Mitigating Negative Style Transfer in Hybrid Dialogue System

Shimin Li[1], Qinyuan Cheng[1], Linyang Li[1], Xipeng Qiu[1,2] *,

[1] School of Computer Science, Fudan University
[2] Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
{smli20, linyangli19, xpqiu}@fudan.edu.cn, chengqy21@m.fudan.edu.cn

Abstract

As the functionality of dialogue systems evolves, hybrid dialogue systems that accomplish user-specific goals and participate in open-topic chitchat with users are attracting growing attention. Existing research learns both tasks concurrently utilizing a multi-task fusion technique but ignores the negative transfer phenomenon induced by the unique textual style differences. Therefore, contrastive learning based on the latent variable model is used to decouple the various textual genres in the latent space. We devise supervised and self-supervised positive and negative sample constructions for diverse datasets. In addition, to capitalize on the style information contained in the decoupled latent variables, we employ a style prefix that incorporates latent variables further to control the generation of responses with varying styles. We performed extensive experiments on three dialogue datasets, including a hybrid dialogue dataset and two task-oriented dialogue datasets. The experimental results demonstrate that our method can mitigate the negative style transfer issue and achieves state-of-the-art performance on multiple dialogue datasets.

Introduction

Previous research on dialogue systems has been distinctly divided into task-oriented dialogue systems (TOD) (Lee 2021; Su et al. 2022; He et al. 2021; Hosseini-Asl et al. 2020) and open-domain dialogue systems (ODD) (Roller et al. 2021) based on their application. Whereas task-oriented is intended for the successful completion of specific goals and instructions from the user, open-domain dialogue systems engage in open-ended chitchat with the user on various topics. Furthermore, with the development of dialogue systems and the excellent transferability and generalization of pre-trained models to downstream tasks (Qiu et al. 2020), there is a trend towards a progressive integration between different functional dialogue systems (Zhao et al. 2022). Some datasets that fuse various dialogue tasks have also emerged (Young et al. 2021; Chiu et al. 2022; Sun et al. 2021). Nevertheless, different dialogue systems vary in the textual style of their responses to facilitate the completion of a particular dialogue task (Gao et al. 2019). As illustrated in Figure 1, task-oriented dialogue systems use refined discourse to complete the user's task accurately and adequately. In contrast, open-domain dialogue systems generate more varied responses to enhance user engagement and interest.

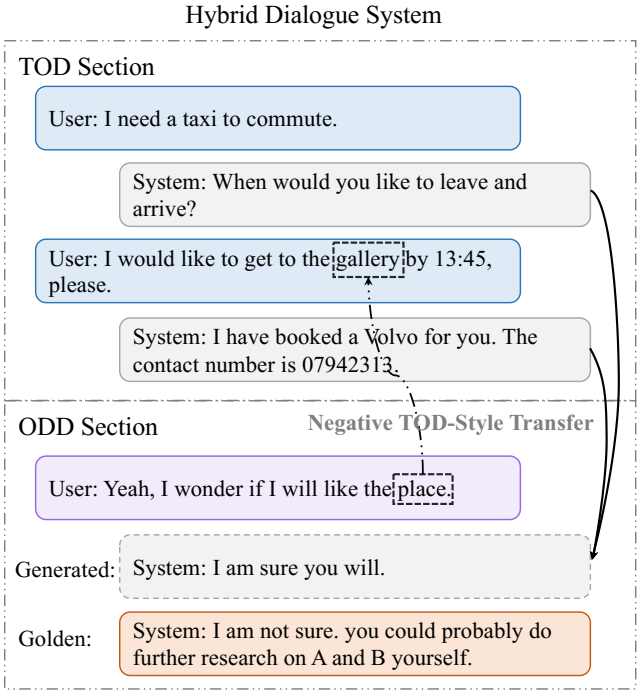

Figure 1: The different styles of responses in the hybrid dialogue system, where TOD responses are in grey, and ODD responses are in orange. The TOD style responses in the ODD section are affected, resulting in mediocre ODD responses.

Existing hybrid dialogue systems learn different dialogue tasks through a multi-task approach or by constructing uniform data patterns (Su et al. 2022; Zhao et al. 2022). However, they ignore the inconsistency in text style between task-oriented and open-domain dialogue responses, leading to the negative transfer of different dialogue tasks in the hybrid dialogue system. Even while the encoder can recognize both dialogue styles exceptionally well, the decoder cannot

*Corresponding author

construct a particular style of discourse during the generation phase (Young et al. 2021). This phenomenon is evidenced by decreased task success rates for task-oriented dialogue and vacuous generic responses for open-domain dialogue. Moreover, some datasets (Budzianowski et al. 2018) do not provide explicit text style labels or relevant indication information, rendering the modeling of text style more intractable (Balasubramanian et al. 2021).

Motivated by the above, our proposed Hybrid-Style-Controlled Dialogue System (HiS-Dialog)[1] is designed with contrastive latent variables to model and distinguish implicit text style information. Also, the style prefix is employed to guide the dialogue system in generating responses of different styles with the modeled style information.

Firstly, to generate diverse text and model text styles, we take variational encoder-decoder (VED) (Serban et al. 2017; Chen et al. 2022) as the base architecture and a pre-trained T5 (Raffel et al. 2020) as the backbone model, which maps deterministic encoder hidden states to latent variables with randomness. Then, during the response generation process, the decoder relies on the different latent variables sampled to generate various styles of dialogue responses. However, VED alone can only model undirected and random implicit text styles, failing to steer the model to generate different response styles with directionality depending on the dialogue task (Balasubramanian et al. 2021).

