A ETHICAL CONSIDERATIONS

In this work, we use the publicly released datasets to train/valid/test our models. Generally, these previous works have considered the ethical issues when creating the datasets. For the datasets we used in this work, we have manually checked some samples, and do not find any obvious ethical concerns, such as violent or offensive content. We will also release the source code and the well-trained models along with friendly instructions to support its correct use. However, we still need to emphasize that the text generation is not as controllable as we think. It still would generate some novel or unexpected words occasionally. We may take the actions to decrease the generation diversity to alleviate this problem.

B HUMAN EVALUATION INSTRUCTIONS

Please rate the quality of the generated response based on the given dialogue context and the target response over following aspects: (1) Fluency; (2) Informativeness; (3) Coherence; (4)Semantic Coverage. We provide some instructions for your rating.

B.1 FLUENCY

This measures whether the generated text has no formatting problems, capitalization errors or obviously ungrammatical sentences (e.g., fragments, missing components) that make the text difficult to read. The definitions of different scores are:

- **5**: The text is fluent, grammatically correct, and has no errors. It is easy to read.
- **4**: The text is grammatically correct, but has a few spelling or capitalization errors, which does not affect your understanding.
- **3**: The text has minor errors in both grammar and spelling. The errors slightly affect your understanding.
- **2**: The text has major errors in both grammar and spelling. The errors make the text hard to read.
- **1**: The text does not make sense and it is unreadable.

B.2 INFORMATIVENESS

This measures whether the generated text has diverse, informative, novel or logically related contents. The definitions of different scores are:

- **5**: The text contains very diverse, informative and novel contents. It is enjoyable to read the text.
- **4**: The text contains many informative and novel contents. (Choose this score when you hesitate between 3 and 5.)
- **3**: The text contains some new information but also contains a few repetitions of the context.
- **2**: The text only contains a few informative and new terms. (Choose this score when you hesitate between 1 and 3.)
- **1**: The text is dull, repetitive and has no new information. All contents are from the dialogue context.

B.3 COHERENCE

This measures whether the generated text is semantically and factually consistent with the dialogue context. The definitions of different scores are:

- **5**: The text is semantically, factually and topically consistent with the dialogue context. All contents of the text are related to the source text or can be inferred from the source.

- **4**: The text is very related to the context but has minor inconsistency or contradictions that do not affect its overall relevance.
- **3**: The text is related to the context but has some obvious inconsistency and contradictions.
- **2**: The text is slightly consistent with the context. Many inconsistency and contradictions to the context can be found.
- **1**: The text is totally inconsistent with the context. It is semantically or factually contradicted to the context.

B.4 SEMANTIC COVERAGE

This measures how many semantic content units from the target response are covered by the generated text. The definitions of different scores are:

- **5**: All semantic content units of the target text can be found from the generated text. They are semantically consistent.
- **4**: Most of the content units of the target text can be found from the generated text while a few missing units do not affect the overall coverage.
- **3**: Some semantic content units can be found from the generated text but also miss some important units.
- **2**: Most of semantic content units are not covered. Only a few insignificant units can be found in the generated text.
- **1**: The text does not have any overlapping semantic content units with the target text.

We recruit five human workers to annotate 24,000 samples. To make sure the workers are fairly paid, we pay 0.1 dollars for each sample. Therefore, the total amount spent on participant compensation is 2,400 dollars. The annotators take 48 hours to finish the task, suggesting the hourly wage for each worker is 10 dollars.

C MORE DETAILS OF THE TASKS

C.1 MULTI-TURN DIALOGUE RESPONSE GENERATION

Training We fine-tune the models on the DailyDialog and LCCC datasets for 8k and 40k steps, respectively. We use a batch size of 128 and truncate the training samples to a maximum length of 256. The parameters of the models are initialized from Huggingface Libraries (Wolf et al., 2019a) and updated by Adam optimizer (Kingma & Ba, 2015) with a learning rate of 3e-5. The margin values of SimCTG and SimDRC are set to 0.5 and 0.7, respectively. The loss weight α is 0.3. All hyper-parameters are selected from the development set. The training process on the DailyDialog and LCCC datasets takes 0.7 hours and 4 hours on four A100 GPUs, respectively.

