Planning with Diffusion Models for Target-Oriented Dialogue Systems

Hanwen Du[♣] **Bo Peng**[♣] **Xia Ning**[♣♠♡✉]
[♣]Department of Computer Science and Engineering, The Ohio State University, USA
[♠]Department of Biomedical Informatics, The Ohio State University, USA
[♡]Translational Data Analytics Institute, The Ohio State University, USA
{du.1128,peng.707,ning.104}@osu.edu

Abstract

Target-Oriented Dialogue (TOD) remains a significant challenge in the LLM era, where strategic dialogue planning is crucial for directing conversations toward specific targets. However, existing dialogue planning methods generate dialogue plans in a step-by-step sequential manner, and may suffer from compounding errors and myopic actions. To address these limitations, we introduce a novel dialogue planning framework, `DiffTOD`, which leverages diffusion models to enable non-sequential dialogue planning. `DiffTOD` formulates dialogue planning as a trajectory generation problem with conditional guidance, and leverages a diffusion language model to estimate the likelihood of the dialogue trajectory. To optimize the dialogue action strategies, `DiffTOD` introduces three tailored guidance mechanisms for different target types, offering flexible guidance toward diverse TOD targets at test time. Extensive experiments across three diverse TOD settings show that `DiffTOD` can effectively perform non-myopic lookahead exploration and optimize action strategies over a long horizon through non-sequential dialogue planning, and demonstrates strong flexibility across complex and diverse dialogue scenarios. Our code and data are accessible through https://github.com/ninglab/DiffTOD.

1 Introduction

Target-Oriented Dialogue (TOD) systems can assist users in accomplishing specific targets through interactive natural language conversations (Deng et al., 2023a; Qin et al., 2023), such as completing a transaction (He et al., 2018) and providing personalized recommendations (Wang et al., 2023b). With the rise of Large Language Models (LLMs), TOD systems have undergone a paradigm shift toward LLM-integrated architectures, which are highly capable of generating high-quality, human-like responses that enhance user engagement experiences. (Ou et al., 2024; Deng et al., 2024b).

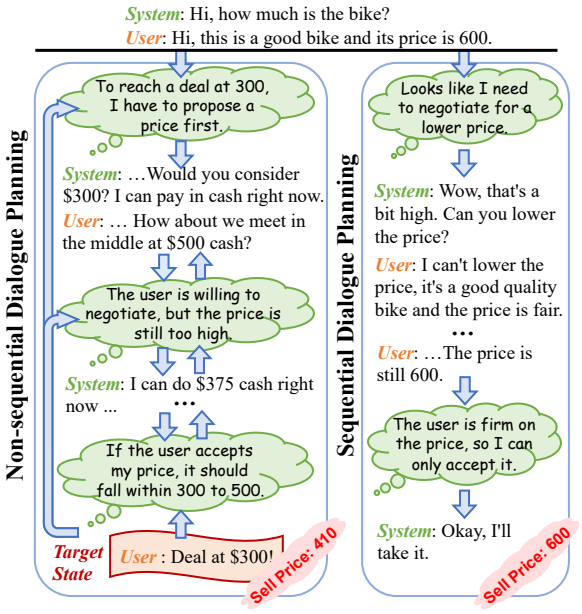

Figure 1: Using sequential and non-sequential dialogue planning methods for negotiation dialogues.

However, since LLMs are typically trained to follow instructions passively (Ouyang et al., 2022), they often lack the proactivity to plan and guide the conversation toward the intended target (Hao et al., 2023; Deng et al., 2025), a crucial property for the successful achievement of targets in TOD (Wang et al., 2023c; He et al., 2024). Therefore, how to develop effective dialogue planning methods for LLM-integrated TOD agents that can strategically guide the conversation toward the target remains an ongoing challenge (Wang et al., 2023c; Deng et al., 2024b).