Therefore, we propose extending the application of contrastive learning (Schroff, Kalenichenko, and Philbin 2015; Chen et al. 2020) to VED to enable decoupling the stylistic latent space into partially distinct textual style subspaces. Precisely, we control the relative distances between different latent variables in the latent style space based on similarity, i.e., latent variables with similar text styles are closer in the style subspace, while the distances between different latent variables are relatively farther apart. Nevertheless, the previous work (Balasubramanian et al. 2021) only applied to cases where style labels were available. For instances without text style descriptors but with alternative subjects or domains, we developed self-supervised techniques with greater application for constructing positive and negative samples (Gao, Yao, and Chen 2021). That is, for all samples in a batch, by reparameterizing twice from its posterior distribution, two variables in the same position of two batches are regarded as positive samples, and variables in other positions are considered negative samples of each other. The method can be extended to some dialogue datasets without explicit textual style information.

Secondly, to exploit the extracted text style variables more adequately and steer the subsequent generation of responses in varying text styles, we designed style prefix as a continuous instruction incorporating style variables for controlled text generation (Li and Liang 2021). Contrary to previous work, which only incorporated latent variables in the self-attention calculation at the decoder side (Chen et al. 2022), HiS-Dialog adds a set of discrete-continuous vectors to both the self-attention and cross-attention calculations at the encoder and decoder. It enables the style prefix to extract more relevant information about the text style from the encoder side and steer the decoder to generate text in the target style.

We validated our approach on a hybrid dialogue dataset and two task-oriented datasets. The experimental results indicate that HiS-Dialog achieves state-of-the-art results in all three datasets.

The contributions of this paper can be summarised as follows:

- We explore the negative text style transfer issue in hybrid dialogue systems for the first time and propose a conditional dialogue style generation model based on contrastive latent variables.
- Depending on the dataset's availability of response style labels, we propose both supervised and self-supervised contrastive latent variable modeling approaches.
- We introduce a style prefix based on style variables, which empowers style information to be extracted and fully utilized by the model to steer the generation of different response styles.

Related Work

Pre-trained Dialogue Models

With the advancement of generative pre-training models (Raffel et al. 2020; Lewis et al. 2020) and increasingly more dialog data (Young et al. 2021; Budzianowski et al. 2018; Chiu et al. 2022), end-to-end pre-trained dialog models are gaining more attention (Su et al. 2022; Hosseini-Asl et al. 2020; Lee 2021). Much of this work addresses the adaptation of pre-trained models to downstream dialogue tasks by further pre-training the models (Zhang et al. 2020; Chen et al. 2022) with additional dialogue data. For instance, PPTOD (Su et al. 2022) extends the idea of the unified paradigm of T5 (Raffel et al. 2020) to the dialogue domain by constructing prompts for further pre-training of different dialogue tasks. Regarding task format construction, work also exists to model different subtasks in dialogue modules into end-to-end form (Hosseini-Asl et al. 2020; Peng et al. 2021; Yang, Li, and Quan 2020). In addition, works also exist that design multiple decoders for different dialogue subtasks (Lee 2021; Zhang, Ou, and Yu 2020; Lin et al. 2020).

However, such approaches to adding decoders can significantly increase the model parameters, thus limiting scalability. Instead, our approach perpetuates the end-to-end setting. It enables the construction of a hybrid dialogue system by introducing a small number of style prefixes without requiring further pre-training or additional decoders.

Hybrid Dialogue Systems

With the gradual unification of the task paradigm (Sun et al. 2022), some studies have attempted to unify several different dialogue systems. In terms of data processing, some research enables the models with some capability of chatting on top of accomplishing the task by inserting open domain dialogues into the task-oriented dialogues (Sun et al. 2021). Regarding the model training approach, most existing works design multiple dialogue tasks as generation tasks and train them in a multi-task fashion (Su et al. 2022; Zhao