Evaluation We conduct automatic evaluations for the response generation task on following metrics: 1) **BERTScore** (Zhang et al., 2019) which calculates the similarities of token representations between the generated response and the target response using the pre-trained BERT model; 2) **BARTScore** (Yuan et al., 2021) which estimates the difficulties of converting the generated text to the reference output by the text generation method; 3) **BLEURT** (Sellam et al., 2020) which is a reference-based text generation metrics that is robust to both domain and quality drifts; and 4) **Distinct2/4** (Li et al., 2016) which computes the generation repetition at different n-gram levels.

C.2 CONVERSATIONAL RESPONSE RETRIEVAL

Training For each dataset, we fine-tune the model on the training set for 3 epochs, and save the best model according to the performance on the development set. The model parameters are initialized from the post-training checkpoint released by Han et al. (2021) and updated by Adam optimizer with a learning rate of $1e^{-5}$. The margin values in SimCTG and SimDRC are set to 0.5 and 0.8, respectively. The value of α in SimDRC is set to 0.6. All hyper-parameters are selected on the development set. The training process on each dataset takes around 10 hours on a single A100 GPU.

Evaluation Following previous work (Zhang et al., 2018; Han et al., 2021), we use *Recall* (i.e. $R_{10}@k$, $k = (1, 2, 5)$) as our evaluation metric, which indicates the probabilities of whether the correct answer stands in the top k candidates given 10 samples. For the Douban benchmark, we also compute the values of mean average precision (*MAP*) and mean reciprocal rank (*MRR*), and precision at one ($P@1$) since the context in this dataset may contain multiple positive responses.

C.3 Conversational semantic role labeling

Training We follow the previous work (Wu et al., 2021) and solve the CSRL task as the sequence labeling problem. We keep the training settings same to CSAGN's. The parameters of the model are initialized from the pre-trained BERT, and updated by Adam optimizer with a linear learning rate schedule. The margin values of SimCTG and SimDRC are set to 0.2 and 0.5, respectively. The loss weight α in SimDRC is set to 0.2. All hyper-parameters are selected on the development set. The training process on the DuConv training set takes around 2 hours on two A100 GPUs.

Evaluation Following (Xu et al., 2021; Wu et al., 2021), we report the F1_{all}, F1_{intra} and F1_{inter} scores over the (predicate, argument, label) tuples. The arguments are categorized into two types, i.e., intra-arguments and cross-arguments, according to whether the argument appears in the same turn with the predicate or not. Therefore, we calculate the F1_{intra} and F1_{inter} scores on intra- and cross-arguments, respectively.

C.4 Results of DialoGPT on response generation task

D More evaluation results

Model	Method	BERTScore			BARTScore↑	BLEURT↑	Dis2/4↑
		P↑	R↑	F↑			
DialoGPT	greedy	12.13	10.22	10.96	-3.82	0.382	0.303/0.695
	beam	12.18	12.63	11.71	-3.90	0.383	0.300/0.671
	nucleus	12.22	12.65	12.05	-3.70	0.386	0.306/0.692
	contrastive	10.14	11.92	10.56	-4.19	0.247	0.288/0.653
SimCTG (ρ=0.6)	greedy	11.31	9.69	10.00	-3.98	0.371	0.271/0.622
	beam	12.01	12.46	12.15	-3.73	0.375	0.273/0.632
	nucleus	10.54	11.63	10.65	-3.79	0.366	0.269/0.627
	contrastive	12.12	12.62	12.22	-3.71	0.375	0.274/0.631
SimDRC (δ=0.5, α=0.2)	greedy	12.05	12.10	12.01	-3.90	0.322	0.271/0.639
	beam	12.63	13.75	12.65	-3.69	0.385	0.277/0.634
	nucleus	**12.74**	**13.88**	**12.78**	-3.65	**0.399**	0.317/0.713
	contrastive	12.62	13.15	12.45	**-3.64**	0.392	**0.322/0.744**

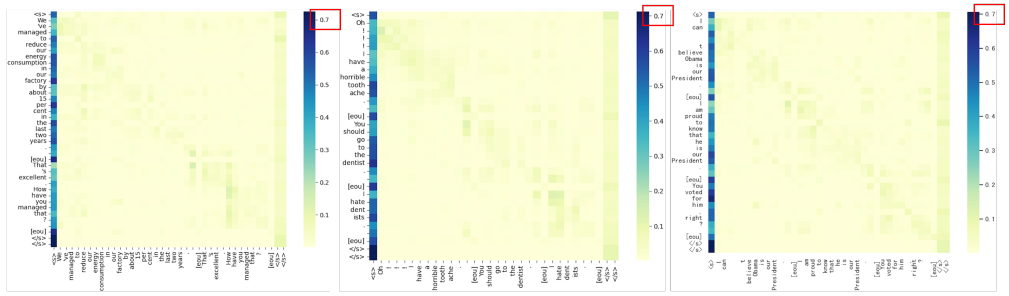
Table 6: Results of automatic evaluation on the DailyDialog dataset.

