A Zero-Shot Approach for Multi-User Task-Oriented Dialog Generation

Shiv Surya **Yohan Jo** **Arijit Biswas** **Alexandros Potamianos**
Amazon
{shisurya,jyoha,barijit,potamian}@amazon.com

Abstract

Prior art investigating task-oriented dialog and automatic generation of such dialogs have focused on single-user dialogs between a single user and an agent. However, there is limited study on adapting such AI agents to multi-user conversations (involving multiple users and an agent). Multi-user conversations are richer than single-user conversations containing social banter and collaborative decision making. The most significant challenge impeding such studies is the lack of suitable multi-user task-oriented dialogs with annotations of user belief states and system actions. One potential solution is multi-user dialog generation from single-user data. Many single-user dialogs datasets already contain dialog state information (intents, slots), thus making them suitable candidates. In this work, we propose a novel approach for expanding single-user task-oriented dialogs (e.g. MultiWOZ) to multi-user dialogs in a zero-shot setting.

SINGLE-USER DIALOG

U: I am looking for a place to dine. The restaurant should serve Mediterranean food and should be in the south.
A: I'm afraid there aren't any. Would you like to try another area or cuisine?
U: What restaurants are cheap?

GENERATED MULTI-USER DIALOG

U: I am looking for a place to dine. The restaurant should serve Mediterranean food and should be in the south.
A: I'm afraid there aren't any. Would you like to try another area or cuisine?
U0: Another area? I don't want to go to a place that isn't in the south. It's important to me that the restaurant be in the centre of the city. If you don't have a place in the south, I'll go to a restaurant in the centre. It'll be a new experience for me.
U1: Maybe you are right. Let me see what I can find. There are some places in the centre that I would like to try. If you can't find anything in the centre, then try the south. There are many good restaurants in the south.
U0: Yes, I would like a restaurant in the South. Cheap is fine.

Figure 1: An example of a multi-user dialog generated by our approach is shown here along with the reference single-user dialog. The single-user utterance shown in red is expanded as a multi-user dialog (also in red).

1 Introduction

As AI assistants are increasingly becoming a staple in our social environments like homes, offices and classrooms, it is imperative that these assistants are capable of comprehending conversations and understand requests in social settings. Consider the conversations in Figure 1 between human users and agents. The first conversation is an example of a single-user conversation, involving a single human user interacting with an AI agent. The second conversation is a conversation between two users that convey the same intent as the first conversation but in a social setting. In contrast with the single-user conversation, the multi-user conversation shows social banter and collaborative decision making arising from the deliberation of a suggestion between the human users, acceptance or rebuttal of suggestions and social banter. These social and conversational cues distract AI agents from tracking the task-relevant information most relevant to them. Therefore, it is important to develop AI agents that are robust to multi-user conversations.

The development of dialog systems that support multi-user conversations is hindered by the lack of suitable multi-user task-oriented dialog datasets with annotations of user belief states and system actions. Collecting such datasets is expensive and it is impractical to support all domains of interest (a typical digital assistant can support thousands of domains). Training people to conduct multi-user dialogs is nontrivial, and annotating the dialogs with dialog states is time-consuming.

To solve this problem, we propose a zero-shot approach for expanding single-user task-oriented dialogs to automatically generate multi-user task-oriented dialogs (Section 3). We do this with two main goals in mind: (1) leverage existing annotations of dialog states in the source single-user dialogs, and (2) ensure a balance between the task-oriented tone and a social tone in the generated multi-user dialogs. To that end, our model expands each user utterance in a single-user dialog to a multi-user dialog that leads to the same user belief state as the source single-user utterance. Specifically, our model consists of two dialog generators and a turn planner. A task-oriented generator is trained on single-user task-oriented dialogs (e.g., MultiWOZ) responsible for generating utterances that contain task-relevant information (e.g. informed or requested slots), and a social generator trained on everyday dialogs (e.g., DailyDialog) responsible for generating utterance that express social dynamics of the speakers. We fuse these dialog generators to generate utterances in a multi-user dialog that is conversational, social, and relevant to the task at hand. The turn planner decides on an appropriate mix of dialog generators for generating a particular utterance in multi-user dialog.

Our automatic and human evaluation (Section 4) shows that our approach generates multi-user dialogs that contain social chatter consistent with dialog history and reflect user belief states consistent with source utterances. The main contributions of our work are:

- This is, to our knowledge, the first approach for generating multi-user dialogs from single-user task-oriented dialogs in a zero-shot setting.
- We show that the fusion of dialog generators is effective in maintaining topic relevance of social turns and improving quality of multi-user dialogs.
- We propose rule and model-based turn planners that select an appropriate combination of dialog generators and generate a multi-user dialog.
- Automated metrics and human quality evaluations show that dialog generator fusion and turn planning improves conversational fluidity in multi-user dialogs. It also better reflects the user belief.