[1]Code: https://github.com/whatissimondoing/HiS-Dialog

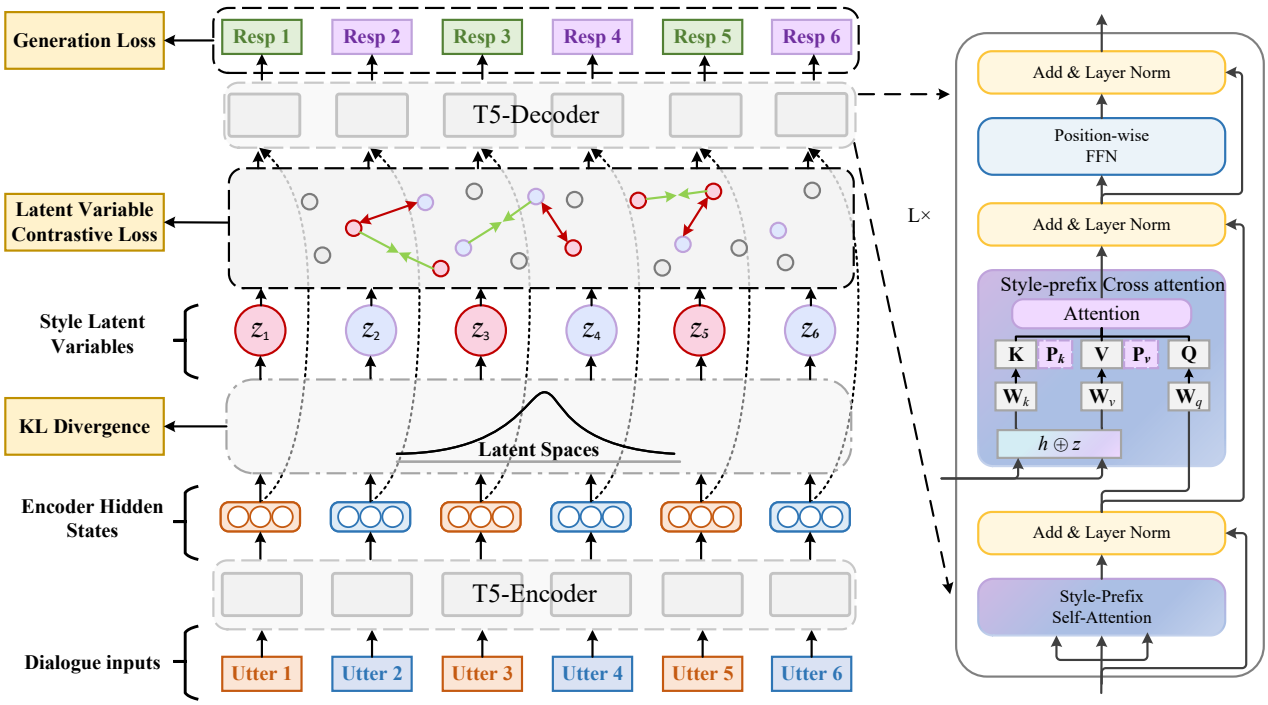

Figure 2: The overall framework of HiS-Dialog consists of the following steps: (1) The dialogue in a batch is encoded to obtain the hidden state and mapped to latent space for latent variables. (2) Contrastive loss is calculated to increase the distance between latent variables of different styles. (3) The style prefix combined with the latent style variable is applied to control the generation of responses with different styles.

et al. 2022). However, they do not account for the relevance of contextual semantics in multiple turns of dialogue when constructing the hybrid dialogue data, making the data incompatible with real-life scenarios. Moreover, multi-task training only extends their model's functionality but neglects the issue of negative transfer across tasks (Pan and Yang 2009).

Unlike previous work, our model adopts a contextually coherent hybrid dialogue dataset for training while introducing contrastive loss in the latent space to mitigate the negative transfer phenomenon of multi-task training for hybrid dialogue tasks.

Methodology

This section describes the basic definition and composition of a hybrid dialogue system and how to constrain the latent space by constructing a contrastive loss on top of the latent variable encoder-decoder model. Style-controlled text generation is then conducted with style prefix incorporating the latent style variable. The overall architecture of our approach is illustrated in Figure 2.

Hybrid Dialogue System

The objective is to model the task-oriented dialogue system and the open-domain dialogue system as a unified end-to-end format (Su et al. 2022). Nevertheless, there are certain discrepancies between the forms of the two tasks. In particular, TOD can be divided into three sub-tasks: dialogue state tracking, dialogue act decision making, and response generation. The ODD, however, contains only dialogue understanding and response generation modules. Therefore, we start by aligning the data schema and training tasks for TOD and ODD.

For a particular turn of conversation t in a multi-turn dialogue, define the user input as U_t and the system-generated responses of different styles $s \in S$ as R_t^s. To model the dialogue history for multiple turns, HiS-Dialog takes all historical turns of conversation as context and generates the belief state B_t:

$$C_t = [U_0, R_0^S, \cdots, U_t], \tag{1}$$

$$B_t = \text{HiS-Dialog}(C_t). \tag{2}$$

Depending on the dialogue belief state B_t, the result D_t, which satisfies the B_t restriction, can be retrieved from the database. Then the previously obtained C_t, B_t and D_t are aggregated to generate the dialogue action A_t:

$$A_t = \text{HiS-Dialog}([C_t, B_t, D_t]). \tag{3}$$

We added act $[chat]$ to the behavioral decision space for the ODD scenario to indicate the system decision for open-domain chit-chat. The information obtained from the individual subtasks can thus be aggregated into a sequence of dialogue messages, termed Dial-INFO $= [C_t, B_t, D_t]$, and is adapted to generate system responses R_t^s with diverse styles:

$$R_t^S = \text{HiS-Dialog}(\text{Dial-INFO}). \tag{4}$$

Style Modeling

Style Latent Variables To alleviate the negative style transfer in hybrid dialogue systems, we should first extract the higher-order style information embedded in the dialogue before the style-controlled text generation. Therefore we utilize the encoder-decoder architecture of the pre-trained model T5 (Raffel et al. 2020) as the backbone model. Then, based on the pre-trained T5, we further introduced latent variables for modeling the styles of different responses. The dialogue information Dial-INFO is first encoded into sentence representations via HiS-Dialog's encoder:

$$h = \text{HiS-Dialog}_{\text{ENC}}(\text{Dial-INFO}). \tag{5}$$

A mapping function is subsequently applied to convert the sentence representation h into the latent variable z containing implicit style information:

$$p_\theta(z|h) = \mathcal{N}(\mu(h), \sigma(h)), \tag{6}$$

where $\mathcal{N}(\mu, \sigma)$ is a normal distribution with mean μ and covariance σ, attained by a multi-layer feed-forward network with different parameters θ. The model can then generate different styles of dialogue responses $p_\theta(R^s|h, z)$ from the different latent variables z sampled from the distribution $p_\theta(z|h)$.

