x_T based on context $X_{0:T-1}$. First, f_θ takes $X_{0:T-1}$ and outputs corresponding μ_t for sampling and mixup in the encoder, where $t \in \{0, 1, \ldots, T-1\}$. Next, the decoder receives the encoder output and an inferred μ_T to decode the response in an autoregressive manner. To obtain the value of μ_T, we do not require additional prediction networks. Instead, we can directly derive its value based on the property of Brownian Bridge. Specifically, given the context, we know that for any t:

$$\mu_t = \mu_0 + \frac{t}{T-1}(\mu_{T-1} - \mu_0). \tag{10}$$

If μ_T is already known, a Brownian bridge established on μ_T and μ_0 would yield the same μ_t values. Consequently, we can establish an equality and derive the value of μ_T as follows:

$$\begin{aligned} &\mu_t = \mu_0 + \frac{t}{T}(\mu_T - \mu_0) = \mu_0 + \frac{t}{T-1}(\mu_{T-1} - \mu_0) \\ &\Rightarrow \mu_T = \frac{T}{T-1}\mu_{T-1} - \frac{1}{T-1}\mu_0. \end{aligned} \tag{11}$$

We find that there is hardly a difference in evaluation results when conducting mixup operations with either expectations μ or sampled variables z. To reduce randomness for easier analyses, experiments in below use expectations μ to mixup. Nonetheless, sampling variables gives DialoGPS the ability to generate diverse responses to an arbitrary context and we will discuss it in § 5.4.

4 Experimental Settings

Datasets We conduct multi-turn dialogue generation experiments on two public datasets: DailyDialog (Li et al., 2017) and PersonaChat (Zhang et al., 2018a). DailyDialog contains high-quality multi-turn dialogues collected from daily conversations, and it has many multi-reference versions (Sai et al., 2020; Gupta et al., 2019) denoted by humans, which makes it possible for us to compare DialoGPS with human annotators. Besides, it is more reliable to evaluate the generalization and performance with multiple references. PersonaChat collects dialogues based on chatters' profiles. Profiles are not shown to models, so it is more challenging and open to generate proper responses, measuring generalization capacity better.

Baselines and Parameters We compare DialoGPS with (1) Transformer (Vaswani et al., 2017). (2)DD++ (Sai et al., 2020): it is a variant of DailyDialog in which each context has five manually denoted responses. We train a vanilla Transformer on it. (3) TSA (Xie et al., 2022): it is an unsupervised augmentation method in the decoder side. It uses its decoder's output to construct pseudo-target data which is used to train the model for another round. From a dialogue generation viewpoint, it is a one-to-many method that bootstraps based on one-to-one data. (4) M&D-D (Zhang et al., 2020a): it uses a pre-trained model and BM-25 algorithm to construct new context-response pairs from unpaired sentences. Since it is a single-turn augmentation, given a multi-turn dialogue, we only apply this method to the last two turns. (5) ResBag (Qiu et al., 2019): an augmented VAE-based model. It captures the common feature in the bag of plausible responses and then adds the specific feature to obtain the final output, which utilizes the multiple references better.

Because DialoGPS is a plug-and-play method, we add it to a BART$_{\text{Large}}$ (Lewis et al., 2020) and compare with DialoFlow$_{\text{Large}}$ (Li et al., 2021). DialoFlow is one of the state-of-the-art pre-trained models in open-domain dialogue generation. It augments the model by modeling the dialogue flow. More details on the implementation and hyper-parameters are in Appendix A.1.