2 Related Work

While there are public datasets of task-oriented dialogs annotated with dialog states (Andreas et al., 2020; Byrne et al., 2019; Rastogi et al., 2020; Zhu et al., 2020), most of them (if not all) are focused on single-user transactions, i.e., an agent converses with one user at a time. As a result, dialogs in these datasets do not reflect important dynamics of users making decisions together while interacting with an agent. Some dialog datasets cover such dynamics, like social banter (Li et al., 2017) and deliberation (Karadzhov et al., 2021; He et al., 2018), but they are not task-oriented or annotated with dialog states important for training dialog systems. While the FusedChat data (Young et al., 2022) contains task-oriented dialogs with social chatter, the dialogs are still single-user. By contrast, our goal is to build multi-user task-oriented dialogs reflecting social dynamics simultaneously.

The most similar line of work to ours is dialog generation from summaries, as a means for data augmentation for dialog summarization. In one framework (Gunasekara et al., 2021), a conversation generator (a seq2seq model) generates a conversation, and a summary generator (a seq2seq model) summarizes the generated conversation. The similarity between the generated summary and the original summary is used as a reward to inform the conversation generator in a reinforcement learning fashion. Another approach is to take a seed pair of a dialog and its summary, and gradually modify them iteratively (Liu et al., 2022). Specifically, an utterance of the dialog is replaced with a new utterance that is generated based on its context by a seq2seq model. Next, the summary is updated based on the updated dialog. Data augmentation using these methods improves dialog summarization accuracy in few-shot settings. The main difference between these approaches and ours is that they require seed dialogs. By contrast, we assume the more challenging scenario of zero-shot dialog generation, where we do not have enough dialogs to start with. This setting is more realistic because it is impractical to collect seed conversations of more than 100K domains in case of popular voice assistants.

Some studies have addressed generation of single-user task-oriented dialogs. One approach is to iteratively generate a user utterance and predict the resulting user belief state using two models (Kim et al., 2021). Another approach is to use a variational hierarchical dialog autoencoder that generates dialogs and their underlying dialog states simultaneously (Yoo et al., 2020). Importantly, these studies still tackle single-user dialogs,

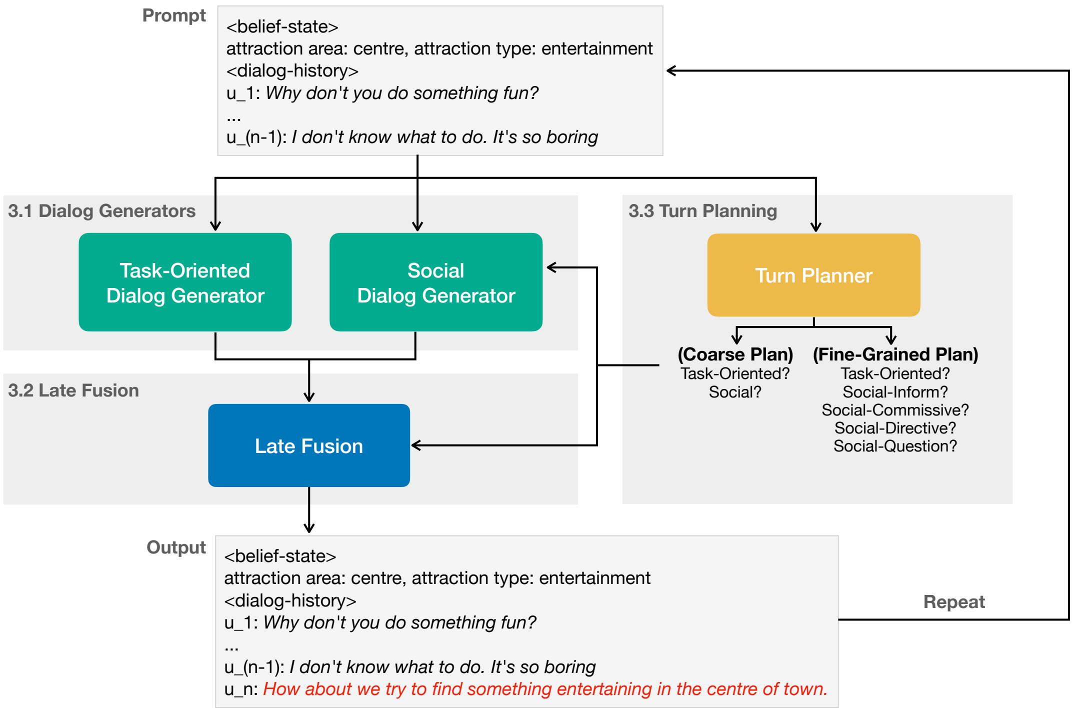

Figure 2: Architecture of the proposed multi-user dialog generator is shown here.

whereas our task of multi-user dialog generation is new and arguably more challenging as it requires resulting dialogs to reflect the social dynamics of decision-making between multiple users.

