Seen to Unseen: Exploring Compositional Generalization of Multi-Attribute Controllable Dialogue Generation

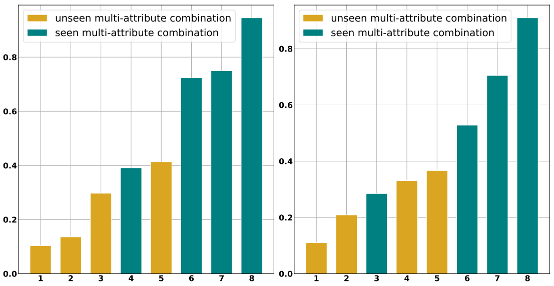
Weihao Zeng[1*], Lulu Zhao[1*], Keqing He[2], Ruotong Geng[1]
Jingang Wang[2], Wei Wu[2], Weiran Xu[1*]
[1]Beijing University of Posts and Telecommunications, Beijing, China
[2]Meituan, Beijing, China
{zengwh,zhaoll,ruotonggeng,xuweiran}@bupt.edu.cn
{hekeqing,wangjingang,wuwei}@meituan.com

Abstract

Existing controllable dialogue generation work focuses on the single-attribute control and lacks generalization capability to out-of-distribution multiple attribute combinations. In this paper, we explore the compositional generalization for multi-attribute controllable dialogue generation where a model can learn from seen attribute values and generalize to unseen combinations. We propose a prompt-based disentangled controllable dialogue generation model, DCG. It learns attribute concept composition by generating attribute-oriented prompt vectors and uses a disentanglement loss to disentangle different attributes for better generalization. Besides, we design a unified reference-free evaluation framework for multiple attributes with different levels of granularities. Experiment results on two benchmarks prove the effectiveness of our method and the evaluation metric.

1 Introduction

Recently, large pre-trained language models (PLMs) like DialoGPT (Zhang et al., 2020), BlenderBot (Roller et al., 2020) and Meena (Adiwardana et al., 2020) can produce fluent and relevant responses for dialogue contexts. However, the generated responses are often uninformative and factual inconsistent. Hence, controllable dialogue generation (CDG) is proposed to guide dialogue generation towards the desired attributes such as emotions (Zhou et al., 2018), acts (Li et al., 2017), and personas (Zhang et al., 2018). Previous work focused on directly fine-tuning the large-scale PLMs (Keskar et al., 2019) or using an extra attribute discriminator (Krause et al., 2021; Dathathri et al., 2019) to guide generation. The former is expensive and requires extensive annotated attribute labels. The decoding of the latter is computationally intensive, reducing the response fluency and generation speed.



(a) E-ACC (b) A-ACC

Figure 1: The difference of controllability scores on seen and unseen multi-attribute combinations of CTRL (Keskar et al., 2019). E-ACC and A-ACC denote emotion accuracy and act accuracy.

Although these methods have made some progress in CDG, most of them focus on single-attribute generation where there is only one attribute label like *happiness* in emotion and pay less attention to the multi-attribute generation, which is a more practical setting. Therefore, we are committed to filling this gap in CDG. Noted that different from single-attribute, the control signal of the multi-attribute generation is a combination of multiple values from different attributes, which faces the challenge of lacking sufficient annotated attribute-specific data. We also find state-of-the-art methods for multi-attribute controllable text generation (Yang et al., 2022; Qian et al., 2022), which combine controllers learned from single-attribute, only suitable for discrete attributes with specific labels (Li et al., 2017) but not for continuous attributes (Zhang et al., 2018). More importantly, we further show directly applying all existing models achieves superior attribute accuracy on seen attribute combinations but drops significantly on unseen combinations, as shown in Figure 1. It proves that previous work lacks compositional generalization capability from seen attribute values to unseen combinations. Besides, the evaluation of controllability in CDG is severely limited by attribute types and annotated attribute data (Du and Ji, 2021), which is not ap-

[*] The first two authors contribute equally. Weiran Xu is the corresponding author.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 14179–14196
July 9-14, 2023 ©2023 Association for Computational Linguistics

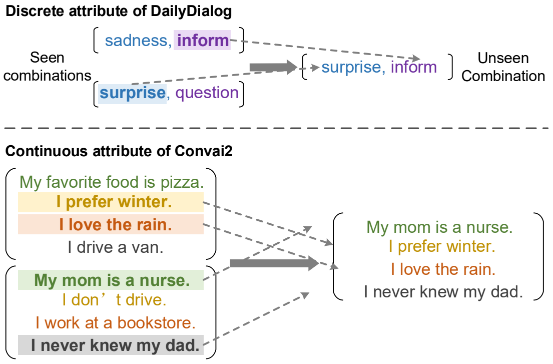


Figure 2: Examples of the compositional generalization for coarse-grained discrete attributes and fine-grained continuous attributes.

plicable to all cases. Therefore, it is valuable to explore a unified and efficient evaluation metric.

In this paper, we try to explore the compositional generalization for multi-attribute controllable dialogue generation where a model could learn from seen attribute values and generalize to unseen combinations. Figure 2 shows two granularities of multi-attribute compositional generalization, where the token-level attribute labels are regarded as coarse-grained discrete attributes and the sentence-level attribute descriptions are regarded as fine-grained continuous attributes. Specifically, we propose a **D**isentangled **C**ontrollable **G**eneration model (**DCG**), for compositional generalization in multi-attribute controllable dialogue generation. Inspired by prompt learning (Lester et al., 2021), we adopt the attribute values in a combination as attribute-oriented prompts to elicit knowledge from PLMs where the prompts for all instances learn a shared transformation layer, instead of learning an independent prompt representation for each attribute value (Clive et al., 2022; Qian et al., 2022; Yang et al., 2022). Our method helps transfer attribute concepts from seen values to unseen combinations by learning different prompt embeddings and is easily applied to attribute combination with a huge number of discrete or continuous attribute values. To further disentangle different attribute values, we construct a set of pseudo combinations and design a novel objective of controllable attribute combinations for prompt-tuning, which separates desired attribute combination from others.

Furthermore, to unify the evaluation of different granularity attributes, we design a novel and general reference-free evaluation framework, i.e. **M**ultiple **A**ttribute **E**valuation (**MAE**), to measure the consistency between desired seen/unseen attribute combinations and generated responses. Specifically, the evaluation of each attribute is converted to a text-to-text generation task based on T5 (Raffel et al., 2020) with handcrafted templates, and the generated probability of "yes" is regarded as the controllability score. To mitigate the potential bias of different handcrafted modalities (Zhao et al., 2019; Ke et al., 2022), we add a trainable continuous prompt to improve stability and robustness. Through human evaluation, we show that our proposed evaluation metric can handle both coarse-grained discrete attributes and fine-grained continuous attributes well.

Our contributions are as follows: (1) To the best of our knowledge, we are the first to explore the compositional generalization for multi-attribute controllable dialogue generation and find existing models lack generalization capability to out-of-distribution multi-attribute combinations. (2) We propose a disentangled controllable generation, DCG, which learns attribute concepts from seen values to unseen combinations via a shared mapping of attribute-oriented prompts and uses a disentanglement loss to disentangle different attribute combinations. (3) We introduce a unified reference-free evaluation framework, MAE, for different granularities of attributes. Two benchmarks are established and sufficient experiment results prove the effectiveness of our method and evaluation metric.