Evaluation Metrics We consider three automatic evaluation metrics: BLEU (Papineni et al., 2002), Distinct (DIST) (Li et al., 2016), and BLEURT (Sellam et al., 2020). BLEU measures the word overlap between generated responses and the ground truth. DIST measures the ratio of unique n-grams in the generated responses. Because these two metrics are only sensitive to lexical variation, we evaluate BLEURT, an advanced learned semantic-sensitive evaluation metric based on BERT (Devlin et al., 2019). On the evaluation of fine-tuning pre-trained models, we follow (Li et al., 2021) to report METEOR (Lavie and Agarwal, 2007) and

Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	DIST-1	DIST-2.	BLEURT
			PersonaChat Dataset				
Transformer	$17.79_{[0.14]}$	$6.93_{[0.06]}$	$3.03_{[0.08]}$	$1.41_{[0.06]}$	$0.82_{[0.01]}$	$6.60_{[0.05]}$	$30.16_{[0.05]}$
ResBag	$17.82_{[0.17]}$	$6.88_{[0.12]}$	$3.04_{[0.09]}$	$1.37_{[0.11]}$	$0.85_{[0.02]}$	$6.83_{[0.02]}$	$30.25_{[0.17]}$
TSA	$17.76_{[0.19]}$	$6.92_{[0.16]}$	$2.97_{[0.15]}$	$1.35_{[0.10]}$	$0.85_{[0.02]}$	$6.56_{[0.01]}$	$30.66_{[0.09]}$
M&D-D	$18.42_{[0.13]}$	$7.25_{[0.09]}$	$3.23_{[0.11]}$	$1.44_{[0.07]}$	$0.80_{[0.01]}$	$6.55_{[0.01]}$	$30.46_{[0.13]}$
DialoGPS$_{K=1}$	$18.29_{[0.08]}$	$7.21_{[0.05]}$	$3.14_{[0.03]}$	$1.44_{[0.05]}$	$\mathbf{1.05}_{[0.01]}$	$\mathbf{7.97}_{[0.07]}$	$30.54_{[0.06]}$
DialoGPS$_{K=2}$	$18.96_{[0.15]}$	$7.61_{[0.09]}$	$3.32_{[0.04]}$	$1.54_{[0.02]}$	$0.84_{[0.00]}$	$7.10_{[0.04]}$	$\mathbf{30.77}_{[0.14]}$
DialoGPS$_{K=4}$	$\mathbf{19.05}_{[0.18]}$	$\mathbf{7.70}_{[0.16]}$	$\mathbf{3.41}_{[0.09]}$	$\mathbf{1.61}_{[0.07]}$	$0.91_{[0.01]}$	$7.45_{[0.09]}$	$30.29_{[0.12]}$
DialoGPS$_{K=8}$	$19.04_{[0.08]}$	$7.64_{[0.11]}$	$3.40_{[0.10]}$	$1.60_{[0.08]}$	$0.93_{[0.01]}$	$7.64_{[0.06]}$	$30.39_{[0.14]}$
			Multi-reference DailyDialog Dataset				
Transformer	$33.93_{[0.26]}$	$12.32_{[0.25]}$	$4.93_{[0.23]}$	$2.14_{[0.14]}$	$2.59_{[0.03]}$	$20.62_{[0.12]}$	$35.79_{[0.15]}$