3 Our Approach: Multi-user Dialog Generation

Our proposed approach automatically generates multi-user task-oriented dialog. The utterances in a multi-user task-oriented dialog can be task-oriented, social or a mix of both. Training a single dialog generator capable of handling different kind of dialog turns e.g social or task-oriented requires an aligned dataset. In the absence of such a dataset, we use a bottom up approach of combining dialog generators capable of generating different kind of utterances (e.g social/task-oriented) to generate a multi-user dialog turn-by-turn. The architecture of this approach is shown in Figure 2. We train a task-oriented generator on single-user task-oriented dialogs responsible for generating task-oriented utterance, and a social generator trained on dialogs from everyday scenarios for generating social utterances (Section 3.1). We control turn level dynamics by training a turn planner. We use a turn planner (Section 3.3) that switches between the two dialog generators utterance-by-utterance to generate a multi-user rewrite. Additionally, we propose a mechanism to blend generations from the two dialog generators (Section 3.2).

3.1 Dialog Generators

Our model consists of two dialog generators. The first of these dialog generators is a task-oriented generator trained on single-user task-oriented dialogs (e.g., MultiWOZ). This generator is responsible for generating task-oriented utterances. We also train a social generator trained on dialogs from everyday scenarios (e.g., DailyDialog) capable of generating social utterances. We use a combination of these dialog generators to generate the utterances in a multi-user dialog. Training and inference details of dialog generators are provided in Appendix A.

3.1.1 Task-Oriented Dialog Generator

The task-oriented dialog generator is responsible for generating task-oriented utterances. This dialog generator is trained to generate an utterance that reflects a user belief state when prompted with the relevant dialog history between user and agent along with the user belief states.

Let $U = \{u_1, \dots u_n\}$ be a sequence of dialog turns consisting of both user and agent utterances. For a given user utterance $r = u_j, j \in \{1, ..., n\}$ we define $X^h = \{x_1^h, ..., x_{|h|}^h\}$ as the tokens that make up the dialog history of utterances $h = u_{<j}$.

Let $X^b = \{x_1^b, ..., x_{|b|}^b\}$ be the tokens that make up the user belief state b corresponding to user utterance r and $X^r = \{x_1^r, ..., x_{|r|}^r\}$ be the tokens that make up target utterance r.

We train the task-oriented dialog generator to produce the tokens X^r in target utterance r while being prompted with the concatenation of tokens in user belief state b and dialog history h, $[X^b, X^h]$. Let $X = \{x_1, x_2, ..., x_t, ..., x_{|h|+|b|+|r|}\}$ be the tokens in concatenation of user belief state b, dialog history h and target utterance r. The task-oriented dialog generator is trained to maximize the conditional log-likelihood of predicting the target tokens x_t for all $t > |h| + |b|$ to achieve this goal.

$$\mathcal{L} = -\sum_{t=|h|+|b|+1}^{|h|+|b|+|r|} \log P_{task}(x_t|x_{<t}) \quad (1)$$

where $|h|$,$|b|$ and $|r|$ are the number of tokens in dialog history, belief state and target respectively. P_{task} is the likelihood of the task-oriented dialog generator predicting a token given prompt and generated tokens as context.

3.1.2 Social Dialog Generator

The social dialog generator is responsible for generating utterances in social chit-chat. This dialog generator is trained to generate an utterance that best continues a dialog. It is prompted with the relevant dialog history between two users and the corresponding social act while generating this utterance. Given a sequence of dialog turns $U = \{u_1, \ldots u_n\}$ consisting of utterances in a dialog from everyday conversations. For a given utterance $r = u_j, j \in \{1, ..., n\}$, we define $X^h = \{x_1^h, ..., x_{|h|}^h\}$ as the tokens that make up the dialog history of utterances $h = u_{<j}$. X^h can optionally include the social act α for r. $X^r = \{x_1^r, ..., x_{|r|}^r\}$ are the tokens that make up target utterance r.