However, since the estimation of the true posterior distribution $p_\theta(z|h, R^s)$ is intractable, the variational posterior distribution $q_\phi(z|h, R^s)$ is introduced to approximate the true posterior distribution. The procedure can be trained by maximizing the variational lower bound on the marginal likelihood as follows:

$$\begin{aligned}\log p_\theta(R^s|h) \geq &-\text{KL}[q_\phi(z|h, R^s)\|p_\theta(z|h)] \\ &+ \mathbb{E}_{q_\phi(z|h,R^s)} \log p_\theta(R^s|z, h),\end{aligned} \tag{7}$$

where the first term KL denotes optimizing the KL divergence $\mathcal{L}_{\text{KL}}$ between the prior and posterior distributions, the latter term represents the maximum likelihood estimate $\mathcal{L}_{\text{MLE}}$ of the generated responses, and ϕ denotes the variational parameters of the multi-layer feed-forward network for estimating the posterior distribution.

Style Disentanglement with Contrast The style information obtained using only latent variables is not attribute-specific oriented, i.e., all possible styles are entangled. It cannot obtain a specific style variable for a particular task (Chen et al. 2019). Therefore, a regularization term is introduced to decouple the styles in the latent space. Specifically, for multiple latent variables representing different text styles sampled in a batch, we expect a certain degree of differentiation between latent variables and retain information on the topic and domain consistency, and therefore introduce a contrastive loss $\mathcal{L}_{\text{CL}}$ with margin. In addition, for the presence or absence of explicit information on text style labels, we design both supervised and self-supervised forms of contrastive loss for different datasets.

For the supervised form, a triplet (z_a, z_p, z_n) can be constructed based on the text style labels (Schroff, Kalenichenko, and Philbin 2015), where z_a denotes the latent variable acting as the anchor, z_p indicates the positive latent variable with the same text style as the anchor, and z_n represents the negative latent variable with a different text style. For all triples in a batch, the contrastive loss can be calculated as:

$$\mathcal{L}_{\text{CL}} = \sum_{i=1}^{|triplets|} \max(0, \lambda + dis), \tag{8}$$

$$dis = d(z_a, z_p) - d(z_a, z_n), \tag{9}$$

$$d(z_i, z_j) = 1 - \frac{z_i z_j}{\|z_i\|\|z_j\|}, \tag{10}$$

where $d(\cdot)$ denotes the function used to measure the cosine distance between two latent variables, and λ indicates the margin between the positive and negative samples.

Furthermore, a self-supervised approach is designed to construct positive and negative samples for the datasets without style labels. Specifically, the posterior distribution $q_\phi(z|h, R^s)$ is re-parameterised twice to obtain z and $\hat{z}$. Thus, for an in-batch contrastive loss, two batches of latent variables at the same index are mutually positive samples. The other latent variables of the different indexes in the batch are mutually negative samples. Then, the triples for computing the contrastive loss are obtained from the two spliced latent variables, $\tilde{triplet} \leftarrow [z, \hat{z}]$, and the contrastive loss is computed with Eq. 8 for $\tilde{triplet}$.

Generation with Style Prefix

Once decoupled style latent variables are obtained, it is essential to consider how the latent variables can be more fully integrated with the pre-trained model to steer the generation of varying text styles. First, the dialogue history information Dial-INFO is encoded by the HiS-Dialog encoder to receive the context representation h and is re-parameterized to obtain the style variable z. Next, the key and value of the decoder in computing the cross-attention are the sum of the context representation h and the latent variable z. For cross-attention in the decoder, the process is computed as:

$$\text{K} = \text{V} = z \oplus h. \tag{11}$$

The distribution of K, V differs from the pre-training stage due to the introduction of z and the relatively small portion of stylized data, which leads the original W^K, W^V to focus primarily on the text content and neglect the stylized information. Therefore, to further enhance the decoder's procedure for different text generation styles using style latent variables, we concatenated a small number of trainable vectors P^K and P^V in front of the keys and values of the attention head as a continuous instruction:

$$head_i = \text{AT}(\text{QW}_i^Q, [\text{P}_i^K, \text{KW}_i^K], [\text{P}_i^V, \text{VW}_i^V]),$$
$$\text{MHA(Q,K,V)} = \text{Concat}(head_1, \cdots, head_n)\text{W}^O,$$

where AT denotes attention computation and MHA indicates multi-head attention (Lin et al. 2021). Thus, for various styles of responses $r_j \in R^s$ of length N, the decoding process for each time step $j \in \{1, \cdots, N\}$ is:

$$h_j^d = \text{HiS-Dialog}_{\text{DEC}}(h, z, h_{<j}^d), \tag{12}$$

where h_j^d denotes the decoder hidden state of each token output during decoding. Then the entire sentence is generated based on the encoder hidden state and the representation generated in the previous decoding time step:

$$p(r_j|r_{j-1}, z) = \text{Softmax}(\text{W}_d \cdot h_j^d), \tag{13}$$

where W_b is the parameter matrix that maps the decoder output to the vocabulary distribution. Finally, the overall training loss is:

$$\mathcal{L} = \mathcal{L}_{\text{MLE}} + \alpha\mathcal{L}_{\text{KL}} + \beta\mathcal{L}_{\text{CL}}, \tag{14}$$

where α and β are coefficients that control the scale of different losses.

Experiments

Evaluation Datasets

We conducted experiments in end-to-end settings under three dialogue datasets containing one hybrid dialogue dataset, FusedChat, and two task-oriented datasets, MultiWOZ2.0, and MultiWOZ2.1.

FusedChat (Young et al. 2021) This dataset expands or rewrites each conversation based on the task-oriented task. By annotating the original TOD data with additional rounds of context-semantically relevant open-domain dialogue, a conversation can contain multiple dialogue patterns while increasing the average number of turns by 5.8.