ResBag	$34.10_{[0.27]}$	$12.61_{[0.18]}$	$4.82_{[0.17]}$	$2.13_{[0.13]}$	$2.98_{[0.06]}$	$24.44_{[0.17]}$	$35.22_{[0.15]}$
TSA	$36.14_{[0.11]}$	$13.21_{[0.15]}$	$5.43_{[0.14]}$	$2.46_{[0.13]}$	$3.56_{[0.04]}$	$26.89_{[0.21]}$	$35.37_{[0.13]}$
DD++	$36.87_{[0.32]}$	$14.09_{[0.24]}$	$6.13_{[0.23]}$	$2.91_{[0.17]}$	$3.84_{[0.03]}$	$28.58_{[0.38]}$	$\underline{37.04}_{[0.14]}$
M&D-D	$36.97_{[0.12]}$	$14.28_{[0.09]}$	$6.50_{[0.19]}$	$3.28_{[0.17]}$	$3.65_{[0.03]}$	$25.35_{[0.21]}$	$36.02_{[0.15]}$
DialoGPS$_{K=1}$	$37.21_{[0.12]}$	$14.72_{[0.14]}$	$6.65_{[0.12]}$	$3.29_{[0.11]}$	$4.25_{[0.05]}$	$28.39_{[0.14]}$	$36.14_{[0.08]}$
DialoGPS$_{K=2}$	$38.01_{[0.13]}$	$14.79_{[0.07]}$	$6.52_{[0.06]}$	$3.20_{[0.04]}$	$4.34_{[0.06]}$	$29.04_{[0.25]}$	$36.15_{[0.16]}$
DialoGPS$_{K=4}$	$38.27_{[0.20]}$	$14.77_{[0.13]}$	$6.62_{[0.15]}$	$\mathbf{3.33}_{[0.20]}$	$\mathbf{4.53}_{[0.07]}$	$\mathbf{30.18}_{[0.17]}$	$36.09_{[0.08]}$
DialoGPS$_{K=8}$	$\mathbf{38.46}_{[0.18]}$	$\mathbf{15.05}_{[0.23]}$	$\mathbf{6.70}_{[0.24]}$	$3.30_{[0.14]}$	$4.32_{[0.06]}$	$28.35_{[0.14]}$	$35.82_{[0.16]}$
DialoGPS$_{K=16}$	$38.38_{[0.14]}$	$14.89_{[0.06]}$	$6.62_{[0.13]}$	$3.30_{[0.15]}$	$4.41_{[0.05]}$	$29.84_{[0.08]}$	$35.81_{[0.05]}$
			Component Ablation on Multi-reference DailyDialog (K=4)				
–M.E.	$38.04_{[0.17]}$	$15.00_{[0.12]}$	$6.63_{[0.12]}$	$3.21_{[0.11]}$	$4.22_{[0.03]}$	$28.05_{[0.10]}$	$35.96_{[0.09]}$
–M.D.	$34.62_{[0.12]}$	$12.71_{[0.13]}$	$5.20_{[0.08]}$	$2.33_{[0.08]}$	$3.19_{[0.04]}$	$24.65_{[0.16]}$	$35.14_{[0.13]}$
–Brown.	$38.05_{[0.22]}$	$14.68_{[0.05]}$	$6.36_{[0.04]}$	$3.01_{[0.10]}$	$4.05_{[0.09]}$	$27.58_{[0.18]}$	$35.52_{[0.11]}$
–M.E. –Brown.	$38.42_{[0.13]}$	$14.76_{[0.15]}$	$6.55_{[0.05]}$	$3.17_{[0.12]}$	$4.11_{[0.03]}$	$27.64_{[0.16]}$	$36.12_{[0.12]}$
–M.D. –Brown.	$34.49_{[0.31]}$	$12.68_{[0.28]}$	$5.15_{[0.23]}$	$2.29_{[0.17]}$	$2.97_{[0.45]}$	$24.46_{[0.15]}$	$35.11_{[0.12]}$
–M.E. –M.D.	$33.93_{[0.26]}$	$12.32_{[0.25]}$	$4.93_{[0.23]}$	$2.14_{[0.14]}$	$2.59_{[0.03]}$	$20.62_{[0.12]}$	$35.79_{[0.15]}$