Table 6 shows the results of DialoGPT models on DailyDialog. Since there are no suitable DialoGPT models for Chinese, we only evaluate on the DailyDialog dataset here.

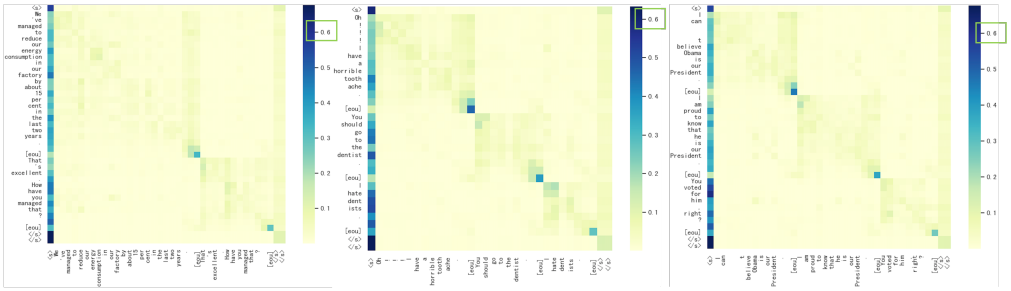
E More In-depth Analyses

E.1 Visualization of Self-Attention Weights

In this part, we visualize the self-attention weights of vanilla BART and BART+SimDRC trained on DailyDialog. As shown in Figure 5, we can see that 1) in vanilla BART, all tokens are primarily attended to the dialogue representative token, i.e. <s> in our cases. This would lead to the problem that all tokens receive much the same information and be nearby to each other on representations. This is exactly the problem of *anisotropy*; 2) with the aid of SimDRC, the tokens are encouraged to be concentrated on the tokens within the same utterance and discriminative to tokens in different

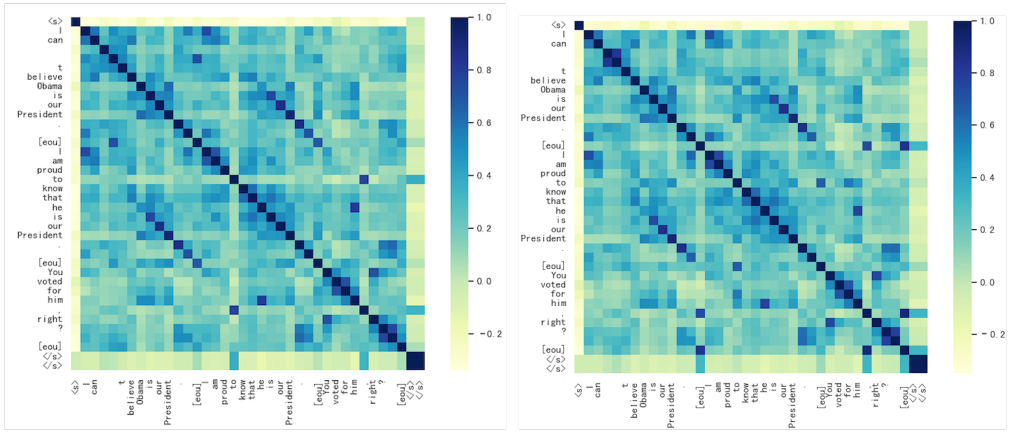


(a) Vanilla BART on Dailydialog



(b) SimDRC on Dailydialog

Figure 5: Attention weights visualization of different models trained on DailyDialog.



(a) Vanilla BART (b) BART + DiaSE

Figure 6: Comparison of vanilla BART and BART with dialogue state embeddings.

utterances. Therefore, we believe that SimDRC also essentially learns locality and isotropy on attention weights, while it explicitly calibrates on token embeddings.