MultiWOZ (Budzianowski et al. 2018; Eric et al. 2020) This dataset is one of the most prevalent datasets in task-oriented dialogue systems, collected via Wizard-of-Oz, and contains a total of 8438/1000/1000 multiturn dialogues. Among them, MultiWOZ 2.0 is the initially proposed multi-domain goal-directed task-oriented dataset, while MultiWOZ 2.1 is a version with several buggy annotations fixed. We evaluate both datasets to assess the robustness of the model. The reason for analyzing these datasets is that they do not provide text style information but implicitly reflect the text style of different domains or themes, allowing the method's effectiveness to be validated in a self-supervised setting.

Metrics

For the assessment metrics in task completion, we utilized the evaluation scripts provided by the dataset to evaluateitepmultiwoz of the experimental results. In particular, **Inform** was used to evaluate whether the system generated the entities mentioned by the user, **Success** measured whether all the slots requested by the user were fulfilled, and **BLEU** assessed the fluency of the system in generating responses. The final combined score is calculated as Combined = (Inform + Success) * 0.5 + BLEU.

To assess the efficacy of open-domain response generation, we calculated the number of unique 1/2/3-grams in a sentence using Distinct-1/2/3 (Li et al. 2016) to measure the diversity of responses generated by the system.

Moreover, we calculated BLEU for both task-oriented and open-domain dialogue responses to evaluate the model's linguistic quality in generating task-related and open-domain responses.

Implementation Setup

HiS-Dialog and other baselines based on pre-trained models are implemented with HuggingFace's Transformers. We employ AdamW as the optimizer and configure the warmup rate to 0.1. For FusedChat, the learning rate is 6e-4 and the batch size is set to 12 for 12 epochs. For MultiWOZ, we set the learning rate 5e-4, epoch 10, and batch size 12. All experiments were performed on a GeForce RTX 3090 GPU (24G), and the mean of the results from three different random seeds was selected as the final result.

Compared Baselines

We conducted comparative experiments for the three benchmark datasets with the following robust and relatively new end-to-end dialogue systems based on pre-trained models.

Encoder-decoder architectures: (1) **DAMD** (Zhang, Ou, and Yu 2020) augments dialogue actions with additional data to provide more diverse responses. (2) **MinTL** (Lin et al. 2020) generates dialogue states and responses in succession with two decoders. (3) **PPTOD** (Su et al. 2022) models multiple dialogue tasks uniformly as generative tasks by constructing task-specific prompts. (3) **MTTOD** (Lee 2021) introduces an additional span prediction task on the encoder side on top of the two-decoder structure. (5) **T5-CVAE** (Du et al. 2022) models additional latent variables on top of T5. (6) **DoTS** (Jeon and Lee 2021) reduces the requirement for historical dialogue length by additionally modelling the domain state.

Auto-aggressive architectures: (1) **SimpleTOD** (Hosseini-Asl et al. 2020) models all tasks in TOD as autoregressive generation with GPT2 (Radford et al. 2019). (2) **SOLOIST** (Peng et al. 2021) further pre-trains the model with heterogeneous dialogue datasets. (3) **UBAR** (Yang, Li, and Quan 2020) additionally adding dialogue states, database information, and dialogue actions to the dialogue history.

Overall Performance

Table 1 illustrates the results of HiS-Dialog with other baselines on the hybrid dialogue dataset FusedChat. It is evident from the results that our approach achieves state-of-the-art results in both the TOD segment and the ODD segment. In the TOD part, HiS-Dialog's Inform and Success scores for task completion improved by 0.4 and 1.5 points, respectively, compared to the T5-CVAE (Du et al. 2022). In the ODD section, HiS-Dialog obtained the latest results on both Distinct-2 and BLEU, indicating our method achieves higher diversity and quality of open domain text. Combining the results of TOD and ODD, it is observed that HiS-Dialog achieves a significant performance improvement in both subtasks compared to the other baselines. The results

Dataset	FusedChat							
	TOD				ODD			
Metrics	Inform	Success	BLEU	Combined Score	DIST-1	DIST2	DIST-3	BLEU
MinTL (Lin et al. 2020)	80.80	74.40	16.20	93.80	0.03	0.15	0.28	9.02
PPTOD (Su et al. 2022)	90.40	82.50	14.32	100.77	0.03	0.14	0.26	9.00
MTTOD (Lee 2021)	90.50	82.50	17.23	103.73	0.04	0.16	**0.36**	9.07
T5 (Raffel et al. 2020)	90.30	81.90	17.48	103.58	0.03	0.15	0.29	9.26
T5-CVAE (Du et al. 2022)	91.60	82.90	16.88	104.13	0.04	0.15	0.28	9.11
HiS-Dialog	**92.00** (±0.78)	**84.40** (±0.65)	**17.58** (±0.82)	**105.73** (±0.93)	**0.04** (±0.01)	**0.17** (±0.04)	0.31 (±0.08)	**9.51** (±0.26)

Table 1: Overall results of HiS-Dialog compared to other methods under two subtasks of FusedChat.