Table 1: Automatic evaluation and ablation results on multi-reference DailyDialog and PersonaChat. We apply Top-5 Sampling decoding scheme. The standard deviation $[\sigma]$ (across 5 runs) is also reported. In the ablation results table, M.E/D. stands for applying mixup in the encoder/decoder, and Brown. stands for optimizing f_θ with Eq. 4. When there is no mixup in either encoder or decoder, the model degenerates into a vanilla transformer.

Entropy (Zhang et al., 2018b). For human evaluation, we recruit five evaluators to manually judge 200 samples from each experiment in blind testing, where we set three metrics to comprehensively evaluate the generation quality: whether a response is *readable* (**Read.**), *coherent* (**Coh.**), and *informative* (**Info.**). For each aspect, evaluators can score at 'bad', 'borderline' and 'good'.

5 Results

Table 1 shows the automatic evaluation results. On PersonaChat, without access to chatters' profiles, conversations are so open that there is so much noise in data for models to learn. Therefore, models prefer safe responses and thus DISTs are relatively low. However, DialoGPS still improves by about 20% in DISTs than the best-performing baseline. Also, BLEU and BLEURT scores imply that DialoGPS matches references more lexically and more semantically. On the multi-reference DailyDialog dataset, DialoGPS gains improvement by a large margin than other strong baselines. Also, most baselines suffer a trade-off between matching the references and diversifying responses. By contrast, DialoGPS performs evenly well on all metrics. DialoGPS also wins 6 out of all 7 metrics compared with the model trained on DD++, the human-written multi-reference training set. Our

Models	DailyDialog				PersonaChat			
	BLEU-2	**BLEU-4**	**METEOR**	**Entropy**	**BLEU-2**	**BLEU-4**	**METEOR**	**Entropy**
BART	27.87	10.85	14.69	9.29	9.95	3.38	8.69	6.55
DialoFlow	28.02	11.57	**16.40**	9.46	10.46	3.03	**9.32**	**6.89**
BART + DialoGPS	**29.18**	**12.05**	15.30	**9.73**	**10.97**	**4.08**	9.26	6.70

Table 2: Automatic evaluation results on fine-tuning pre-trained models (beam search with width 5).

Models	DailyDialog			PersonaChat		
	Read.	**Coh.**	**Info.**	**Read.**	**Coh.**	**Info.**
Transformer	70/8	69/9	73/12	53/14	51/11	52/9
ResBag	58/13	60/11	64/14	51/14	50/19	51/16
TSA	59/15	57/16	60/16	48/20	47/22	43/20
DD++	53/24	55/20	51/17	-	-	-
M&D-D	56/19	47/20	52/16	44/21	46/18	45/17
BART	40/34	42/23	44/26	39/31	41/26	34/20
DialoFlow	36/32	40/29	43/27	39/34	35/28	35/25

Table 3: Human evaluation results (rounded). Compared with each baseline, we report our win/lose percentage. Evaluators achieve substantial agreement with kappa value 0.62 on experiments trained from scratch and 0.70 on pre-trained experiments.

results in bold pass the significance test p < 0.01. In Table 2, when adding DialoGPS$_{K=2}$ to a pre-trained BART and fine-tuning on two datasets, it achieves competitive performance as one of the SOTA dialogue generation pre-trained models, DialoFlow. DialoFlow augments the generation with the help of 'flow', i.e., the difference of adjacent utterances in continuous space. Their flows are not as flexible as paths sampled from the Brownian Bridge, which is one of the reasons that DialoGPS outperforms DialoFlow in five out of all eight metrics. Table 3 shows human evaluation results. In three metrics, DialoGPS achieves the top rank with solid agreement among evaluators. More evaluation details are in Appendix A.2.

5.1 Study on Dialogue Paths

We conduct an ablation study on the number of sampled dialogue paths K, results are shown in Table 1. On both datasets, with the increase of K, various metrics increase and then reach the bottleneck or slightly decrease. This phenomenon mainly dues to that different from discrete data, sampled paths in continuous space have a information bottleneck, i.e., if K is big enough to cover the most samplable area in the Brownian Bridge, then increasing K further may cause little improvement or even decrease due to more noise. We visualize the sampled paths of a conversation with 5 utterances during training in Figure 3. A sample at each time step is denoted as a point and paths are depicted. We can see that the Brownian Bridge area covered by paths is significantly increased when K increases from 1 to 8, but there is a slight difference when K further increases to 16. The visualization confirms automatic evaluation results in Table 1.