To enhance the capability of LLMs for dialogue planning, existing methods prompt LLMs to generate dialogue plans through reflection (Zhang et al., 2023a; Deng et al., 2023b) or demonstration (Zheng et al., 2024) mechanisms, or formulate TOD as a Markovian Decision Process (Bellman, 1957) (MDP) and train policy-based agents

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 20307–20329
July 27 - August 1, 2025

via reinforcement learning to learn dialogue strategies (Deng et al., 2024b). However, as LLMs generate text in an autoregressive manner and policy gradient relies on the sequential assumption in MDP, all these methods generate dialogue plans in a step-by-step sequential manner. As a result, they can only plan the next dialogue action based on the observation of previous responses without looking ahead, which may suffer from compounding errors and myopic actions (Jiang et al., 2016; Ma et al., 2025). By contrast, non-sequential dialogue planning methods can generate dialogue actions by considering both the past and possible future responses with iterative refinement, offering desirable properties such as lookahead reasoning and maintaining global consistency for overall achievement of the target (Janner et al., 2022; Zhang et al., 2023b). For example, in the negotiation dialogue in Figure 1, the non-sequential planning method can generate a dialogue action (propose a price of $375) by considering both the previous history and the future target, and make dynamic adjustments to the price range accordingly. By comparison, the sequential planning method gets stuck in suboptimal dialogue actions that fail to negotiate a lower price.

In this work, we aim to develop a non-sequential dialogue planning framework, called `DiffTOD`, to address the limitations mentioned above. To achieve this, we first demonstrate that dialogue planning can be transformed into a trajectory generation problem. By relaxing the sequential constraint in MDP to allow for non-sequential generation, we reveal a strong connection between the likelihood of the generated trajectory and the denoising process in diffusion models (Ho et al., 2020). Based on this insight, we adopt a masked diffusion language model (Lou et al., 2024) to estimate the likelihood of the trajectory by fine-tuning it on the dialogue history from the training dataset. Furthermore, to ensure the optimality of the action strategies in the generated trajectories, we decompose the likelihood of trajectory generation into (1) an unconditional part generated by the diffusion model, and (2) a conditional part that allows for flexible guidance at test time to direct the trajectory sampling process toward the desired dialogue target. Based on this decomposition, we design three guidance mechanisms tailored to different types of targets in TOD, which can be applied separately or combined to effectively guide the dialogue toward the target. Extensive experiments across three diverse TOD settings show that `DiffTOD` substantially outperforms baselines on target achievement success and demonstrates strong flexibility across complex and diverse dialogue scenarios.

Our contributions are summarized as follows:

- We present `DiffTOD`, a novel dialogue planning framework that leverages a diffusion language model for non-sequential dialogue planning.
- We design three guidance mechanisms tailored to different types of TOD targets, enabling effective and flexible control at test time to direct the dialogue toward diverse and complex targets.
- Our extensive experiments show that `DiffTOD` outperforms baseline methods and demonstrates strong flexibility across diverse scenarios.

2 Related Works

2.1 LLM-Integrated Dialogue Planning

To enhance the dialogue planning capability of LLMs, several approaches have been proposed along various dimensions, such as intricate prompt engineering to elicit the planning and reflection of LLMs (Zhang et al., 2023a; Deng et al., 2023b), improving the planning capability of LLMs through demonstrations (Zheng et al., 2024), integrating LLMs with a plug-and-play policy planner (Deng et al., 2024b), and applying dual-process theory to guide dialogue planning (He et al., 2024). Despite promising, all these methods generate dialogue plans in a step-by-step sequential manner. Such sequential approach may struggle with targets that require complex planning and reasoning over multiple conversational turns (Kambhampati et al., 2024; Ye et al., 2025). *In contrast,* `DiffTOD` *leverages a diffusion model for non-sequential dialogue planning, which can effectively optimize dialogue actions for overall target achievement and allow for flexible guidance at test time.*