Dataset	MultiWOZ 2.0				MultiWOZ 2.1			
Metrics	Inform	Success	BLEU	Combined Score	Inform	Success	BLEU	Combined Score
DAMD (Zhang, Ou, and Yu 2020)	76.33	60.40	16.60	84.97	-	-	-	-
SimpleTOD (Hosseini-Asl et al. 2020)	84.40	70.10	15.01	92.26	85.00	70.50	15.23	91.98
DoTS (Jeon and Lee 2021)	86.59	74.14	15.06	95.43	86.65	74.18	15.90	96.32
SOLOIST (Peng et al. 2021)	85.50	72.90	16.54	95.74	-	-	-	-
MinTL-BART (Lin et al. 2020)	84.88	74.91	17.89	97.79	-	-	-	-
UBAR (Yang, Li, and Quan 2020)	95.40	80.70	17.00	105.10	**95.70**	81.80	16.50	105.25
MTTOD (Lee 2021)	91.00	82.60	**21.60**	108.30	91.00	82.10	**21.00**	107.50
PPTOD (Su et al. 2022)	89.20	79.40	18.62	102.92	87.09	79.08	19.17	102.26
T5 (Raffel et al. 2020)	90.70	81.30	18.94	104.94	91.10	82.00	18.34	104.89
HiS-Dialog	**92.85**	**84.30**	20.12	**108.70**	92.30	**83.90**	19.76	**107.86**

Table 2: HiS-Dialog results in an end-to-end paradigm on two different versions of the MultiWOZ dataset.

also indirectly indicate that HiS-Dialog is more capable of mitigating the negative transfer phenomenon between tasks, thus improving the overall effectiveness of the model.

We evaluated the performance of the proposed method in a self-supervised setting without text style labels, and exploratory experiments were done using MultiWOZ with the different implicit domain or topic styles. In addition, we assume that `inform` and `success` are utilized to signify task completion, and `BLEU` is employed to measure the quality of text generation. Hence, a rise in each of the two metrics indicates that the negative style transfer issue has been alleviated. Table 2 indicates that HiS-Dialog achieved the latest combined scores on the MultiWOZ 2.0 and 2.1 datasets. Of these, MTTOD has the relatively highest BLEU scores on both datasets since MTTOD employs two decoders for dialogue states and response generation, respectively, but with a 50% increase in the overall number of model parameters. Nonetheless, HiS-Dialog achieves the latest combined scores with only contrastive latent variable and style prefix that adds a small number of parameters, demonstrating that our approach mitigates the effects of negative style transfer, resulting in enhancements to both task completion rate and text generation quality.

Effect on Mitigating Negative Style Transfer

Ablation This section experiments on the effects of latent variable contrastive learning and style prefix in HiS-Dialog on mitigating the negative style transfer problem in hybrid dialogue systems. The experimental results in Table 3 illustrate that adding additional contrastive loss to the optimization process in the latent space for constraint leads to a 1.0 and 1.4 point improvement in task success rate and text generation quality, thus corroborating the gain of contrastive learning in mitigating negative style transfer.

	Inform	Success	BLEU	Comb
HiS-Dialog	90.20	81.80	18.60	104.60
+CLV	91.20 (↑1.00)	83.20 (↑1.40)	19.05 (↑0.45)	106.25 (↑1.65)
+SP	92.50 (↑2.30)	83.75 (↑1.95)	19.63 (↑1.03)	107.76 (↑3.16)
+ CLV+SP	92.60 (↑**2.40**)	84.60 (↑**2.80**)	19.79 (↑**1.19**)	108.39 (↑**3.79**)

Table 3: Efficacy of contrastive latent variables (CLV) and style prefix (SP) in mitigating negative style transfer.

Furthermore, by additionally introducing a style prefix into the attention calculation, the impact of negative style transfer of responses of different styles can also be mitigated, thus significantly improving the overall performance of the hybrid dialogue system. Lastly, adding the two mechanisms together improves the overall score of the hybrid dialogue system by 3.79 points, suggesting that style prefixing enables further style-controlled text generation by fully using the well-defined style variables obtained from latent space modeling.

	FusedChat		MultiWOZ	
Metric	Informative ↓	Differential ↓	Informative ↓	Differential ↓
PPTOD	3.27	3.57	2.75	3.26
MTTOD	2.89	2.56	2.71	**2.24**
HiS-Dialog	**2.16**	**2.54**	**2.63**	2.39
Golden	1.68	1.23	2.01	2.11

Table 4: The responses generated by the different approaches were ranked by human evaluation. Informative indicates the informative adequacy of the generated sentences, Differential denotes the distinguishability of the different text styles, and Golden refers to the correct responses used for reference.

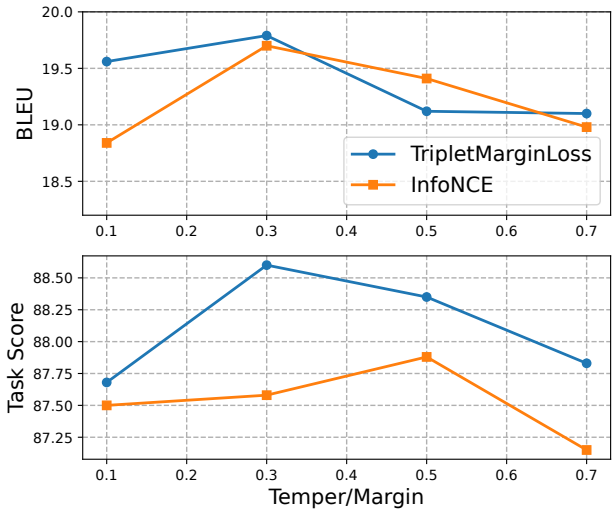

Figure 3: The effect of various contrastive loss functions and varying margin/temperature coefficients on performance.