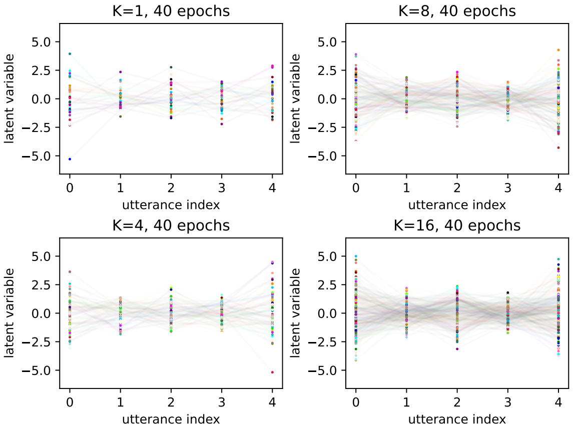


Figure 3: The visualization of sampled dialogue paths (normalized expectations) for a 5-utterance dialogue, training with varying K.

5.2 Component Ablation

We study the effect on the performance of the following components in DialoGPS: mixup in the encoder (M.E.), mixup in the decoder (M.D.), and constraints from Eq. 4 that is the optimization of the mapping function (Brown.). The results are reported at the bottom of Table 1. Removing mixup in the decoder (–M.D.) degenerates DialoGPS to a many-to-one mode and thus the performance degrades much, confirming the intuition mentioned in §1. Removing mixup in the encoder(–M.E.) degenerates DialoGPS to a one-to-many pattern which is insufficient compared with the many-to-many pattern, and DIST drops while the BLEU maintains. Nonetheless, the performance is still

Method	BLEU-2	BLEU-4	DIST-1	DIST-2
Avg.	14.77	3.33	4.53	30.18
Avg. + Pos.	14.41	2.89	4.19	29.22
GPT-2	15.13	3.28	4.23	29.55

Table 4: Experimental results with different utterance representation methods (K=4).

competitive with the best one-to-many baseline. Without constraints from Eq. 4 (–Brown.), there is no context-wise correlation among sampled latent variables and the mixup turns to introduce noise. This variant resembles sampling each utterance with a VAE (Bowman et al., 2016; Miao et al., 2016). However, Eq. 11 does not hold anymore so there exist gaps between the inference and the training, and results drop compared to the variant with Eq. 4. Overall, this variant still plays a positive role because adding noise during training is proved to be effective in improving the robustness and generalization of the model (Srivastava et al., 2014; Gao et al., 2021). When there is neither M.D. nor M.E., the method becomes a vanilla transformer.

$X_{0:2}$ x_3	A: Excuse me, sir. Is there a barber near here? B: Yes, the nearest one is at the third cross of this road. A: I'm a stranger here. How can I get there, please? B: ______________________
Transformer	Thank you very much.
ResBag	Two stops at the next door.
TSA	Let me see. It's about ten minutes.
DD++	Sure.
M&D-D	You can take the subway to get there.
DialoGPS	You have to go to the next stop. (×2) You get off at the next stop. (×2) You have to change. (×2) You have to go to the hotel. (×1) It's not easy. You have to go. (×1) You have to go to the airport. (×1) Then, you have to go to the hotel. (×1)

Table 5: 10 outputs given by DialoGPS when adopting sampling then mixup during inference. To avoid the randomness introduced by the decoding strategy, responses are decoded by Beam Search with width 5.

5.3 Study on Utterance Representation

In §3.3, we defer details on obtaining utterance representations of each turn in a dialogue. We study three variants of encoding an utterance: (1) average embeddings of each token in an utterance (Avg.), (2) average embeddings of each token in an utterance along with position embeddings (Avg. + Pos.), and (3) encode utterances by a GPT-2 (Radford et al., 2019). We conduct this study on the multi-reference DailyDialog dataset and the results are in Table 4. The simplest method (Avg.) achieves first place. With extra positional information, the performance drops a little, and in this experiment, we observed that the $\mathcal{L}_\beta$ term in the overall training objective Eq. 9 maintains steadily, but other terms increase a little. An explanation is that features to be mixed with latent variables (e and d) have included positional information and positional information in latent variables introduces redundancy. For (GPT-2), we add a special token '<eou>' at the end of an utterance and view its corresponding output as the utterance representation. (GPT-2) costs much more training time and only beat (Avg.) in one metric. We guess there is an expression capacity gap so we try to (1) train a 4-layer language model to replace the GPT-2 and (2) apply GPT-2 in pre-trained experiments. In both experiments, we do not observe improvement than (Avg.). To sum up, the simplest (Avg.) achieves the best trade-off between performance and costs so in DialoGPS, we adopt this scheme by default.