2 Related Work

Controllable Dialogue Generation Currently, there have existed many studies on CDG (Zhou et al., 2018; Li et al., 2017; Zhang et al., 2018). CTRL (Keskar et al., 2019) used 55 kinds of attribute control codes to finetune an LM which is expensive and requires extensive annotated attribute labels. Krause et al. (2021); Dathathri et al. (2019); Yang and Klein (2021); Lin and Riedl (2021) addressed these limitations by employing an attribute discriminator to update the hidden activations or re-weight the next token distributions, resulting in a slow inference speed. Despite the progress, these models all focus on the single-attribute CDG where the attribute only contains coarse-grained discrete values, such as *happiness* in emotion-controlled generation. It is also vital to explore multi-attribute CDG with multi-granularity attributes. Recently, some works (Yang et al., 2022; Qian et al., 2022) extend to multi-attribute controllable text generation by simply concatenating the prefixes trained for single attribute. However, they are only suitable for discrete attributes but not for fine-grained continuous attributes like personas (Zhang et al., 2018). Besides, we find all these methods have a large performance drop from seen attribute values to unseen combinations. Therefore, in this paper, we are the first to explore the compositional generalization for multi-attribute CDG where a model could learn from seen attributes and generalize to out-of-distribution (OOD) combinations.

Compositional Generalization in NLP Compositional generalization has gradually attracted the interest of NLP researchers. The main application is in semantic parsing, involving grammar-based approaches (Herzig and Berant, 2021), data augmentation strategies (Oren et al., 2020), disentangled representations (Zheng and Lapata, 2022), etc. Recently, a large-scale benchmark, STYLEPTB, is constructed to advance the development of compositional style transfer (Lyu et al., 2021), and a template-based input representation is also performed on the data-to-text task (Mehta et al., 2022). Overall, the application of compositional generalization in NLP tasks is not widespread and there is no related work on CDG at all.

Prompt Learning Prompt-based methods have achieved significant success in many NLP fields (Lester et al., 2021; Schick and Schütze, 2021). Li and Liang (2021) proposed the task-specific continuous prompts to finetune a NLG model. For controllable generation, Clive et al. (2022); Qian et al. (2022); Yang et al. (2022) applied the prompt learning to represent each attribute value as an independent prefix. However, those methods are impractical for fine-grained attributes with a large value set. In contrast, we use the control codes to generate attribute-oriented prompts to guide the generation via a shared MLP layer.

3 Problem Formulation

Given a predefined set of attributes $\mathcal{X} = \{A, B, C, ...\}$, each attribute contains various values $A = \{a_1, ..., a_k\}$ and k is the number of values of attribute A. Multi-attribute controlled dialogue response generation aims to generate responses r that satisfy multiple desirable attributes $c = (a_1, b_2, ...)$ conditioned on the dialogue history d, where a_1 and b_2 are one value of the attribute A and B, and $c \in C_v$ is a combination of attribute values. It can be symbolized as $p(r|d, a_1, b_2, ...)$, $(a_1 \in A, b_2 \in B, ...)$.

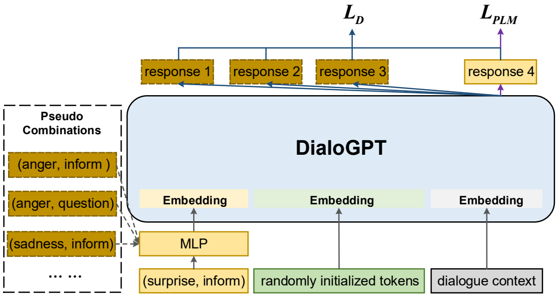


Figure 3: Overall architecture of our DCG model.

In this paper, we further focus on the multi-attribute compositional generalization, where the combinations of multiple attribute values for the training set and the test set are disjoint, i.e., $C_{v,train} \cap C_{v,test} = \varnothing$.

4 Methodology

As shown in Figure 3, our model is on the basis of the framework of DialoGPT (Zhang et al., 2020) with the compositional prompt module.

4.1 Compositional Prompt

4.1.1 Prompt Design

To better use the control signals, we design two types of prompts to elicit the attribute-related information from the PLM:

Attribute-oriented Prompt We use the combination of controlled attribute values corresponding to each instance as prompts to guide the model to focus on the controlled information in the dialogue. Here, the controlled attribute values are discrete attribute labels in DailyDialog or continuous attribute descriptions in ConvAI2. The multiple attribute values $a_{i,\cdot}$ in the corresponding combination c are simply concatenated as an attribute-oriented prompt sequence, i.e., $p_{att} = [a_1, b_2, ...]$. We encode the prompt tokens using the word embedding layer of a pre-trained DialogGPT and then employ a shared MLP_{θ_1} to generate the embeddings E_{att} of the attribute-oriented prompts. Note that we don't require independent parameters for each attribute value like Clive et al. (2022); Qian et al. (2022); Yang et al. (2022), but only a shared transformation MLP layer.

Task-oriented Prompt Although attribute-oriented prompts capture the instance-specific control signals, the dialogue response generation task also is guided by the instance-independent global features. Following Lester et al. (2021), we adopt a series of randomly initialized tokens as the task-oriented prompt, i.e., $p_{task} = [p_1, ..., p_m]$, where m is the length of the task-oriented prompt sequence. We look up this prompt sequence in the randomly initialized embedding table M_{θ_2} and get the prompt embeddings E_{task}.

Finally, we concatenate the two prompt embeddings as the whole prompt embeddings, i.e., $E_p = [E_{att}; E_{task}]$.

4.1.2 Disentanglement Learning

Given an instance (d, c), d is the dialogue history and c is the combination of controllable attribute values. To force the model to distinguish different combinations of multiple attribute values, we design some pseudo combinations to enhance the diversity of the prompts, which improves the generalization ability of our model. A disentanglement loss $\mathcal{L}_D$ is further introduced to disentangle the combination representations and train multiple compositional prompts simultaneously:

$$\mathcal{L}_D = -log\frac{P(r|d,c)}{P(r|d,c) + \sum_{c^{'} \in C_{pse}} P(r|d,c^{'})} \tag{1}$$

where C_{pse} is the set of pseudo combinations and at least one value in the combination $c^{'}$ is different from the corresponding value in the golden combination.[1] Here, we maximize the generated likelihood of the desirable positive combination $P(r|d,c)$ against the generated likelihood of pseudo combinations $P(r|d,c^{'})$ to generate more controllable responses relevant to given attributes.

4.2 Training Strategy

We use DialoGPT (Zhang et al., 2020) as the backbone of our model. Given the dialogue history d, the embedding E_d is obtained by DialoGPT. Then, the embeddings of the prompt sequence E_p are prepended to the E_d as a whole input embedding matrix. Overall, the PLM loss is calculated as:

$$\mathcal{L}_{PLM} = -\sum_{t=1}^{T} \log p_{\theta_1, \theta_2, \varphi}(y_t | y_{<t}, d, p_{att}, p_{task}) \tag{2}$$

where T is the length of generated sequence, i.e., the dialogue history and response. φ is the parameter of the PLM and is fixed. The parameters of two prompts, θ_1 and θ_2, are the only updated parameters. Therefore, the training loss $\mathcal{L}$ is the weighted sum of the disentanglement loss and the PLM loss:

$$\mathcal{L} = \alpha \mathcal{L}_D + (1 - \alpha)\mathcal{L}_{PLM} \tag{3}$$

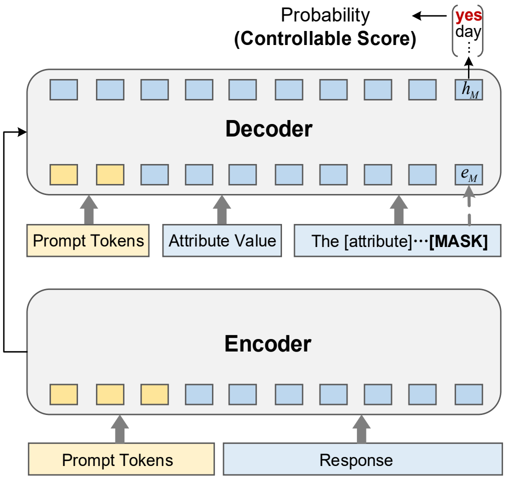


Figure 4: Overview of our evaluation model, MAE.