Human Evaluation To further assess the quality of the generated responses from different models, we evaluated the informativeness and distinguishability of their generated responses on FusedChat and MultiWOZ. We randomly selected 50 dialogues from both datasets and invited five domain-related practitioners to rank the responses generated by these methods. Table 4 clearly illustrates that the informativeness of HiS-Dialog-generated responses receives the highest-ranking scores among the baselines in both datasets. Regarding distinguishability of response styles, HiS-Dialog obtains the highest ranking in the hybrid dialogue dataset FusedChat, while MTTOD obtains the highest-ranking score on MultiWOZ. Also, FusedChat provides explicit text style labels, allowing HiS-Dialog's contrastive learning to learn better style variables. In contrast, MultiWOZ does not provide style labels, so the two decoder architectures of MTTOD can generate more distinguishable sentences.

Analysis

Efficacy of Latent Variable Contrastive Learning We explored the efficacy of latent variable contrastive learning with two forms of loss functions and varying sizes of hyper-parameters. TripletMarginLoss (Schroff, Kalenichenko, and Philbin 2015) is the triplet loss employed in our method, and InfoNCE (Gao, Yao, and Chen 2021) is another widely used loss function for contrastive learning. The hyperparameter explored in the triplet loss is the relative distance margin between positive and negative samples, and the temperature coefficient τ is analyzed in InfoNCE.

To explore the influence of different settings on task completion and generation quality, we define `Task Score` = (`Inform` + `Success`) * 0.5 for task completion and BLEU to assess the quality of text generation. As illustrated in Figure 3, the overall results of triplet loss compared to InfoNCE are generally better under different hyperparameters. A possible explanation is that InfoNCE only pulls in the distance between the anchor point and the positive sample. In contrast, triplet loss further considers the relative distance between the anchor point and the negative samples. Also, 0.3 was also chosen as the best margin value in our method.

Prefix type	Prefix-length	Inform	Success	BLEU	Combine
Enc+Dec+Cross	30	92.45	83.30	19.35	107.22
	50	92.60	**84.60**	**19.79**	**108.39**
	70	91.50	83.20	19.52	106.87
	100	92.60	84.60	19.32	107.92
Dec	50	91.60	83.40	19.61	107.11
Enc+Dec	50	**93.00**	84.35	19.60	108.27

Table 5: Impact of different lengths and positions of style prefix on the effectiveness of hybrid dialogue systems.

Analysis of Style Prefix We investigate the effect of different lengths and positions of style prefixes on the quality of text generation for style controllable. From Table 5, we can observe that HiS-Dialog achieves the highest combined score when the size of the style prefix is set at 50. As the prefix length increases, the system performance degrades to some extent, suggesting that excessively long style prefixes can cause the model to over-fit to style-specific information, thereby reducing the diversity of the generation. Style prefix being added to both the encoder side and the cross-attention computation results in better performance than adding style prefix to the decoder side alone. It indicates that the style prefix on the encoder side can further extract style-related information, and style prefixes in the cross-attention computation can better guide the decoder's controlled style generation with the extracted style information.

Conclusion

We propose a novel latent variable encoder-decoder model, HiS-Dialog, which incorporates contrastive loss to constrain the latent space for better control of varying styles of text generation and to mitigate negative style transfer. We also introduce a style prefix to fully exploit the modeled style latent variables to guide the process of generating diverse styles of responses. Empirical results on three dialogue datasets demonstrate that our approach mitigates the occurrence of negative transfer in a better way than previous baselines and achieves improvements in several aspects.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (No.2020AAA0108700) and National Natural Science Foundation of China (No.62022027).

References

Balasubramanian, V.; Kobyzev, I.; Bahuleyan, H.; Shapiro, I.; and Vechtomova, O. 2021. Polarized-VAE: Proximity Based Disentangled Representation Learning for Text Generation. In Merlo, P.; Tiedemann, J.; and Tsarfaty, R., eds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, 416–423. Association for Computational Linguistics.

Budzianowski, P.; Wen, T.; Tseng, B.; Casanueva, I.; Ultes, S.; Ramadan, O.; and Gasic, M. 2018. MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, 5016–5026. Association for Computational Linguistics.

Chen, M.; Tang, Q.; Wiseman, S.; and Gimpel, K. 2019. A Multi-Task Approach for Disentangling Syntax and Semantics in Sentence Representations. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 2453–2464. Association for Computational Linguistics.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. E. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, 1597–1607. PMLR.

Chen, W.; Gong, Y.; Wang, S.; Yao, B.; Qi, W.; Wei, Z.; Hu, X.; Zhou, B.; Mao, Y.; Chen, W.; Cheng, B.; and Duan, N. 2022. DialogVED: A Pre-trained Latent Variable Encoder-Decoder Model for Dialog Response Generation. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, 4852–4864. Association for Computational Linguistics.

Chiu, S.; Li, M.; Lin, Y.; and Chen, Y. 2022. SalesBot: Transitioning from Chit-Chat to Task-Oriented Dialogues. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, 6143–6158. Association for Computational Linguistics.

Du, W.; Zhao, J.; Wang, L.; and Ji, Y. 2022. Diverse Text Generation via Variational Encoder-Decoder Models with Gaussian Process Priors. *CoRR*, abs/2204.01227.

Eric, M.; Goel, R.; Paul, S.; Sethi, A.; Agarwal, S.; Gao, S.; Kumar, A.; Goyal, A. K.; Ku, P.; and Hakkani-Tür, D. 2020. MultiWOZ 2.1: A Consolidated Multi-Domain Dialogue Dataset with State Corrections and State Tracking Baselines. In Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Moreno, A.; Odijk, J.; and Piperidis, S., eds., *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, 422–428. European Language Resources Association.