The social dialog generator is trained to produce the tokens X^r in target utterance r given the tokens in h, $[X^h]$ as input. Let $X = \{x_1, x_2, ..., x_t, ..., x_{|h|+|r|}\}$ be the tokens in concatenation of dialog history h and target utterance r. To achieve this goal, the social dialog generator is trained to to maximize the conditional log-likelihood of predicting the target tokens x_t for all $t > |h|$ in X.

$$\mathcal{L} = -\sum_{t=|h|+1}^{|h|+|r|} \log P_{social}(x_t|x_{<t}) \quad (2)$$

where $|h|$ and $|r|$ are the number of tokens in dialog history, and target respectively. P_{social} is the likelihood of the social dialog generator predicting a token at time step t given the prompt and generated token history as context.

3.2 Late Fusion of Dialog Generators

We propose a mechanism to blend generations between task-oriented and social dialog generators. Utterances generated by the task-oriented dialog generator can be command like. Similarly, utterances generated by the social dialog generator can be unrelated to the task in the utterance. We hypothesize that fusing the generators would allow us to achieve more conversational task-oriented dialog and increase the topic relevance of social dialog turns. Formally, at each step in generating a token from our dialog generator, we combine the token probabilities P_{task} and P_{social} from the task-oriented generator and social generator respectively. Formally, this is defined as:

$$p(x_t|x_{<t}) = \beta \cdot P_{task}(x_t|x_{<t}) + (1-\beta) \cdot P_{social}(x_t|x_{<t}) \quad (3)$$

where $\beta \in [0, 1]$ is a scalar blending factor.

The blending factor allows us to bias the decoding of the utterance towards the social dialog generator or the task-oriented dialog generator. This allows us to generate dialog that is a mix of both social and task-oriented dialog. Thus, we can achieve more natural transitions between turns and lessen abrupt topic changes at an utterance level in the multi-user dialog generation. Examples of how late fusion of dialog generators affects dialog generated are provided in Appendix A.

3.3 Turn Planning

We train a turn planner to predict the type of utterance we expect in the multi-user dialog and select the appropriate combination of dialog generators for use in generating a particular utterance. This model based planner predicts the type of utterance (social/task) we use in the next utterance of multi-user dialog, based on history of utterances in the dialog.

We design two variants of the turn planner. The first version, a coarse turn planner, is capable of predicting turn type as social or task. The fine-grained variant of the turn planner is trained to predict social acts of the utterance along with the type of turn. Training and inference details of the turn planners are provided in Appendix A.

3.3.1 Coarse Turn Planner

Given previous dialog turns as context, the coarse planner is trained to predict the turn type of the next turn in the dialog as social or task. The coarse turn planner is a pre-trained BERT model (Devlin et al., 2019) tuned in a binary classification setting. Following BERT, we use a fully-connected layer over the final hidden representation corresponding to the special classification token ([CLS]). The coarse turn planner is trained on as subset of the FusedChat dataset (Young et al., 2022).

3.3.2 Fine-Grained Turn Planner

The fine-grained turn planner is a classifier trained to predict turn type of the next utterance as social or task. In addition, it also predicts the social act of this utterance. DailyDialog dataset (Li et al., 2017) has annotations for the social act $\alpha \in \{$inform, commissive, directive, question$\}$ for each utterance. This social act α is given as an additional input prompt during finetuning of the social dialog generator. Specifically, the input prompt has the corresponding social act prepended at the beginning. This allows the social dialog generator to learn to maximize $P_{social}(x_t|x_{<t}, \alpha)$ enabling finer control of the social dialog generator via these social dialog acts.

Similar to the setting of the coarse turn planner, the fine-grained turn planner is a BERT classifier. However, this classifier is trained in a multi-label setting. The multi-label setting allows turn type prediction(social/task) along with prediction of the social act. However, annotations for these social acts are not present in the FusedChat dataset used to train our model turn planner. To circumvent this lack of annotations,we label each utterance in the FusedChat data with social acts using distance supervision as follows: (1) First, we train a multi-class BERT classifier on DailyDialog dataset with labels for each type of utterance. (2) We then use this trained multi-class classifier to label every utterance in FusedChat. This propagates distant labels of the utterance type to social and task utterances in FusedChat. The distant labels allow us to train the fine-grained turn planner on a subset of FusedChat to jointly predict the nature of the next turn in the dialog as social or task along with the type of utterance in the next turn.