When the training is completed, we save all parameters of the prompt module. During the inference, the data from the test set is mapped to the representations of prompts only via the embedding matrices, where the features of the attributes seen in the training set can be transferred to the unseen combinations.

5 Method of MAE

To fill the gap in metrics for multi-attribute controllable dialogue generation, we propose a unified and efficient evaluation framework without additional large-scale labeled data, as shown in Figure 4, which converts the evaluation of each attribute to a unified text-to-text generation task, just like Gu et al. (2022). T5 (Raffel et al., 2020) is used as the base model for our work. A template is designed as discrete prompts, i.e., "The emotion/act/persona controls the response [MASK]". To alleviate the potential bias of different handcrafted patterns (Ke et al., 2022), we further add a trainable continuous task-oriented prompt to improve stability and robustness.

Specifically, the continuous prompt sequence is prepended to the response as a prefix, which makes up the input of the encoder. Another continuous prompt sequence, the attribute values, and the template are concatenated and fed to the decoder. We take the probability of generating "yes" corresponding to [MASK] token as the controllability score. In training process, only embeddings of continuous prompts are updated and the parameters of T5 are fixed. Note that our model-based evaluation ap-

[1] We find constructing pseudo combinations with at least one different attribute value is slightly better than with all different attributes in the experiments.

Split	DailyDialog-CG					ConvAI2-CG				
	Size	Turn.num	Att_com.num	Dial.len	Res.len	Size	Turn.num	Att_com.num	Dial.len	Res.len
Train	12,504	6.8	18	77.6	12.9	18,000	5.0	11,566	46.5	11.7
Validation	1,390	6.5	18	75.0	13.0	2,000	5.0	1,883	46.8	11.6
Test	1970	6.0	6	69.6	13.9	2,000	5.0	873	46.1	11.6

Table 1: Statistics of DailyDialog-CG and ConvAI2-CG ("CG" means compositional generalization). "Size" and "Att_com.num"denote the numbers of examples and attribute combinations. "Turn.num" are the average number turns per example. "Dial.len" and "Res.len" are the average lengths of dialogue history and response.

proach gets rid of the reliance on golden response when tested and can be uniformly applied to various granularities of attributes.

6 Experiments

6.1 Datasets

We construct two datasets based on DailyDialog (Li et al., 2017) and ConvAI2 (Dinan et al., 2020) for compositional generalization in multi-attribute controllable dialogue response generation.

DailyDialog-CG DailyDialog is an open-domain dialogue dataset with two controllable attributes: emotion and act. Here, we treat the labels of the two attributes as an attribute combination, e.g., (surprise, inform). For dialogues, each utterance with two attribute labels is regarded as the response and all preceding texts of this utterance are considered as the corresponding dialogue history. In this way, we get 14,879 examples. We count the attribute combinations labeled in all examples, 18 of which are selected as $C_{v,train}$ and the other 6 are $C_{v,test}$. Then, the examples are divided into the training set and test set according to the combination set. We also extract 10% samples from the training set as the validation set.

ConvAI2-CG ConvAI2 is a persona-based dialogue dataset in which the persona profile of each dialogue is consisting of 4 or 5 personalized sentences. We treat each sentence as an attribute value and the sentences in the same position belong to the same attribute. The persona profile is regarded as an attribute combination, e.g., ("My mom is my best friend.", "I've four sisters.", "I believe that mermaids are real.", "I love iced tea."). For each dialogue, we choose the first 4 utterances as the dialogue history and the 5th utterance as the response. Consistent with the processing method of DailyDialog-CG, we select 11,566 combinations as $C_{v,train}$[2] and the other 873 combinations as $C_{v,test}$. After that, we obtain the corresponding training set, validation set, and test set.

The statistics about the two datasets are shown in Table 1.

6.2 Baselines

We compare our methods with several competitive baselines. The common dialogue generation models are included: (1) DialoGPT-Ori (Zhang et al., 2020); (2) FUDGE (Yang and Klein, 2021); (3) PPLM (Dathathri et al., 2019); (4) Cocon (Chan et al., 2020); (5) Fine-tuning; (6) CTRL (Keskar et al., 2019). We also implement some prompt-based methods for comparison: (1) Prompt-tuning (Lester et al., 2021); (2) CatPrompt (Yang et al., 2022). More details can be seen in Appendix A[3].

6.3 Evaluation Metrics

In this work, we focus on evaluating the attribute controllability and text quality for different controllable generation methods.

Attribute Controllability It aims to evaluate whether the method can generate responses constrained by multiple attributes successfully.

1. For the control of coarse-grained discrete attributes in DailyDialog-CG, we use the classification accuracy, i.e., E-ACC and A-ACC, for each attribute computed by an independently trained Roberta classifier (Liu et al., 2019), respectively.

2. For the control of fine-grained continuous attributes in ConvAI2-CG, we calculate the cosine similarity between the representations of attribute sentences and the generated response, i.e., P-SIM(Du and Ji, 2021). We also evaluate the model by measuring the consistency of attribute sentences with the generated response via a Roberta-based Natural Language Inference (NLI) model, i.e., P-NLI(Madotto et al., 2019).

3. We propose a unified model-based evaluation metric, i.e., MAE, for various granularities of

[2] The 1,883 combinations of the validation set are included in the 11,566 combinations of the training set.

[3] Our code, models and other related resources are publicly available at https://github.com/Zeng-WH/Seen-to-Unseen.

Method	Controllability				Text Quality		
	E-ACC ↑	E-MAE ↑	A-ACC ↑	A-MAE ↑	BLEU-1 ↑	BLEU-2 ↑	METEOR ↑
DialoGPT-Ori	50.36	60.46	27.82	31.61	11.53	1.58	9.03
FUDGE	60.10	64.29	27.21	29.21	12.24	1.13	8.67
PPLM	51.57	56.87	33.60	33.71	11.77	1.34	9.26
CoCon	52.79	59.99	29.44	34.51	6.91	0.42	11.50
Fine-tuning	62.74	66.77	35.66	37.02	21.64	10.19	19.15
CTRL	67.34	69.55	33.50	36.15	24.76	11.42	20.45
Prompt-tuning	57.06	62.78	30.36	32.53	19.71	7.36	15.13
CatPrompt	60.91	66.50	36.75	38.43	24.07	11.17	20.72
DCG (ours)	**70.66**	**72.61**	38.98	**41.63**	**26.33**	**14.16**	**24.57**
DCG w/o AOP (Prompt-tuning)	57.06	62.78	30.36	32.53	19.71	7.36	15.13
DCG w/o TOP	66.80	68.02	**41.83**	41.50	19.18	6.74	15.63
DCG w/o DL	60.41	64.57	38.07	39.45	22.45	9.20	19.55

Table 2: The performance of compositional generalization in multi-attribute controllable dialogue generation for DailyDialog-CG. "E" and "A" denote controllable attributes of "Emotion" and "Act". "AOP", "TOP", and "DL" mean attribute-oriented prompt, task-oriented prompt, and disentanglement learning. Results are averaged over three random runs. ↑ means a higher score is better. ($p < 0.01$ under t-test)

Method	Controllability			Text Quality		
	P-SIM ↑	P-NLI ↑	P-MAE↑	BLEU-1↑	BLEU-2↑	METEOR↑
DialoGPT-Ori	60.16	72.47	23.12	12.33	1.54	8.95
PPLM	59.90	75.98	25.03	13.20	1.65	9.06
Fine-tuning	65.48	69.50	19.21	16.53	2.40	10.96
CTRL	65.20	77.65	26.12	18.39	**3.12**	12.23
Prompt-tuning	64.84	74.30	24.56	17.59	2.60	11.22
DCG (ours)	**69.03**	**81.20**	**30.42**	**19.55**	2.68	**12.42**
DCG w/o AOP (Prompt-tuning)	64.84	74.30	24.56	17.59	2.60	11.22
DCG w/o TOP	67.35	78.50	28.44	12.18	1.05	7.61
DCG w/o DL	68.25	79.00	28.53	18.34	2.39	11.63

Table 3: The performance of compositional generalization in multi-attribute controllable dialogue generation for ConvAI2-CG. "P" denotes controllable attribute of "Persona". Results are averaged over three random runs. ↑ means a higher score is better. ($p < 0.01$ under t-test)

attributes, the details can be seen in Section 5.