5.4 What Does the Model Learn from Augmented Data?

If we mixup with sampled variables instead of expectations during inference, the model obtains the ability to generate diverse responses. Although we do not know what discrete labels augmented data have, to some extent the diverse outputs during inference reflect semantics that augmented data have during training. We provide a case in Table 5. Transformer and ResBag generates incoherent responses, and TSA answers the arrival time but not the way. DD++ reply to the context but does not leads to the follow-up dialogue. M&D-D responds properly but can only provide one answer. We let DialoGPS generate 10 times and report all the outputs along with their respective frequency.

The frequency, the semantics, and lexical features of responses resemble a Gaussian distribution. In this case, 'you have to go to (get off at) the next stop' is close to the expectation. As the semantics get farther away, the frequency of other responses are lower. Overall, DialoGPS provides diverse choices to arrive at the barber. This case shows that continuous augmented data do have open dialogue knowledge which is conducive to model generalization.

6 Conclusion

We propose DialoGPS that first augments open-domain and multi-turn dialogue generation from a many-to-many perspective. Specifically, We map dialogues into the continuous semantic space which is modeled by our extended Brownian Bridge and sample dialogue paths to augment training. We propose a self-distillation framework to utilize augmented data despite the inaccessible discrete labels. Empirically, we prove the effect of DialoGPS and study its characteristics. DialoGPS could be a general method that suits seq2seq tasks where the source has multiple sentences and the target is different from the source in semantics, like summarization. However, DialoGPS should be modified according to the unique properties of the task, which is left to study in the future.

Limitations

Similar to other augmentation methods, DialoGPS demands high requirements for computing resources. The training is performed on up to 8 V100 GPUs. On DailyDialog: a vanilla transformer only needs 50 minutes while a non-pretrained DialoGPS takes about 80 minutes when $K = 1$. Other baselines take about the same amount of time as DialoGPS $K = 1$. But when DialoGPS achieves its performance peak ($K = 16$), the training takes 4 hours. Most of time cost comes from sampling which is difficult to be accelerated by GPUs.

Acknowledgement

This work was supported by National Natural Science Foundation of China (NSFC Grant No. 62122089), Beijing Outstanding Young Scientist Program NO. BJJWZYJH012019100020098, and Intelligent Social Governance Platform, Major Innovation & Planning Inter-disciplinary Platform for the "Double-First Class" Initiative, Renmin University of China.

References

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany. Association for Computational Linguistics.

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.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskenazi, and Jeffrey Bigham. 2019. Investigating evaluation of open-domain dialogue systems with human generated multiple references. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 379–391, Stockholm, Sweden. Association for Computational Linguistics.

Shaojie Jiang and Maarten de Rijke. 2018. Why are sequence-to-sequence models so dull? understanding the low-diversity problem of chatbots. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, pages 81–86, Brussels, Belgium. Association for Computational Linguistics.

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 '19, page 2879–2885, New York, NY, USA. Association for Computing Machinery.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech 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.

Juntao Li, Lisong Qiu, Bo Tang, Dongmin Chen, Dongyan Zhao, and Rui Yan. 2019. Insufficient data can also rock! learning to converse using smaller data with augmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6698–6705.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021. Conversations are not flat: Modeling the dynamic information flow across dialogue utterances. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 128–138, Online. Association for Computational Linguistics.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, Barcelona, Spain.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *International conference on machine learning*, pages 1727–1736. PMLR.