3.3.3 Rule Planner

As a baseline, we design a rule based turn planner. The rule turn planner predicts a random number of social turns followed by a single task turn. In such a plan, each social and task turn is generated by the social or task generator respectively. While rule planners could append social turns, we only consider the setting in which social turns are prepended to obtain a setting similar to the user dynamics modeled by our model based turn planners. This enables a fairer comparison to the settings of the trained model planners in Section 3.3.1–3.3.2.

4 Experiments

We compare four variants of our proposed approach for multi-user dialog generations in our evaluations. These variants ablate the effect of (1) late fusion (2) type of turn planner. The variants are described below:

- **Rule Planner (RTP)**: This approach makes a hard choice between the social and task-oriented generator while generating a multi-user dialog. The type of utterance at each turn is determined by a rule planner (see Section 3.3.3)

- **Rule Planner w/ Late Fusion (RTP+LF)**: This approach blends the social and task-oriented generator using late fusion(see Section 3.2) while generating a multi-user dialog. The type of utterance at each turn is determined by a rule planner (see Section 3.3.3).

- **Coarse Planner w/ Late Fusion (CTP)**: This approach blends the social and task-oriented generator using late fusion(see Section 3.2) while generating a multi-user dialog. The type of utterance at each turn is determined by a coarse planner (see Section 3.3.1).

- **Fine-grained Planner w/ Late Fusion (FTP)**: This approach blends the social and task-oriented generator using late fusion(see Section 3.2) while generating a multi-user dialog. The type of utterance and social act at each turn is determined by a fine-grained model planner(see Section 3.3.2).

4.1 Datasets

4.1.1 MultiWOZ

We use this dataset as our primary task-oriented dialog dataset for generating and evaluating multi-user dialog generations. It is also used for training the task-oriented dialog generator. This dataset (Zang et al., 2020) contains multi-turn dialog between a single user and agent spanning multiple task domains. The dataset consists of 8,438 training dialogs and 1,000 validation and test dialogs each. In our experiments, we exclude the police, bus and taxi domains from the training set as they have very few dialogs in training.

4.1.2 DailyDialog

This is a corpus (Li et al., 2017) containing dialogs centered around daily life communications written by humans. We use the DailyDialog twice in our proposed approach. Firstly, it is used to train the social dialog generator. It is also used to train a classifier that labels utterances in FusedChat with social acts using distance supervision. The dataset contains 13,118 multi-turn dialogs spanning 10 daily life topics. This dataset is of particular interest to us as it has annotations for social act at an utterance level. We infer speaker turns based on utterance turns as the corpus does not explicitly indicate speaker information.

4.1.3 FusedChat

We use FusedChat to train the coarse and fine-grained turn planners. This is a dataset (Young et al., 2022) based on MultiWOZ (Zang et al., 2020). This work adds expanded social turns between a single user and agent. The dataset consists of MultiWOZ dialogs with prepended and appended social turns. We only consider the subset of FusedChat with social turns prepended to task-oriented turns. In this setting, the intent of the task-oriented turn following the social turns is strictly dependent on the topic of conversation in the social turns. This results in 3670 training dialogs and 500 validation and testing dialogs each.

4.2 Evaluation Measures

4.2.1 Automatic evaluation

Our multi-user rewrite of a single user task-oriented dialog should reflect the same user intent as the single user utterance, reliably cover the user belief states expressed in user utterance being rewritten while exhibiting high lexical diversity. We define the following automated metrics targeted at measuring semantic similarity between source utterance and utterances in multi-user rewrite and lexical diversity.

- **Semantic Similarity (SS)**: For a multi-turn rewrite of a single utterance, the cosine similarity of sentence embedding from transformers (Reimers and Gurevych, 2019) is computed between the average embedding of all generated utterances and the single-user utterance being rewritten. This measures the topical alignment of the multi-user dialog with the single user utterance.

- **Slot Coverage (SC)**: We measure the fraction of user belief states expressed in single-user utterance being mentioned in the multi-turn rewrite via a sub-string match. This metric captures the coverage of user belief states from user utterance in the generated multi-user rewrite.

- Additionally, we measure **Distinct-n/D-n (n=1,2,3)** (Li et al., 2015) and **Repetition-n/R-n (n=3)** which measures the percentage of unique n-grams and the fraction of copied 3-gram from prompt in a multi-user rewrite. These metrics measure lexical diversity in the generated multi-user dialog.

4.2.2 Human Evaluation

We conduct a human evaluation on Amazon Mechanical Turk. For this, we randomly sample 150 single-user utterances and corresponding generated multi-user dialogs from our proposed approach. We compare 4 variants of our proposed approach in a staggered manner which results in 450 multi-user rewrites. We conduct pairwise comparisons of generated multi-user rewrites, and each pair was evaluated by three human labelers. We ask humans to pick the better dialog in a pairwise evaluation by assessing:

- **Content Quality**: Consider the fluency, grammatically correctness, logical coherence and utterance ordering to determine the better multi-user dialog.