E.2 Dialogue State Embedding vs. SimDRC

A straightforward idea to distinguish different utterance might be simply adding a dialogue state embedding (DiaSE) (Wolf et al., 2019b; Xu et al., 2021) to token embeddings. To this end, we introduce an additional randomly initialized dialogue state embedding layer into vanilla BART to differentiate different turns. During encoding, these additional embeddings are added with word embeddings and positional embeddings. As shown in Figure 6(b), even with the help of the turn indicator, the problem of anisotropy still remains. We think the reason is that DiaSE is essentially another kind of "segment embedding" which is not capable of capturing deep semantic or conversational features.

E.3 EXAMPLES OF GENERATED TEXTS

Context	SimDRC	Label
A: 刺猬很可爱！以前别人送了只没养，味儿太大！ **B**: 是很可爱但是非常臭。	是的，我家的刺猬也是，太臭没办法养	是啊，没办法养
A: 还发说说感冒早睡的人都几点了 **B**: 哈哈，睡不着 **A**: 看电视	哈哈，看电视也睡不着	看你
A: 都有什么好玩的哇 **B**: 没啥好玩的、一点儿意思都没有 **A**: 那跟沈阳差不多，还是大连好	大连好玩的地方太多了！	嗯呐、大连比这地儿好多了
A: puma的板鞋就挺好看的 **B**: 我看了，我这里差不多随处就能看到一双	哈哈哈，我这里也有一双，但是不好看	但是确实很百搭
A: 当你遇到一个不讲道理的女票又不忍心吵架时，你会发现沉默挺好的	不讲道理的女朋友是不会和你吵架的	讲道理，如果说自己女朋友无理取闹，当初为什么又要追她呢？其实并不矛盾。

Table 7: More generated examples on LCCC.

Context	**A**: Have you got any experience in advertising? **B**: Yes, I have been working in the Public Relations Section of a company in the past two years. I plan the advertising campaign and cooperate the work of artists and typographers. **A**: Do you have any practical experience as a secretary? **B**: Yes, sir. I acted as a secretary in an insurance company. **A**: Have you had any experience with computer?
BART	Yes, I have used a computer several times in the past two years. I've learned how to use it very well.
SimCTG	Yes, I have worked in a computer company for two years and I have a computer of my own.
SimDRC	Yes, I have worked in a computer company for two years and I have used Microsoft Word and Excel.
Context	**A**: Listen, Karen, I need your help. I don't know anyone here yet. **B**: I'm glad to help you. What's wrong? **A**: My mother-in-law just went into the hospital in L.A. Hank and I will be flying there tonight. **B**: I'm sorry to hear it. What's wrong with her? **A**: Doctors aren't sure yet. But the real problem is Suzy. She has a bad cold, and we don't want to take her with us. But we don't know who can stay with her. **B**: You need a babysitter. **A**: Yes, I'd ask Jill, the girl I've had before, but I need someone overnight.Maybe even for two nights.
BART	That's a good idea. I'll ask Jill. She's very good.
SimCTG	I'm sorry to hear that, but I don't think Jill would be able to do that.
SimDRC	I don't think Jill would be able to do it. She's not good at babysitting because she's only 17.
Context	**A**: 911 emergency. What is the problem? **B**: I would like to report a break-in. **A**: Where was the break-in? **B**: It happened at my house. **A**: Was anything stolen? **B**: I haven't checked yet. **A**: Why do you think there was a break-in?
BART	I don't know, but I think it was a robbery.
SimCTG	There was a break-in at my house last night.
SimDRC	I believe it was a break-in because the door was left open.
Context	**A**: Excuse me. I have an appointment with Mr. Li at nine. May I come in? **B**: Yes, come in please. I am Mr. Li. You must be My Liu, right? **A**: Yes, I am My Liu. Thanks. **B**: I'd like to start this interview with some questions. Why do you think you are qualified for this position?
BART	I am very interested in this position and I am very competent in this area.
SimCTG	I think I am qualified for this position because I have worked in this company for more than five years and I have experience in this field.
SimDRC	Well, first of all, I think I am a good team-player and I have a lot of experience in this line of work. Second, I enjoy working with people and have a good working relationship with my colleagues. Finally, I am very interested in this position.
Context	**A**: How do you get road news? **B**: I've a driver mate in my car! **A**: So you have all the news on broadcast?
BART	Yes, I've got all the news in my car!
SimCTG	Yes, I've got it on the TV, too.
SimDRC	Yes, I've got all the news on the radio in my car.

Table 8: More generated examples on DailyDialog.