Text Quality We use the BLEUs (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) to measure the match scores between generated responses and ground-truth references.

6.4 Main Results

Results on DailyDialog-CG Table 2 presents the results of controllable dialogue generation about unseen attribute combinations for DailyDialog-CG. [4] We conduct experiments based on some strong controllable dialogue generation models and novel prompt-based methods. In general, our DCG outperforms all other baselines in terms of attribute controllability and text quality. Compared to CTRL, our model improves by 1.6%, 2.7%, 4.1% in BLEU-1, BLEU-2, METEOR for text quality, and 3.3%, 3.1%, 5.5%, 5.5% in E-ACC, E-MAE, A-ACC, A-MAE for attribute controllability. We also find the FUDGE and PPLM, two methods based on the decoding strategy, perform poorly here, especially in text quality, which illustrates the incompatibility of these decoding strategies for combinatorial generalization. Besides, as observed, Catprompt is a relatively well-performing prompt-based baseline, but it is still far worse than our method. This is because it directly concatenates all trained single-attribute prompts as the multi-attribute prompt for test. This inconsistency between training and testing stages decreases the performance. Different from these methods, our method optimizes the language modeling loss only based on discrete prompts for attribute combination and continuous task-oriented prompt, which can focus on the features of multiple attributes at the same time also during the training and achieve a better transfer via a learnable mapping.

Besides, we also concern whether DCG benefits from attribute-oriented prompt, task-oriented prompt, and disentanglement learning. We find that DCG w/o AOP is the same with Prompt-tuning and it performs poorly in attribute controllability, which shows attribute-oriented prompt plays an important role in guiding the model to focus on the controlled information. After removing the task-oriented prompt, the DCG w/o TOP decreases to

[4] Our DCG improves text quality and controllability. The BLEUs seem low because we adopt the same calculation as ParlAI (Miller et al., 2017), which is lower than results in (Li et al., 2017) for different smooth functions.

19.18%, 6.74%, and 15.63% on text quality, but still maintains high controllability. It proves task-oriented prompt helps improve text quality. We also conduct experiments to prove that TOP can improve text quality when combined with other methods. (See Appendix H). Besides, after removing disentanglement learning, the DCG w/o DL drops significantly, which shows disentanglement learning effectively disentangles attribute combinations and improves the ability of compositional generalization.

Results on ConvAI2-CG Table 3 presents the results of generalization on unseen attribute combinations for ConvAI2-CG. Due to the diversity of attribute values and attribute combinations, it is very difficult to implement CatPrompt in ConvAI2-CG. Therefore, we remove this baseline. We also remove FUDGE and Cocon for their poor generation quality and slow decoding speed, which is shown in Table 2 and Table 5. We can observe that the trend of overall performance is consistent with that of DailyDialog-CG. Compared to CTRL, our model achieves a great improvement in attribute controllability and text quality, which proves the generality of our methods on the coarse-grained discrete attribute control and fine-grained continuous attribute control. It also shows the effectiveness of our method when more attributes are combined. However, all BLEU scores are low, which is because the ConvAI2-CG has more diverse and complex attribute combinations and leads to the instability of models facing new attribute combinations. Generally, the results show that the compositional generalization for multi-attribute controllable dialogue generation is necessary and meaningful. Noted that we also conduct experiments on the setting with changed number of attributes from training to inference (See in Appendix G).

7 Qualitative Analysis

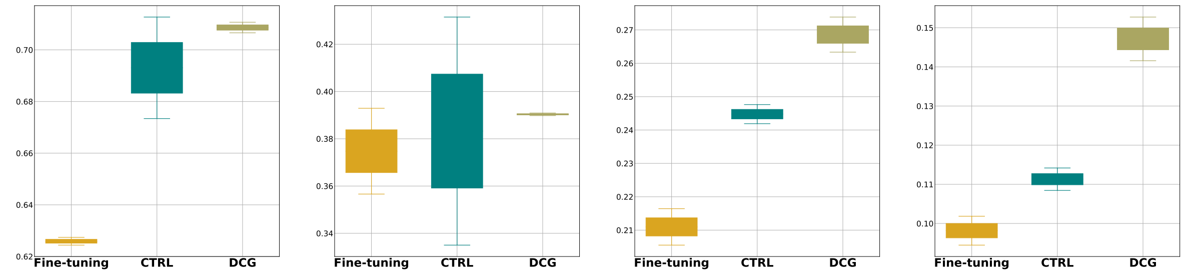
7.1 Comparison between Seen and Unseen Attribute Values

Figure 5 displays the comparison of the performance on seen and unseen attribute combinations for DailyDialog-CG. We report the controllability metrics, E-ACC (emotion) and A-ACC (act), and the BLEUs of the Fine-tuning, CTRL, and our DCG. The top of each box denotes the result of seen attribute combinations and the bottom represents unseen attribute combinations. We find all methods achieve significantly superior performance on seen attribute combinations than on unseen combinations. For example, CTRL achieves 71.27% E-ACC and 43.15% A-ACC on seen attribute combinations but drops to 67.34%(-3.93) and 33.50%(-9.65) on unseen combinations. It strongly proves previous methods suffer from the difficulty of compositional generalization for the multi-attribute controllable dialogue generation. However, we find our proposed DCG can greatly alleviate this gap. The DCG has a smaller drop of 0.41% and 0.11% for E-ACC and A-ACC, and it also outperforms CTRL on both controllability and text equality of unseen attribute combinations. The results confirm the effectiveness of our method for transferring seen attributes to unseen combinations. We find CTRL achieves a higher A-ACC on seen combinations but a lower score on unseen combinations than Fine-tuning, which demonstrates directly adding control codes may cause overfitting to seen attribute combinations.

7.2 Correlation Results on Metrics

Following Guan and Huang (2020), we adopt Pearson (r), Spearman (ρ), and Kendall (τ) correlation coefficients between our proposed automatic metric, MAE, and human judgments (details can be seen in Appendix D) to measure the quality of different metrics. Table 4 shows the overall results on the controllability of coarse-grained discrete attributes, emotion and act, and the fine-grained continuous attributes, persona description. We can observe that our MAE outperforms classic metrics, E-ACC, A-ACC, P-SIM, and P-NLI, by a large margin, indicating the effectiveness of our unified metric on different granularities. We also conducted experiments on some variants of MAE. After the removal of continuous prompts, the correlation scores decrease. It is because the task-oriented prompts are the only parameters can be fine-tuned, which is important for MAE. We also implement MAE on another PLM, BART, to demonstrate generality for our model.