OpenAI. 2022. Chatgpt. `https://openai.com/blog/chatgpt`.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

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

Lisong Qiu, Juntao Li, Wei Bi, Dongyan Zhao, and Rui Yan. 2019. Are training samples correlated? learning to generate dialogue responses with multiple references. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3826–3835, Florence, Italy. Association for Computational Linguistics.

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

D. Revuz and M. Yor. 2013. *Continuous Martingales and Brownian Motion*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg.

Ananya B. Sai, Akash Kumar Mohankumar, Siddhartha Arora, and Mitesh M. Khapra. 2020. Improving dialog evaluation with a multi-reference adversarial dataset and large scale pretraining. *Transactions of the Association for Computational Linguistics*, 8:810–827.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Chongyang Tao, Changyu Chen, Jiazhan Feng, Ji-Rong Wen, and Rui Yan. 2021. A pre-training strategy for zero-resource response selection in knowledge-grounded conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4446–4457.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Rose E Wang, Esin Durmus, Noah Goodman, and Tatsunori Hashimoto. 2022. Language modeling via stochastic processes. In *International Conference on Learning Representations*.

Shufang Xie, Ang Lv, Yingce Xia, Lijun Wu, Tao Qin, Tie-Yan Liu, and Rui Yan. 2022. Target-side input augmentation for sequence to sequence generation. In *International Conference on Learning Representations*.

Zheng Ye, Liucun Lu, Lishan Huang, Liang Lin, and Xiaodan Liang. 2021. Towards quantifiable dialogue coherence evaluation. *CoRR*, abs/2106.00507.

Rongsheng Zhang, Yinhe Zheng, Jianzhi Shao, Xiaoxi Mao, Yadong Xi, and Minlie Huang. 2020a. Dialogue distillation: Open-domain dialogue augmentation using unpaired data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3449–3460, Online. Association for Computational Linguistics.

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

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018b. Generating informative and diverse conversational responses via adversarial information maximization. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

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 *ACL, system demonstration*.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664, Vancouver, Canada. Association for Computational Linguistics.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390, Online. Association for Computational Linguistics.

A Appendix

A.1 Model Implements

In pre-process, we truncate the original long conversations in the dataset with the window size 5. Table 6 shows the dataset statistics.

Datasets	**Train**	**Valid**	**Test**
DailyDialog	44050	4176	6740(Multi-ref)
PersonaChat	68859	8593	8239

Table 6: Dataset statistics.

For non-pretrained experiments, our code is based on fairseq (Ott et al., 2019). We adopt grid search to tune hyper-parameters. On the DailyDialog dataset, the search ranges for learning rate and batch size are $\{0.00008, 0.00010, 0.00012, 0.00015\}$ and $\{112, 160\}$, respectively. On the PersonaChat dataset, the search ranges for learning rate and batch size are $\{0.00010, 0.00012, 0.00015\}$ and $\{32, 64\}$, respectively. We choose the parameter combination with the lowest perplexity in the validation set. Table 7 shows the searched results for each experiment.

Method	**LR(DD)**	**Batch size(DD)**	**LR(PS)**	**Batch size(PS)**
Transformer	1e-4	112	1e-4	32
ResBag	8e-5	160	1e-4	64
TSA	8e-5	160	1.5e-4	32
DD++	8e-5	112	-	-
M&D-D	1e-4	112	1e-4	64
DialoGPS$_{K=1}$	1.5e-4	160	1.5e-4	64
DialoGPS$_{K=2}$	1.5e-4	160	1e-4	64
DialoGPS$_{K=4}$	1.5e-4	112	1.2e-4	64
DialoGPS$_{K=8}$	1.5e-4	160	1.2e-4	64
DialoGPS$_{K=16}$	8e-5	160	-	-

Table 7: Learning rate and batch size in each experiment.