- **Belief State faithfulness**: Choose the multi-user dialog generation system that better reflects user beliefs expressed in the single user utterance.

```
DIALOG HISTORY:                                     D.1

U: Your reservation has been booked. Your
reference number is Q9DRQ55J . Please be aware
that your table will only be reserved until
16:30. Can i provide any more assistance?
A: I'm also looking for things to do in the
same general area as the restaurant.
U: Sure, are you interested in museums, sports,
parks, entertainment, swimming or boating?

USER BELIEF STATE:

attraction area: east, attraction type: cinema

UTTERANCE BEING REWRITTEN:

Are there any cinemas?If there isn't what about
something that has to do with multiple sports?

GENERATED MULTI-USER DIALOG:
U0: I think I'd like to go to the cinema. What
do you think of this one? It's a new cinema,
isn't it?
U1: Yes. It is a good cinema. You can go
to the movies. It's a new cinema, isn't it?
U0: Yes it is. Is there a cinema in the east?
```

```
DIALOG HISTORY:                                                   D.2
U: Please see what you can find out about jesus college
A: It is a college in the town centre, address Jesus Lane. Postcode
is cb58bl and it has free admission.
U: Do you happen to have the telephone number for the college?I would
like to call the main office there.
A: I can certainly look that up for you. Ok, it's 01223339485. Do you
need any more information or something else I can help with?

USER BELIEF STATE:

train day: tuesday, train destination: stansted airport

UTTERANCE BEING REWRITTEN:

Thank you.I'm looking to leave on Tuesday by train. Arrival will be
Stansted Airport.

GENERATED MULTI-USER DIALOG:
U0: Yes. The college has a large population. It is a very good
college. It has a lot of students. It is very easy to  get into the
college. It is in the centre of town. It is not very expensive to get
into college.It's very easy to find out about it.Thank you very much.
Are there any other places I can get to  for my college search?
U1: There are many places. For example. You can go to the college
website. It has lots of information. And also there are many places
to get a map.
U0: Can you help me find a train to the college on Tuesday?
```

```
DIALOG HISTORY:                                                                         D.3

A: Addenbrookes Hospital Telephone : 01223245151 Address : Hills Rd, Cambridge Postcode : CB20QQ Do you need a
taxi to get there?
U: No, thank you. Can you also find me a restaurant that serves British food in the centre? Postcode, please.
A: Cotto would fit your needs perfectly. Their postcode is CB11BG. Can I help you with anything else?
U: Are there any theatres near the Cotto?
A: Adc theatre is nearby. it is in park street

USER BELIEF STATE:

attraction area: centre , attraction type: theatre, attraction address: requested, attraction phone: requested

UTTERANCE BEING REWRITTEN:

Thank you. I'm also looking to leave on Tuesday by train. Arrival will be Stansted Airport.

GENERATED MULTI-USER DIALOG WITH COARSE TURN PLANNER:

U0: Great. Can I get the phone number and address for
the theatre? I am also looking for a restaurant in the centre.
Thank you.
U1: You can call the theatre from here. And I can give you the
address and phone number for the restaurant. You can book it from
here. It is very easy to do. And I will be happy to help you with
it. You are welcome. Goodbye.
U0: Can I have the address and telephone number for the
theatre, please?

GENERATED MULTI-USER DIALOG WITH FINE-GRAINED TURN
PLANNER:
U0: Thank you. Please get me the address and phone
number for the theatre. I will be happy to go
there.
U1: Cotto is in the centre of town. The theatre is
in the same street as the restaurant. So you can
walk there.
U0: That's great. Can I get the address and
telephone number for the theater?
```

Figure 3: Examples of multi-user task-oriented dialogs generated by our approach is shown here.

For each pairwise setting, we compute the majority vote based on the forced pairwise comparison. We use the Sign Test (Dixon and Mood, 1946) to compute statistical significance for both evaluation criteria. More details of the human evaluation including a screenshot of the evaluation template is available in Section B of the Appendix.