Robustness Analysis To verify the effect of the bias of the handcrafted template, we design another two templates. The Template 1 is "The response is related to the emotion/act/persona [MASK]" and Template 2 is "The response is about the emotion/act/persona [MASK]". As shown in Table 4, MAE (T1) and MAE (T2) achieve similar correlation results (within 0.50%) while the results of MAE w/o Prompt (T1) and MAE w/o Prompt (T2)



(a) E-ACC (b) A-ACC (c) BLEU-1 (d) BLEU-2

Figure 5: Comparision of performance for Fine-tuning, CTRL, and DCG on seen and unseen multi-attribute combinations for DailyDialog-CG in terms of E-ACC, A-ACC, BLEU-1, and BLEU-2.

Metrics	DailyDialog-CG						ConvAI2-CG		
	Emotion			Act			Persona		
	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
ACC	0.5242	0.4936	0.4834	0.3852	0.4077	**0.4027**	\	\	\
P-SIM	\	\	\	\	\	\	-0.0683	0.0065	0.0098
P-NLI	\	\	\	\	\	\	-0.0881	-0.0741	-0.0706
MAE	0.6821	**0.7500**	**0.6242**	0.5446	**0.4661**	0.3936	**0.5793**	0.5768	0.4418
MAE w/o Prompt	0.3665	0.4802	0.3857	-0.2832	-0.2136	-0.1789	-0.0529	0.2591	0.2062
MAE (BART)	**0.6829**	0.7396	0.6102	**0.5478**	0.4358	0.3697	0.5550	**0.5848**	**0.4517**
MAE (T1)	0.6801	0.7661	0.6382	0.5557	0.4661	0.3935	0.6037	0.6235	0.4811
MAE (T2)	0.6758	0.7070	0.5851	0.5357	0.4055	0.3458	0.5724	0.5767	0.4418
MAE w/o Prompt (T1)	0.1158	0.1053	0.0912	-0.3035	-0.2684	-0.2266	0.0835	0.0984	0.0884
MAE w/o Prompt (T2)	0.0417	-0.0257	-0.0210	-0.2680	-0.1040	-0.0835	-0.0512	-0.0199	-0.0295

Table 4: Pearson (r), Spearman (ρ), and Kendall (τ) correlations of attribute controllability evaluation metrics on DailyDialog-CG and ConvAI2-CG. "T1" and "T2" denote the Template 1 and Template 2.

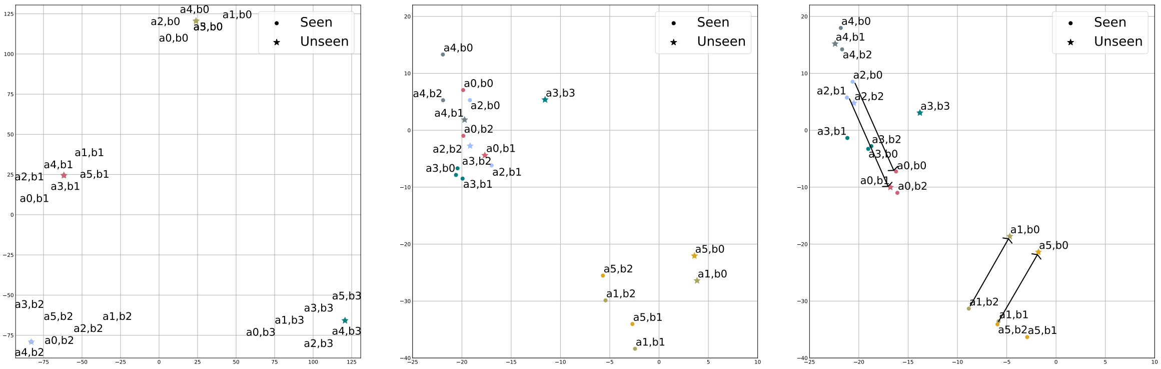
are quite different. It suggests the trainable continuous task-oriented prompt can alleviate the potential bias of different handcrafted templates and further improve the robustness of MAE.

7.3 Prompt Visualization

To show the effect of prompts for compositional generalization, we display a visualization of the concatenated prompt embeddings of two attributes via PCA (Jolliffe and Cadima, 2016) on DailyDialog-CG in Figure 6. For CatPrompt in Figure 6(a), all the multi-attribute combinations $(6(emotion) \times 4(act) = 24)$ almost collapse into four dots where each dot is of the same act attribute value but of different emotion values. We find directly concatenating two single-attribute prompts makes the model only focus on the latter attribute (act), i.e., position sensitive, so that the CatPrompt cannot distinguish different combinations with the other attribute (emotion). Therefore, it's hard for CatPrompt to learn multi-attribute compositional generalization. In Figure 6(b), We find that DCG w/o DL can distinguish different multi-attribute combinations to some extent. However, the combinations of different attribute values are tightly entangled, such as (a0, b2) and (a4, b1). Figure 6(c) shows that our DCG has a close distribution with prompts of the same attribute value, i.e., (a0, b0), (a0, b1), (a0, b2), and a sparse distribution with prompts of different attribute values, e.g., (a0, b2) and (a4, b1). It proves our DCG can disentangle attribute combinations and learn relations between different attributes. Furthermore, DCG learns generalization capability from seen attributes to unseen combinations. For example, (a2, b1) -> (a0, b1) (unseen path) is equal to (a2, b0) -> (a0, b0) (seen path). The results confirm that our proposed attribute-oriented prompt outperforms the models that learn an independent prompt for each attribute value. The shared embedding mapping helps learn attribute concepts from seen values to unseen combinations.

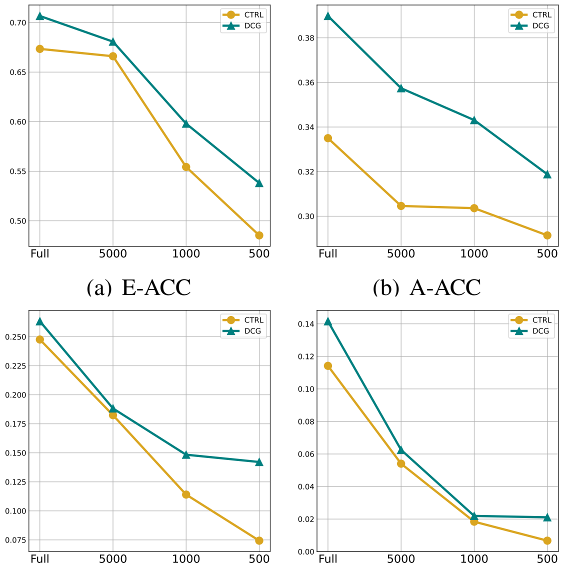
7.4 Few-shot Learning

To study the effect of few-shot learning, we randomly select a ratio of original training data from DailyDialog-CG to train CTRL or DCG in low-resource settings and evaluate the model performance on the original test set. "Full" denotes the same setting as the main results. 5000, 1000, and 500 denote the number of examples chosen from the original training data respectively. The results are shown in Figure 7. Note that we keep the original test set fixed for a fair comparison. As the



(a) CatPrompt (b) CtrlPrompt (c) DCG

Figure 6: Visualization of prompts from different models on DailyDialog-CG. Each dot denotes the prompt embeddings of a multi-attribute combination $(a*, b*)|a* \in A, b* \in B$, where A is the attribute *Emotion* and B is the attribute *Act*. The same color represents that two dots have the same value of *Emotion* and different shaped dots represent seen/unseen combinations. For clarity, we leave out some outliers.