Except for batch size and learning rate, the following important settings: the warmup steps are 4000. We use Adam optimizer with $\beta = (0.9, 0.98)$. Both attention dropout and activation dropout are 0.1. For models trained from scratch, δ on Dailydialog is $\frac{1}{2}$ and $\frac{1}{3}$ on PersonaChat. For fine-tuned models, δ is $\frac{1}{2}$ on two datasets. We select the best checkpoint based on the perplexity in the validation set. Early stop patience is 10 epochs. For pre-trained experiments, on both datasets, the batch size is 64 and learning rate is 0.00002. The training is performed on Nvidia V100 GPU. On DailyDialog: our method takes about 80 minutes when $K = 1$, 4 hours when $K = 16$, and 8 hours

Method	**PersonaChat**	**DailyDialog**
Transformer	2.93	3.08
ResBag	2.93	3.12
TSA	2.92	3.13
DD++	-	**3.24**
M&D-D	2.96	3.13
DialoGPS(K=4)	**3.03**	**3.24**

Table 8: QuantiDCE results on two datasets.

to finetune a $\text{BART}_{\text{large}}$.

Because M&D-D does not suit multi-turn settings, we only use it to modify the last two turns with Okapi BM25 algorithm and we finetune BERT on DailyDialog and PersonaChat respectively to measure the fluency between the last two utterances and the fluency between the penultimate sentence and the above as filtration. In our experiments, on two datasets, the paired sentence set D_p is same as the original training set and the unpaired sentence set D_u is constructed from all sentences in DD++. On DailyDialog, we use multiple references in DD++ as the response bag of ResBag, and on PersonaChat, we use constructed data from M&D-D as its response bag.

A.2 Evaluation Details

Because some evaluation script links of DialoFlow (Li et al., 2021) are out of date, we can not reproduce NIST (Lin and Och, 2004) scores so we do not report it. This issue was also reported by the community [1]. Also, METEOR and Entropy are reproduced. Our reproduced BLEU scores are close to the original paper so we directly quote their results.

Our human evaluators are recruited from Amazon Mturk. In terms of human evaluation, all generated responses are re-capitalized and de-tokenized fairly. The salary for each evaluator is 1 dollar per 10 samples. To give a fair salary, we first evaluate 50 samples by ourselves, calculate the time and effort, and set this amount (samples evaluated by ourselves are just for evaluating the salary, which is not given to evaluators and not reported in the final results).

A.3 QuantiDCE

In addition to the metrics mentioned in the main paper, we further supplement our evaluation with the dialogue-specific metric QuantiDCE (Ye et al., 2021), which measures the coherence between the response and the context. The results show that our proposed DialoGPS outperforms all baseline models.

[1] https://github.com/microsoft/DialoGPT/issues/72

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Following instructions, we add Limitations after Conclusion.

☑ A2. Did you discuss any potential risks of your work?
In Limitations.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
The main claims in the paper are stated in the abstract and in the introduction.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

We use public datasets and open pre-trained models. These are mentioned in many places in the paper such as Introduction and Experiments.

☑ B1. Did you cite the creators of artifacts you used?
We have cited all datasets we use. We have cited open pre-trained models. For example, in Section.1 Introduction and Section.4 Experiments, etc.

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
All open code we use are from github where code is licensed under MIT by default.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
In appendix A1, we report the dataset statistics.

C ☑ Did you run computational experiments?

In Section 4.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
In terms of parameters, we report model structure, e.g., 4-layer transformer, BART large... which have certain parameters. In appendix A1, we report computational budget and GPU version.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
In 4.1.2 and appendix A1, we discuss experimental setup, including hyperparameter search and best-found hyperparameter values.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
We report standard deviation across 5 runs if there's randomness. We report p-value in t-test and kappa value of human evaluation agreement.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
In 4.1.3, we report evaluation metrics. In 4.1.2, 4.1.3, and 4.5, we report pre-trained models we use.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
In 4.1.3 and 4.2.

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
In 4.1.3, we summarized three aspects of evaluation instructions. Also, in appendix A2, before human evaluation, we have de-tokenized and re-capitalized the outputs for a fair and solid evaluation, and thus the instructions are relatively concise.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
We discuss these In appendix A2,

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.