We thank all the reviewers for their positive comments, and address their major questions and comments below. Clarifications will be added in the revision and we will keep improving our draft.

Reviewer #1 We thank the reviewer for the positive reviews. The remarks raised are addressed below.

Q: More details about model training

A: $\lambda_{\max}$ is selected via parameter search for $\lambda_{\max} \in \{0.1,\ 0.5,\ 1.0,\ 2.0,\ 5.0\}$, ending up with $\lambda_{\max} = 1.0$. For $\Gamma_\theta(\cdot)$, we use PPO in our implementation to keep the stability as the parameters are updated within a 'trust region'.

We are happy to release our code for better reproducibility.

Reviewer #2 We appreciate reviewer's acknowledgement of our novelty and suggestions provided.

Thanks for the suggestion on the paper improvements. In our updated version, we will add more results to show how our approach can handle the more useful dialogue generation problem.

Reviewer #3 We thank the reviewer for the positive reviews and appreciate the reviewer's suggestions.

Q: How to ensure that the sentences are short and meaningful.

A: Our paper proposed a general constraint-augmented reinforcement learning framework. In the recommendation task, natural language is only an example way for the user to express the preferences (*i.e.*, constraints), and we do not focus too much on how to handle more complicated or even free-form language. Therefore, similar to [20], our sentences in experiments are in the format of simple sentences with prefix. In this setting, the length of the sentences is implicitly determined by the training data. That is, we train the GRU model (the user simulator) on short sentences collected by human, and thus the trained model usually generates short sentences.

Q: Whether historical behaviors are considered.

A: Our approach considers the user historical behaviours within the current user session, although it does not model the user historical behaviours in previous sessions. The GRU tracks the user behaviours and the discriminator considers all previous user preferences in the current user session. However, we do agree that user historical behaviours from previous sessions can be employed to further enhance the performance.

Q: Comparison with traditional methods considering historical behaviors.

A: Our approach is not directly comparable to the traditional methods considering historical behaviors. Traditional recommendation models are usually trained in an offline manner, *i.e.*, the model is trained on a pre-collected dataset. By contrast, our method is proposed in a different setting, where the pre-collected dataset is not required and our recommender is interactively learned when the user interacts with the system. Moreover, it is not clear how to handle the interactive natural language feedback to provide interactive recommendations by traditional models.

Q: Incorporating traditional methods into our framework.

A: A simple approach can be developed to incorporate the traditional recommendation methods into our framework, to leverage historical behaviours from previous sessions. Assume we have users' historical behavior data and train a traditional model. In each user session, we make initial recommendations by the traditional model, collect natural-language feedback from users, and make further interactive recommendations by our framework. After a number of user sessions, we can update the traditional model based on the recently collected users' feedback to the items.

Thank you very much for providing us with the related work [49, 50, 51, 52]. We will definitely discuss these references in our related work section, and fix all the typos in our minor revision.

[49] Konstantina Christakopoulou, Alex Beutel, Rui Li, Sagar Jain, and Ed H Chi. Q&R: A two-stage approach toward interactive recommendation. In KDD, 2018.
[50] Yu Zhu, Yu Gong, Qingwen Liu, Yingcai Ma, Wenwu Ou, Junxiong Zhu, Beidou Wang, Ziyu Guan, and Deng Cai. Query-based interactive recommendation by meta-path and adapted attention-gru.arXiv:1907.01639, 2019.
[51] Yu Zhu, Hao Li, Yikang Liao, Beidou Wang, Ziyu Guan, Haifeng Liu, and Deng Cai. What to do next: modeling user behaviors by time-lstm. In IJCAI, 2017.
[52] Yu Zhu, Junxiong Zhu, Jie Hou, Yongliang Li, Beidou Wang, Ziyu Guan, and Deng Cai. A brand-level ranking system with the customized attention-gru model. In IJCAI, 2018.