(a) E-ACC (b) A-ACC

(c) BLEU-1 (d) BLEU-2

Figure 7: Few-shot learning of our DCG and CTRL on DailyDialog-CG.

size of training data decreases, the performance of both CTRL and DCG presents a dropping trend and our DCG model is consistently better than CTRL, which confirms our model has a strong capability for multi-attribute controllable dialogue generation.

8 Case Study

Figure 9 (See in Appendix) shows two examples from Dailydialog-CG and ConvAI2-CG, respectively. For example one in the DailyDialog-CG, the CTRL generates the word "great", showing that the generated response is emotionally controllable. However, both sentences in the response are declarative sentences, which does not control the act *question*. As observed, the response generated by our DCG contains the word "Wow", which strongly expresses the emotion of *happiness*. Besides, a question sentence is also generated. Example two in ConvAI2-CG needs to control 5 attributes, of which the golden response contains 2 attributes. The CTRL only controls "like to skate", while our DCG controls "like to write poetry and skate", which is highly consistent with the golden response. Compared with previous models, our model addresses many difficult issues in compositional generalization for multi-attribute controllable dialogue generation. With an attribute-oriented prompt and a task-oriented prompt, our method learns attribute concepts from seen attribute values to unseen attribute combinations. Through a disentanglement learning, some artificial-constructed unseen pseudo combinations are injected into the training process, which greatly improves the generalization ability of our model.

9 Conclusion

In this paper, we study the compositional generalization for multi-attribute controllable dialogue generation. We propose a prompt-based disentangled controllable dialogue generation model which generates attribute-specific prompt vectors from control codes and uses a disentanglement loss to disentangle different attributes. Further, we develop a unified reference-free evaluation framework, MAE, for multi-attribute generation with different levels of granularities. Experiments and analysis show our method achieves better text quality and controllability scores. Moreover, our proposed MAE has a higher correlation with human judgments for evaluation on CDG.

Acknowledgements

We thank all anonymous reviewers for their helpful comments and suggestions. We are also grateful to the track organizers for their valuable work. This work was partially supported by National Key R&D Program of China No. 2019YFF0303300 and Subject II No. 2019YFF0303302, DOCOMO Beijing Communications Laboratories Co., Ltd, MoE-CMCC "Artifical Intelligence" Project No. MCM20190701. Jingang Wang is funded by Beijing Nova Program(Grant NO. 20220484098)

Limitations

Although DCG achieves significant improvements compared with existing baselines, there are still avenues to be explored in future research. (1) DCG in this paper focuses on the compositional generalization for multi-attribute on controllable dialogue generation. We hope to extend the method to other generative tasks, including but not limited to dialogue summarization and story generation. (2) In this paper, we explored the control of coarse-grained discrete attributes and the control of fine-grained ones separately, and we intend to study the combination of these two attributes in future research.

Ethics Statement

Controllable dialogue generation(CDG) is an essential task in Natural Language Processing (NLP) and has been widely studied for decades, which aims to guide dialogue generation toward the desired attributes such as emotions, acts, and personas. In the open-domain dialogue scenario, CDG can generate emotional and diverse responses to enhance the user's sense of participation. In the task-oriented dialogue scenario, CDG can generate responses that meet the user's needs according to the user's intent. However, most previous works focus on single-attribute generation where there is only one attribute label like *happiness* in emotion and pay less attention to the multi-attribute generation, which is a more practical setting. Different from single-attribute, the control signal of the multi-attribute generation is a combination of multiple values from different attributes, which faces the challenge of lacking sufficient annotated attribute-specific data. Therefore, we explore the compositional generalization for multi-attribute controllable dialogue generation where a model could learn from seen attribute values and generalize to unseen combinations. We also design a novel and general reference-free evaluation framework to unify the evaluation of different granularity attributes. The experimental results prove the effectiveness of our model and evaluation framework. Besides, there is no huge biased content in the datasets and the models. If the knowledge base is further used, the biased content will be brought into the generated responses, just like biased content posted by content creators on the Web which is promoted by a search engine. To prevent the technology from being abused for disinformation, we look forward to more research effort being paid to fake/biased/offensive content detection and encourage developers to carefully choose the proper dataset and content to build the knowledge base.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, Jennifer C Lai, and Robert L Mercer. 1992. An estimate of an upper bound for the entropy of english. *Computational Linguistics*, 18(1):31–40.

Alvin Chan, Yew-Soon Ong, Bill Pung, Aston Zhang, and Jie Fu. 2020. Cocon: A self-supervised approach for controlled text generation. *arXiv preprint arXiv:2006.03535*.

Jordan Clive, Kris Cao, and Marek Rei. 2022. Control prefixes for parameter-efficient text generation. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 363–382.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan

Lowe, et al. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition: From Machine Learning to Intelligent Conversations*, pages 187–208. Springer.

Wanyu Du and Yangfeng Ji. 2021. SideControl: Controlled open-domain dialogue generation via additive side networks. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2175–2194, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. 2022. PPT: Pre-trained prompt tuning for few-shot learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8410–8423, Dublin, Ireland. Association for Computational Linguistics.

Jian Guan and Minlie Huang. 2020. UNION: An Unreferenced Metric for Evaluating Open-ended Story Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9157–9166, Online. Association for Computational Linguistics.

Jonathan Herzig and Jonathan Berant. 2021. Span-based semantic parsing for compositional generalization. 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 908–921, Online. Association for Computational Linguistics.

Ian T Jolliffe and Jorge Cadima. 2016. Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202.

Pei Ke, Hao Zhou, Yankai Lin, Peng Li, Jie Zhou, Xiaoyan Zhu, and Minlie Huang. 2022. CTRLEval: An unsupervised reference-free metric for evaluating controlled text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2306–2319, Dublin, Ireland. Association for Computational Linguistics.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. GeDi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. 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 4582–4597, Online. Association for Computational Linguistics.

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.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zhiyu Lin and Mark Riedl. 2021. Plug-and-blend: A framework for controllable story generation with blended control codes. *arXiv preprint arXiv:2104.04039*.

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

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Yiwei Lyu, Paul Pu Liang, Hai Pham, Eduard Hovy, Barnabás Póczos, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2021. Styleptb: A compositional benchmark for fine-grained controllable text style transfer. *arXiv preprint arXiv:2104.05196*.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5459.

Sanket Vaibhav Mehta, Jinfeng Rao, Yi Tay, Mihir Kale, Ankur Parikh, and Emma Strubell. 2022. Improving compositional generalization with self-training for data-to-text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4205–4219, Dublin, Ireland. Association for Computational Linguistics.

Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. ParlAI: A dialog research software platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84, Copenhagen, Denmark. Association for Computational Linguistics.

Inbar Oren, Jonathan Herzig, Nitish Gupta, Matt Gardner, and Jonathan Berant. 2020. Improving compositional generalization in semantic parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2482–2495, Online. Association for Computational Linguistics.

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.

Jing Qian, Li Dong, Yelong Shen, Furu Wei, and Weizhu Chen. 2022. Controllable natural language generation with contrastive prefixes. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2912–2924, Dublin, Ireland. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Kevin Yang and Dan Klein. 2021. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535, Online. Association for Computational Linguistics.

Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Mingfeng Xue, Boxing Chen, and Jun Xie. 2022. Tailor: A prompt-based approach to attribute-based controlled text generation. *arXiv preprint arXiv:2204.13362*.

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

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. 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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

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

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Hao Zheng and Mirella Lapata. 2022. Disentangled sequence to sequence learning for compositional generalization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4256–4268, Dublin, Ireland. Association for Computational Linguistics.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

A Baselines

DialoGPT-Ori: Proposed by (Zhang et al., 2020), this model is a dialogue generative pre-trained transformer. Here, we use the original DialoGPT for open-domain dialogue generation. DialoGPT is the backbone for all other baselines except CoCon.
Fine-tuning: We use dialogue history in datasets to fine-tune the DialoGPT for dialogue generation.
CTRL: Proposed by (Keskar et al., 2019), this method provides attribute control codes for a language model trained from scratch. We concatenate

multi-attribute control codes with dialogue history to fine-tune the DialoGPT.