4.3 Results

Automatic evaluation of variations of our proposed approach are available in Table 1. Overall, using both late fusion and a turn planner achieves better performance. In Row 1, we see that the rule planner with hard choice of generators struggles to maintain relevance to intent in the single-user utterance. This is indicated by much lower seman-

Turn Planner	LF	SC↑	SS↑	D-1↑	D-2↑	D-3↑	R-3↓
RTP	✗	0.358	0.386	**0.797**	**0.960**	**0.986**	0.013
RTP	✓	0.446	**0.478**	0.658	0.898	0.963	0.012
CTP	✓	**0.480**	0.461	0.622	0.883	0.957	0.012
FTP	✓	0.464	0.455	0.747	0.939	0.980	**0.011**

Table 1: Automatic quality metrics on the test set to ablate the effect of turn planner type and fusion of dialog generators. *Legend - ↑: Higher is better, ↓: Lower is better, RTP: Rule Turn Planner, LF: Late Fusion, CTP: Coarse Model Turn Planner, FTP: Fine-Grained Model Turn Planner, SC: Slot Coverage, SS: Semantic Similarity.*

	Win %		
Evaluation	RTP	RTP+LF	Sign Test (p<0.05)
CQ	42	58	✓
BF	38	62	✓
Evaluation	RTP+LF	CTP	Sign Test (p<0.05)
CQ	52	48	✗
BF	54.66	45.33	✗
Evaluation	CTP	FTP	Sign Test (p<0.05)
CQ	62.66	37.33	✓
BF	61.34	38.66	✓

Table 2: Human evaluation on samples from the test set in a forced choice pairwise evaluation. Win% = % times multi-user dialogs from one model was preferred over the other when evaluated against a particular criterion. *Table Legend - RTP: Rule Turn Planner, LF: Late Fusion, CTP: Coarse Model Turn Planner, FTP: Fine-Grained Model Turn Planner, CQ: Content Quality, BF: Belief State Faithfulness*

tic similarity and slot coverage scores along with high n-gram diversity metrics. The addition of late fusion (RTP vs RTP+LF) produces a significant jump in both semantic similarity and slot coverage of the multi-user dialog when compared to the hard choice of generators rule based turn planner. This indicates better relevance to intent in the single-user utterance across the generated multi-user rewrite with late fusion. We also see a reduction in n-gram diversity metrics. This is expected as n-gram diversity would reduce when the social turns are also related to same topic. The replacement of rule turn planner by a coarse turn planner (RTL+FL vs CTP) produces improvements in slot coverage. The fine-grained planner (FTP) gets comparable semantic similarity and slot coverage with higher lexical diversity scores. Holistically, this is indicative of the fine-grained planner showing comparable faithfulness in reflecting the user belief state, while reducing repetition across utterances in the multi-user dialog.

Results of our human evaluation are available in Table 2. From the pairwise evaluation, we see that late fusion outperforms a hard choice of the backbone generators with statistical significance. This is also observed in the automatic evaluation where improved semantic similarity scores and slot coverage indicate that late fusion produces dialog turns that are more related to intent expressed in single-user utterance. The coarse model turn planner (CTP) and rule turn planner with late fusion (RTL+FL) are tied without a statistically significant result on both criteria. This result is aligned with close automatic metric for these models observed in Table 1. This shows the limitations of training turn planners in low data regimes. We expect the performance of our approach with the coarse and fine-grained model planners to improve with training on larger datasets and data augmentation.

Some examples of generations from our approach are shown in Figure 3. The generations show reasonable faithfulness to the intent and user belief in the single user utterance being rewritten across D.1–D.3. However, in example D.1, we see repetition of the phrase "It's a new cinema, isn't it?" across utterances by different users indicating the challenge of consistent Point-of-View (PoV) depiction. In D.2, we find that the role and characteristics of an agent bleeds into the users engaged in a conversation with command like responses despite blending of dialog generators. Example D.3 contrasts multi-user dialog generation with coarse and fine-grained turn planners. Here we see that while both generations cover the user beliefs, the dialog generated with the fine-grained turn planner is more coherent across turns. The coarse-grained planner shows content repetition across turns. Additional examples are shown in Section C of the appendix.

4.4 Challenges and Future work

Despite these promising results, we find that there are considerable challenges to be tackled. Our ability to control the social utterances in the multi-user rewrite is limited to broad social acts, i.e., question or inform. This limitation arises from the dataset we use for fine-tuning the social dialog generator. Using datasets with fine-grained annotations for utterance type or larger language models capable of instruction prompting are potential directions to address this. Another challenge is ensuring consistency of user beliefs across social utterances. We observe challenges in maintaining consistent beliefs across utterances in a dialog sequence for a user. Further, maintaining consistent Point-of-View (PoV) depiction is challenging. We find that the role and characteristics of an agent bleeds into the users engaged in a conversation. Planning approaches like those employed in story and long text generation (Rashkin et al., 2020; Yao et al., 2019) is one possible family of approaches that could reduce inconsistency in user behaviour, PoV and provide more control.