2.2 Diffusion Models

Diffusion models have emerged as an expressive class of generative models known for their ability to generate high-quality data through iterative denoising (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song and Ermon, 2019). They have found widespread applications across various domains, such as image synthesis (Dhariwal and Nichol, 2021; Rombach et al., 2022), protein design (Watson et al., 2023; Gruver et al., 2023), molecular generation (Chen et al., 2025), and trajectory generation for reinforcement learning (Janner et al., 2022; He et al., 2023a). Recently, diffusion models

have also shown remarkable potential in text generation, with approaches ranging from continuous diffusion language models that denoise from a latent space of word embeddings (Gong et al., 2023; Gulrajani and Hashimoto, 2024), and discrete diffusion language models that generate text from a sequence of mask tokens (Austin et al., 2021; He et al., 2023b; Lou et al., 2024). *Different from these methods,* `DiffTOD` *focuses on optimizing dialogue strategies with diffusion models, and can generate dialogue plans that effectively achieve the target across diverse and complex scenarios.*

3 Dialogue Planning for TOD

3.1 Target-Oriented Dialogue

A Target-Oriented Dialogue (TOD) consists of alternating responses between the user and the system, and a target $g \in \mathcal{G}$ (e.g., recommending a specific item, reaching a deal with the user) that the system aims to achieve during the conversation. It can be formulated as follows:

$$\mathcal{D}_g = \{(d_0^s, d_0^u), \cdots, (d_t^s, d_t^u), \cdots, (d_T^s, d_T^u); g\}, \tag{1}$$

where (d_t^s, d_t^u) denotes the t-th conversational turn consisting of the system's response d_t^s and the user's response d_t^u, T denotes the maximum number of conversational turns. The starting conversational turn (d_0^s, d_0^u) is usually initialized with a predefined utterance d_0^s from the system's side (e.g., start the conversation with a greeting from the system) followed by the user's response d_0^u. We denote d_0^s as an empty string if the conversation starts from the user's side.

3.2 Conversational MDP

To establish a principled framework for dialogue planning and optimization, we formulate TOD as a Markovian Decision Process (MDP) (Bellman, 1957). The conversational MDP is defined by a quintuple $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma\}$, where $\mathcal{S}$ denotes the set of states, which summarizes all the information about the conversation history and the dialogue context; $\mathcal{A}$ denotes the set of actions that the system can take at each conversational turn; $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ denotes the transition to the next state after taking an action from the current state; $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ denotes the immediate reward after taking an action; and $\gamma \in (0, 1)$ denotes the discount factor. An illustration of how to calculate the state and the action is presented in Appendix A.

State At each conversational turn t, the state $s_t \in \mathcal{S}$ is defined as a sequence $s_t = (d_0^s, d_0^u, \cdots, d_{t-1}^s, d_{t-1}^u)$ that includes all the user's and the system's responses from previous turns. Besides, the system also has access to all the information about the dialogue context, such as the user's profile and the description of the target item.

Action The set of action $\mathcal{A}$ denotes all the responses the system can take in the conversation. At each conversational turn t, the system takes an action $a_t \in \mathcal{A}$ and generates a response d_t^s.

Transition The transition function $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ denotes the transition to the next state $s_{t+1} \in \mathcal{S}$ from the current state $s_t \in \mathcal{S}$. After the system takes an action $a_t \in \mathcal{A}$, the user will give a response d_t^s, and the next state $s_{t+1} = (d_0^s, d_0^u, \cdots, d_{t-1}^s, d_{t-1}^u, d_t^s, d_t^u)$ is updated to include the user's and the system's responses at the current conversational turn t.

Reward The reward function $\mathcal{R} : r(s, a) \to \mathbb{R}$ denotes the immediate reward after taking an action $a_t \in \mathcal{A}$ in the current state $s_t \in \mathcal{S}$. Usually, we assign a positive reward when the conversation achieves the target (e.g., a deal is reached), a negative reward when the conversation fails to achieve the target (e.g., unable to reach a deal), and no reward is assigned during the conversation.