CoCon: Proposed by (Chan et al., 2020), this method uses a content input to control an GPT's output text at a fine-grained level.

PPLM: Proposed by (Dathathri et al., 2019), this method is a gradient-based baseline that uses a plug-and-play language model(PPLM) to guide the language model. We train a joint classifier of emotion and dialogue act which takes a single response as input and predicts the attribute combination of the emotion and dialogue act on DailyDialog-CG. Noted that the attribute classifiers of PPLM can not directly generalize to unknown attribute combinations, so we use both training data and test data to train the attribute classifiers. We use the bag-of-words attribute model which encodes persona profile to control the DialoGPT on ConvAI2-CG.

FUDGE: Proposed by (Yang and Klein, 2021), this method is a weighted decoding baseline which uses a future discriminator for generation(FUDGE) to guide the DialoGPT. We train a joint discriminator that takes the dialogue history and the current response as input and predicts the attribute combination of emotion and dialogue act on DailyDialog-CG.

Prompt-tuning: Proposed by (Lester et al., 2021), this method uses continue prompts to fine-tune language models. We apply this method to the DialoGPT for dialogue generation.

CatPrompt: Inspired by Yang et al. (2022); Qian et al. (2022), we initialize an unique prompt for each single attribute value and concatenate single-attribute prompts as the multi-attribute prompts. We fine-tune multi-attribute prompts for dialogue generation. Note that CatPrompt is only applied to coarse-grained discrete attributes like emotion and act instead of persona. Because persona has a large value set, resulting in numerous parameters (see Table 6).

B Implementation Details

Our implementation is based on the Hugging Face Transformer models[5]. DialoGPT$_{\text{Small}}$ is used as a backbone and the input sequence length is truncated to 512 tokens. Following the HuggingFace default setup, we use an AdamW optimizer (Loshchilov and Hutter, 2017) and a linear learning rate scheduler with an initial rate of $7.5 \cdot 10^{-5}$, and the batch size is set to 8. The prompt lengths are set to 50 and 150, the attribute-oriented prompt lengths are set to 6 and 100, the disentanglement loss weight is set to 0.1 and 0.03, and the number of Pseudo Combinations is set to 8 and 6 for DailyDialog-CG and ConvAI2-CG, respectively. Our model is trained on Tesla V100 machines, taking 24 minutes per epoch on DailyDialog-CG and 36 minutes per epoch on ConvAI2-CG. For all experiments, we set the number of training epochs to 30. At the decoding phase, we use a greedy search and max generated tokens of 150.

Method	Decoding Speed ↑
DialoGPT-Ori	1.1837x
FUDGE	0.0041x
PPLM	0.0006x
CoCon	0.0044x
Fine-tuning	1.1347x
CTRL	1.1673x
Prompt-tuning	1.0000x
CatPrompt	1.0408x
DCG (ours)	1.0490x
DCG w/o DL	1.0122x

Table 5: The decoding speed of different models, which takes the decoding speed of the model relative to the Prompt as a metric.

C Inference Efficiency

We compare the average inference efficiency of our methods with the baselines. As we can observe from Table 5, the inference speed of PPLM, FUDGE, and CoCon is far slower than the original GPT-2 model. Prompt-based methods are much faster than that decoding strategy based methods. The inference speed of our method is close to the original DialoGPT methods. As shown in Table 6, with the growth of attribute combinations, the trainable parameters of CatPrompt increase rapidly, from 0.84M to 224M, which even exceeds the 117M trainable parameters of full DialoGPT. While our method achieves better results with a lower number of trainable parameters on DialyDialog-CG and ConvAI2-CG.

D Human Evaluation

To validate the good performance of DCG, we further deploy a set of human evaluations to compare the controllability and text quality between several methods. We randomly sample 100 examples from two datasets and collect the corresponding generated responses of CTRL, DCG, and DCG w/o DL. For the controllability, 5 human annotators are invited to evaluate on a scale of 1-3, where score 1

[5] https://github.com/huggingface/transformers

Model	DailyDialog-CG		ConvAI2-CG	
	Traninable Parameters	Percent Trainable	Traninable Parameters	Percent Trainable
Fine-tuning	117M	100%	117M	100%
CTRL	117M	100%	117M	100%
Prompt-tuning	0.13M	0.11%	0.21M	0.18%
CatPrompt	0.84M	0.71%	244M	205%
DCG (ours)	0.66M	0.56%	0.66M	0.56%
DCG w/o DL	0.66M	0.56%	0.66M	0.56%

Table 6: Number of parameters used for different models. Trainable parameters is the number of parameter used for training in models. Percent Trainable is the ratio of trainable parameters to original GPT-2.

Model	DailyDialog-CG				ConvAI2-CG		
	Controllability		Text Quality		Controllability	Text Quality	
	Emo.	Act.	Flu.	Rel.	Per.	Flu.	Rel.
CTRL	2.20	2.05	4.19	3.35	1.70	4.02	3.25
DCG	2.35	2.85	4.42	3.89	2.17	4.03	3.26
DCG w/o DL	1.70	2.30	4.04	3.18	1.61	4.07	3.22

Table 7: Human evaluation on controllability and text quality for DailyDialog-CG and ConvAI2-CG. Emo., Act., and Per. are the attributes of emotion, act, and persona. Flu. and Rel. are the fluency and context relevancy.

means that the generated response is completely inconsistent with the expected attribute label, score 2 denotes that the generated response has the same meaning as the expected attribute label, but no explicit attribute-related words, and score 3 means that the generated response contains some clear attribute words. For the text quality, we ask the annotators to evaluate the fluency and context relevancy of the generated responses on a scale of 1-5, where a higher score indicates better quality. The inter-annotator agreement on the controllability and text quality is 0.63 and 0.61 for DailyDialog-GC, and 0.58 and 0.60 for ConvAI2-CG. For all metrics, the average score of the 5 annotators is treated as the final score.

As shown in Table 7, the text quality scores of all models are high, which is because the models finetuned on contextualized language backbones can generate fluent sentences with relevant information. For controllability, our DCG achieves better performance than CTRL both on the coarse-grained discrete attributes and fine-grained continuous attributes, which suggests that our shared prompt mapping can learn the attribute concepts from seen attribute values to unseen attribute combinations and is useful for diverse attributes. Besides, when removing the disentanglement learning, the scores of our DCG w/o DL drop significantly, which further shows the effectiveness of the combination disentanglement to improve the generation ability.