Gao, T.; Yao, X.; and Chen, D. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In Moens, M.; Huang, X.; Specia, L.; and Yih, S. W., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, 6894–6910. Association for Computational Linguistics.

Gao, X.; Zhang, Y.; Lee, S.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2019. Structuring Latent Spaces for Stylized Response Generation. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, 1814–1823. Association for Computational Linguistics.

He, W.; Dai, Y.; Zheng, Y.; Wu, Y.; Cao, Z.; Liu, D.; Jiang, P.; Yang, M.; Huang, F.; Si, L.; Sun, J.; and Li, Y. 2021. GALAXY: A Generative Pre-trained Model for Task-Oriented Dialog with Semi-Supervised Learning and Explicit Policy Injection. *CoRR*, abs/2111.14592.

Hosseini-Asl, E.; McCann, B.; Wu, C.; Yavuz, S.; and Socher, R. 2020. A Simple Language Model for Task-Oriented Dialogue. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jeon, H.; and Lee, G. G. 2021. Domain state tracking for a simplified dialogue system. *arXiv preprint arXiv:2103.06648*.

Lee, Y. 2021. Improving End-to-End Task-Oriented Dialog System with A Simple Auxiliary Task. In Moens, M.; Huang, X.; Specia, L.; and Yih, S. W., eds., *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, 1296–1303. Association for Computational Linguistics.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J. R., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 7871–7880. Association for Computational Linguistics.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In Knight, K.; Nenkova, A.; and Rambow, O., eds., *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, 110–119. The Association for Computational Linguistics.

Li, X. L.; and Liang, P. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, 4582–4597. Association for Computational Linguistics.

Lin, T.; Wang, Y.; Liu, X.; and Qiu, X. 2021. A Survey of Transformers. *CoRR*, abs/2106.04554.

Lin, Z.; Madotto, A.; Winata, G. I.; and Fung, P. 2020. MinTL: Minimalist Transfer Learning for Task-Oriented Dialogue Systems. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, 3391–3405. Association for Computational Linguistics.

Pan, S. J.; and Yang, Q. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10): 1345–1359.

Peng, B.; Li, C.; Li, J.; Shayandeh, S.; Liden, L.; and Gao, J. 2021. SOLOIST: Building Task Bots at Scale with Transfer Learning and Machine Teaching. *Trans. Assoc. Comput. Linguistics*, 9: 907–824.

Qiu, X.; Sun, T.; Xu, Y.; Shao, Y.; Dai, N.; and Huang, X. 2020. Pre-trained Models for Natural Language Processing: A Survey. *CoRR*, abs/2003.08271.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.*, 21: 140:1–140:67.

Roller, S.; Dinan, E.; Goyal, N.; Ju, D.; Williamson, M.; Liu, Y.; Xu, J.; Ott, M.; Smith, E. M.; Boureau, Y.; and Weston, J. 2021. Recipes for Building an Open-Domain Chatbot. In Merlo, P.; Tiedemann, J.; and Tsarfaty, R., eds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, 300–325. Association for Computational Linguistics.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. FaceNet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 815–823. IEEE Computer Society.

Serban, I. V.; Sordoni, A.; Lowe, R.; Charlin, L.; Pineau, J.; Courville, A. C.; and Bengio, Y. 2017. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. In Singh, S.; and Markovitch, S., eds., *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, 3295–3301. AAAI Press.

Su, Y.; Shu, L.; Mansimov, E.; Gupta, A.; Cai, D.; Lai, Y.; and Zhang, Y. 2022. Multi-Task Pre-Training for Plug-and-Play Task-Oriented Dialogue System. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, 4661–4676. Association for Computational Linguistics.

Sun, K.; Moon, S.; Crook, P. A.; Roller, S.; Silvert, B.; Liu, B.; Wang, Z.; Liu, H.; Cho, E.; and Cardie, C. 2021. Adding Chit-Chat to Enhance Task-Oriented Dialogues. In Toutanova, K.; Rumshisky, A.; Zettlemoyer, L.; Hakkani-Tür, D.; Beltagy, I.; Bethard, S.; Cotterell, R.; Chakraborty, T.; and Zhou, Y., eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, 1570–1583. Association for Computational Linguistics.

Sun, T.; Liu, X.; Qiu, X.; and Huang, X. 2022. Paradigm Shift in Natural Language Processing. *Int. J. Autom. Comput.*, 19(3): 169–183.

Yang, Y.; Li, Y.; and Quan, X. 2020. UBAR: Towards Fully End-to-End Task-Oriented Dialog Systems with GPT-2. *CoRR*, abs/2012.03539.

Young, T.; Xing, F. Z.; Pandelea, V.; Ni, J.; and Cambria, E. 2021. Fusing task-oriented and open-domain dialogues in conversational agents. *CoRR*, abs/2109.04137.

Zhang, Y.; Ou, Z.; and Yu, Z. 2020. Task-Oriented Dialog Systems That Consider Multiple Appropriate Responses under the Same Context. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, 9604–9611. AAAI Press.

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, B. 2020. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In Celikyilmaz, A.; and Wen, T., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, 270–278. Association for Computational Linguistics.

Zhao, X.; He, B.; Wang, Y.; Li, Y.; Mi, F.; Liu, Y.; Jiang, X.; Liu, Q.; and Chen, H. 2022. UniDS: A Unified Dialogue System for Chit-Chat and Task-oriented Dialogues. In Feng, S.; Wan, H.; Yuan, C.; and Yu, H., eds., *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering, DialDoc@ACL 2022, Dublin, Ireland, May 26, 2022*, 13–22. Association for Computational Linguistics.