We thank the reviewers for the constructive feedback. Following are response to some of the comments.

Innovation is not huge (R1, R2, R4) The novelty of the proposed SimpleTOD is based on simplifying task-oriented dialogue as a causal language modeling. Moreover, we show that using a simplified input sequence definition, special tokens, and embedding layer, the model can achieve higher performance compared to all previous models on DST. It also achieved state-of-the-art results on end-to-end setting, without any additional pretraining (used in SOLOIST) or multi-action data augmentation (used by DAMD).

Missing references (R1, R2) ARDM is only cited in the experiment section, line 205, and compared in the Table 7 (oracle information) in appendix B. SOLOIST appeared on arxiv after our paper, and we assumed it's a concurrent work. Detailed description of ARDM and SOLOIST will be added in related section and Tables in the camera ready version.

Weak difference compared to previous pretrained models (R2, R3) The Soloist model and model proposed by Ham et al. [17] are appeared within 30 days of ours. **ARDM** is pretrained model in alternating role for user and system, without using belief state or action annotation. It achieved better results than its predecessor on context-to-response generation (dialogue policy), but it requires oracle belief states to compute inform and success rate and is not applicable to end-to-end evaluation. It is also not designed for DST task. **Ham et al. [17]** used pretrained model trained on serialized sequence of dialogue. As mentioned in our paper, the role of special tokens are crucial in achieving better performance. They employed delimiter tokens for different segments, i.e. <usr>, <sys>, <ds> and <sa>. However, our tokenization is different than delimiter tokens, and is based on choosing semantic words which help the model to learn the semantic of each segments, i.e. belief, action, and response, and end of segment tokens such as <|endofbelief|>. Additionally, they used token-type embedding for user and system sequences, where our model does not use this embedding. Their proposed model achieved lower performance than previous baselines for DST and end-to-end evaluation, whereas we outperforms all previous models. **Pawel et al. [7]** used a different approach for creating input sequence by prepending dialogue context with belief state and DB search results. Their evaluation is only for context-to-response generation, and is not applicable to end-to-end setting. **SOLOIST** model removes action from input sequence and also used token-type embedding. It is pretrained on seven more dialogue datasets, before finetuning on MultiWOZ. Moreover, they used a data augmentation during training (contrastive learning), similar to DAMD paper, where they combined dialogue context and belief with a negative response and used the final token for classification, to improve their end-to-end performance. However, they did not report end-to-end performance without pretraining on other dialogue datasets. Their "w/o pretraining" setting is only evaluated on low resource training. Also, they did not report DST performance. **In summary, our proposed model is much simpler, in terms of (1) input sequence definition, (2) embedding layers, (3) training algorithm, and (4) pretraining, which make it easy to reproduce the results, and outperformed previous models on DST and end-to-end setting.**

Human evaluation (R2): An experiment with human conversing with the model in multi-domain setting to complete the task is conducted. This evaluation will be added in the camera ready version with a live demo.

Small Performance gain on DST (R2): Previous models on DST task have used bidirectional encoder and mostly based on BERT pretrained model. DSTQA, SST, and TripPy also used additional supervisions. However, our proposed model indicates that a unidirectional encoder can achieve state of the art results with no extra supervision. All previous models use a label cleaning, where our model outperforms them without cleaning, according to Table 1. The analysis shows that our model is robust to noisy annotation, especially for Type 2, known as early markup.

Experiment on other datasets (R2, R3, R4): MultoWOZ is the largest dataset that all previous models are evaluated for both DST and end-to-end settings. DAMD is not evaluated on CamRest767, which is a small dataset, and no previous model is evaluated on Schema-guided dialogue (SGD). Therefore, we focus our evaluation on MultiWOZ for a fair comparison. Evaluation on other dataset such as SGD and CamRest676 will be added in the camera-ready version.

Comparison with modular-based models (R4) We run an experiment similar to modular-based setting three language models for belief, action, and response generation, which briefly mentioned in line 218-222. Based on experiment, separate models can improve individual scores, but the combined score remains nearly identical.

Vulnerability to oracle information (R4) As shown in Table 7 of Appendix B, DAMD achieved higher performance when using oracle information. We suspect that due to augmentation, where they exploit single-to-many mapping of dialogue, that combined context and belief states with different true action and responses, they gained a higher performance. This is due to better action/response generation. However, DAMD performs lower in end–to-end setting perhaps due to lower accuracy in generating belief states that affect the subsequent tasks (action/response generation)

Analysis on inference speed (R4): This section will be added to the camera-ready version, with comparison to other DST and end-to-end methods. Moreover, the previous models on DST should classify or generate value for all slots, which is not scalable to large domain-slot situation. However, autoregressive models only generate slot-values that exist in dialogue context.