Ethics Statement

Advances in multi-user dialog generation techniques would aid training of digital assistants. As AI assistants are increasingly becoming a staple in our social environments, synthetic methods of multi-user dialog generation would aid the training of these assistants and ensure they are capable of comprehending human conversations and understand task-oriented requests in social settings. This would help increase human-machine interaction and enhance human productivity in collaborative settings.

Synthetic multi-user dialog generation techniques would also reduce the need of (the gold standard for data collection) crowdsourcing. This would also have a positive effect on human productivity and reduce the need for humans to manually write dialogs for different scenarios.

We use language models as initialisation for our dialog generators. These are trained on data collected from the web. Hence, issues related to bias and abusive language are a potential concern. These concerns of abusive content should be largely mitigated as we fine-tune of the dialog generators on task-oriented and everyday conversation datasets with sanitised data. The generator fine-tuning and prompt structure used for dialog generation should limit unintended consequences as all generations are trained to reflect the intent of the single-user dialog. However, with our proposed method of multi-user dialog generation, any racial, ethnic or other forms of bias present in the datasets used to train the dialog generators is likely to get propagated to the generated multi-user dialog.

References

Jacob Andreas, John Bufe, David Burkett, Charles Chen, Josh Clausman, Jean Crawford, Kate Crim, Jordan DeLoach, Leah Dorner, Jason Eisner, Hao Fang, Alan Guo, David Hall, Kristin Hayes, Kellie Hill, Diana Ho, Wendy Iwaszuk, Smriti Jha, Dan Klein, Jayant Krishnamurthy, Theo Lanman, Percy Liang, Christopher H Lin, Ilya Lintsbakh, Andy McGovern, Aleksandr Nisnevich, Adam Pauls, Dmitrij Petters, Brent Read, Dan Roth, Subhro Roy, Jesse Rusak, Beth Short, Div Slomin, Ben Snyder, Stephon Striplin, Yu Su, Zachary Tellman, Sam Thomson, Andrei Vorobev, Izabela Witoszko, Jason Wolfe, Abby Wray, Yuchen Zhang, and Alexander Zotov. 2020. Task-Oriented Dialogue as Dataflow Synthesis. *Transactions of the Association for Computational Linguistics*, 8:556–571.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a Realistic and Diverse Dialog Dataset. *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 4515–4524.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Wilfrid J Dixon and Alexander M Mood. 1946. The statistical sign test. *Journal of the American Statistical Association*, 41(236):557–566.

Chulaka Gunasekara, Guy Feigenblat, Benjamin Sznajder, Sachindra Joshi, and David Konopnicki. 2021. Summary Grounded Conversation Generation. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3748–3756.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling Strategy and Generation in Negotiation Dialogues. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2333–2343.

Georgi Karadzhov, Tom Stafford, and Andreas Vlachos. 2021. DeliData: A dataset for deliberation in multi-party problem solving. *arXiv*.

Sungdong Kim, Minsuk Chang, and Sang-Woo Lee. 2021. NeuralWOZ: Learning to Collect Task-Oriented Dialogue via Model-Based Simulation. *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 3704–3717.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

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.

Yongtai Liu, Joshua Maynez, Gonçalo Simões, and Shashi Narayan. 2022. Data Augmentation for Low-Resource Dialogue Summarization. *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 703–710.

Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao. 2020. PlotMachines: Outline-conditioned generation with dynamic plot state tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4274–4295, Online. Association for Computational Linguistics.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards Scalable Multi-Domain Conversational Agents: The Schema-Guided Dialogue Dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8689–8696.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *AAAI*, pages 7378–7385.

Kang Min Yoo, Hanbit Lee, Franck Dernoncourt, Trung Bui, Walter Chang, and Sang-goo Lee. 2020. Variational Hierarchical Dialog Autoencoder for Dialog State Tracking Data Augmentation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3406–3425.

Tom Young, Frank Xing, Vlad Pandelea, Jinjie Ni, and Erik Cambria. 2022. Fusing task-oriented and open-domain dialogues in conversational agents. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11622–11629.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117, Online. Association for Computational Linguistics.

Qi Zhu, Kaili Huang, Zheng Zhang, Xiaoyan Zhu, and Minlie Huang. 2020. CrossWOZ: A Large-Scale Chinese Cross-Domain Task-Oriented Dialogue Dataset. *Transactions of the Association for Computational Linguistics*, 8:281–295.