E Effect of Model Parameters

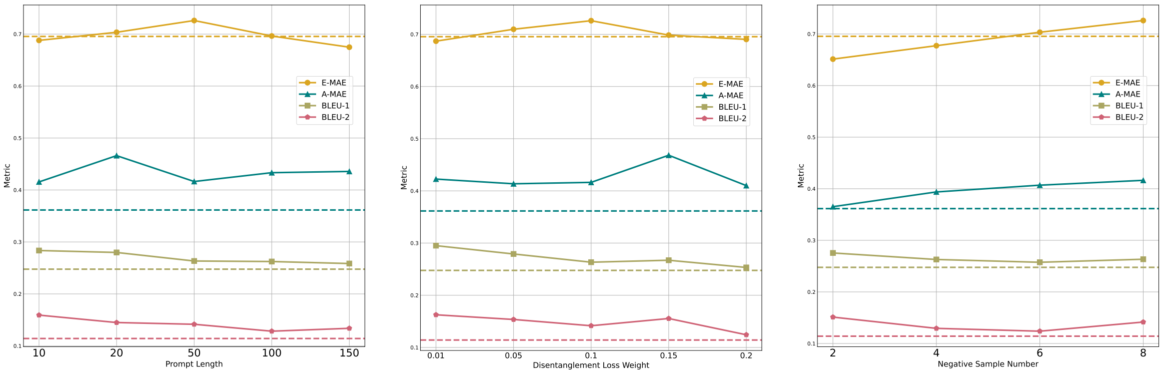
Prompt Length Figure 8 (a) displays the effect of overall prompt lengths of E_p. Since the length of attribute-oriented prompt is fixed to the number of control code, we change the length of the task-oriented prompt. We find that our DCG achieves superior performance when the prompt length is between 20 and 100, and gets the best scores when the prompt length is 50. The DCG outperforms the strong baseline CTRL by the 3.19% (averaged) for MAE and 2.16% (averaged) for BLEUs but uses only 56% trainable parameters of CTRL, which verifies the effectiveness and robustness of our method.

Weight of Disentanglement Loss Figure 8 (b) shows the effect of different weight ratios α for the disentanglement loss $\mathcal{L}_D$. We observe that $\alpha \in (0.05, 0.15)$ achieves consistent improvements than CTRL and we take $\alpha = 0.10$ in all experiments.

Number of Pseudo Combinations Figure 8 (c) shows the effect of the number of pseudo combinations in the disentanglement loss. We find a larger number will improve the controllability of our model. It's because more pseudo attribute values help the model to separate the desired attribute combination from the others.

F Comparison with CTRLEval

Automatic evaluation metrics are important for text generation tasks, including reference-based like BLEU (Papineni et al., 2002), ROUGE (Lin, 2004),



(a) Prompt Length

(b) Disentanglement Loss Weight

(c) Number of Pseudo Combinations

Figure 8: Effect of prompt length, disentanglement loss weight, and number of pseudo combinations for DailyDialog-CG. The dotted lines denote the performance of CTRL. We report the MAE and BLEU scores for all settings.

Metrics	**DailyDialog-CG**						**ConvAI2-CG**		
	Emotion			**Act**			**Persona**		
	Pearson	**Spearman**	**Kendall**	**Pearson**	**Spearman**	**Kendall**	**Pearson**	**Spearman**	**Kendall**
CTRLEval	**0.6927**	0.6994	0.5961	0.1232	0.3391	0.2743	0.4059	0.3622	0.2847
MAE	0.6821	**0.7500**	**0.6242**	**0.5446**	**0.4661**	**0.3936**	**0.5793**	**0.5768**	**0.4418**

Table 8: Pearson (r), Spearman (ρ), and Kendall (τ) correlations of attribute controllability in DailyDialog-CG and ConvAI2-CG. We use the attribute relevance of CTRLEval as the controllability score.

BERTScore (Zhang et al., 2019) and unreferenced like perplexity (Brown et al., 1992), discriminator scores (Dathathri et al., 2019), BARTScore (Yuan et al., 2021). To evaluate controllability, (Dathathri et al., 2019; Yang and Klein, 2021) trained an attribute classifier to predict the probability using labeled external data, which is hard to multi-attribute controllable generation. As a concurrent work, CTRLEval (Ke et al., 2022) proposes an evaluation method for controllable text generation. Different from our MAE, CTRLEval uses handcrafted prompts to evaluate attribute relevance. However, handcrafted prompts are hard to construct for new tasks and cause generation bias. In contrast, our MAE uses a learnable soft prompt based on PLMs to enhance the generalization capability and robustness. We also provide a performance comparison in Table 8. Results show our MAE shows superior correlations of attribute controllability.

G Performance on Number of Attribute

To prove our model still be useful when the number of attributes varies from training to inference, we train CTRL and our DCG with 4 attributes and inference with 5 attributes in ConvAI2-CG. As shown in Table 9, DCG outperforms the strong baseline CTRL by 3.54% , 5.99%, 4.8% in P-SIM, P-NLI and P-MAE on controllability and achieves comparable BLEU scores. It proves DCG can also handle well with changed number of attributes.

H Impact of TOP on Text Quality

We prove that task-oriented prompts (TOP) can also improve text quality when combined with other methods. Specifically, we trained CTRL with TOP in our experiments. As Table 10 shows, the results of CTRL for BLEU-1, BLEU-2, and METEOR are 24.76%, 11.42%, and 20.45%, respectively. Meanwhile, the results of CTRL+TOP for BLEU-1, BLEU-2, and METEOR are 25.88%, 14.36%, and 21.82%. These results indicate that CTRL can utilize TOP to enhance text quality.

Example one in DailyDialog	**Example two in Convai2**
Dialogue History: A: Smells good! What's for breakfast? B: Well, since we are getting up so late, I decided to make a big breakfast!" A: Nice! Brunch!" B: Kind of, I made scrambled and soft boiled eggs, some French toast and buttermilk pancakes!	**Dialogue History:** A: Hi! how are you doing today? B: I am doing good and yourself? A: I am doing alright. What are some of your hobbies? B: I love the outdoors and camping"
Controllable Attribute Combination: (happiness, question)	**Controllable Attribute Combination:** I sometimes write poetry. My favorite color is black. I sing in a punk band. I like to skate. I want to visit the uk.
Golden Response Wow! You really went all out! Did you make any coffee?	**Golden Response** That sounds fun. I like to write poetry and to go ice stating haha.
CTRL That sounds like a great breakfast! I'll make a big omelet, toast, and buttermilk pancakes!	**CTRL** Just like to skate and ski.
DCG (ours) Wow! You really went all out with it! Did you make any coffee?	**DCG (ours)** That is cool. I like to write poetry and skate.

Figure 9: Case study for two examples from DailyDialog-CG and ConvAI2-CG. We present the dialogue history, its corresponding controllable attribute combination, golden response, CTRL prediction, and prediction of our DCG.

Model	**Controllability**			**Text Quality**		
	P-SIM ↑	**P-NLI** ↑	**P-MAE** ↑	**BLEU-1** ↑	**BLEU-2** ↑	**METEOR** ↑
CTRL	67.09	77.21	26.38	19.44	3.20	12.51
DCG	70.63	83.20	31.18	18.63	2.32	11.87

Table 9: The performance of CTRL and DCG for ConvAI2-CG when the number of attributes varies. We train models with 4 attributes and inference with 5 attributes. Results are averaged over three random runs. ↑ means a higher score is better. ($p < 0.01$ under t-test)

Model	**BLEU-1** ↑	**BLEU-2** ↑	**METEOR** ↑
CTRL	24.76	11.42	20.45
CTRL+TOP	25.88	**14.36**	21.82
DCG	**26.33**	14.16	**24.57**

Table 10: The performance of CTRL , CTRL+TOP and DCG for DailyDialog-CG. Results are averaged over three random runs. ↑ means a higher score is better. ($p < 0.01$ under t-test)

ACL 2023 Responsible NLP Checklist

A For every submission:

□ A1. Did you describe the limitations of your work?
Left blank.

□ A2. Did you discuss any potential risks of your work?
Left blank.

□ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

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

B □ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
Left blank.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Left blank.

□ 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)?
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?
Left blank.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
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.
Left blank.

C □ Did you run computational experiments?

Left blank.

□ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Left blank.

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?
Left blank.

□ 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?
Left blank.

□ 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.)?
Left blank.

D □ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ 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.?
Left blank.

□ 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)?
Left blank.

□ 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?
Left blank.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Left blank.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Left blank.