Dialogue Trajectory We define a dialogue trajectory $\tau_{0:t} = (s_0, a_0, \cdots, s_t, a_t)$ as a sequence of states and actions up to and including the t-th turn.

3.3 Dialogue Planning

Based on the definition of MDP, the problem of dialogue planning can be formulated as a constrained trajectory optimization problem. The goal is to find a sequence of optimal actions $a_{1:T}^*$ that maximize the cumulative reward, subject to the constraint that the transition from the current state s_t to the next state s_{t+1} should follow the transition function $\mathcal{T}$ defined in the MDP. Note that we exclude the first action a_0 from the optimization problem, since the conversation usually begins with a predefined opening from the system. Formally, the problem of dialogue planning can be defined as follows:

$$\begin{aligned} a_{1:T}^* &= \arg\max_{a_{1:T}} \sum\nolimits_{t=1}^{T} \gamma^t r(s_t, a_t) \\ \text{s.t.,} \quad & s_{t+1} = \mathcal{T}(s_t, a_t), \quad 0 \le t < T, \end{aligned} \tag{2}$$

where $\gamma, r(s_t, a_t)$ and $\mathcal{T}$ denote the discount factor, the reward and the transition defined in the MDP.

Following the literature (Wang et al., 2023c; Deng et al., 2024b), we decompose the TOD task into two stages: dialogue planning and dialogue generation. After a dialogue plan is constructed, we prompt an LLM to role-play as the system: at each conversational turn, the system will interact with the user and generate a response that strictly follows the action strategies in the dialogue plan.

4 Introducing DiffTOD

4.1 Trajectory Modeling with Diffusion Model

To solve the optimization problem in Equation 2, we define a *planner* $p_{\theta_p}(a_t|s_t)$ that generates an action a_t given the current state s_t, and an *environment* $p_{\theta_e}(s_t|s_{t-1}, a_{t-1})$ that generates the next state s_t given the current state s_{t-1} and action a_{t-1}. The planner and the environment will work together to generate a sequence of states $s_{1:T}$ and actions $a_{1:T}$ that constitute the trajectory $\tau_{0:T}$. The dialogue planning problem can then be transformed into a trajectory generation problem as follows:

$$\begin{aligned} p_{\theta_p,\theta_e}(\tau_{0:T}) &= p(s_0, a_0) \cdot p_{\theta_p,\theta_e}(s_1, a_1, \cdots, s_T, a_T) \\ &= p(s_0, a_0) \cdot \prod_{t=1}^{T} p_{\theta_e}(s_t|s_{t-1}, a_{t-1}) \cdot p_{\theta_p}(a_t|s_t) \\ &= p(s_0, a_0) \cdot \prod_{t=1}^{T} p_{\theta_p,\theta_e}(s_t, a_t|s_{t-1}, a_{t-1}). \end{aligned} \tag{3}$$

Note that we exclude s_0 and a_0 from the optimization problem, since the conversation typically begins with a predefined opening from the system. By Markov property, $p(s_t, a_t|s_{t-1}, a_{t-1}) = p(s_t, a_t|s_{1:t-1}, a_{1:t-1}) = p(s_{1:t}, a_{1:t}|s_{1:t-1}, a_{1:t-1})$. Therefore, the likelihood of the trajectory $p_\theta(\tau_{0:T})$ can be rewritten as follows:

$$p_{\theta_p,\theta_e}(\tau_{0:T}) = p(s_0, a_0) \cdot \prod_{t=1}^{T} p_{\theta_p,\theta_e}(\tau_{1:t}|\tau_{1:t-1}). \tag{4}$$

This formulation shows that the likelihood of the trajectory can be decomposed into a prior distribution $p(s_0, a_0)$ and the product of conditional distributions $p_{\theta_p,\theta_e}(\tau_{1:t}|\tau_{1:t-1})$. The conditional distributions can be interpreted as a trajectory inpainting process, where a partially observed trajectory $\tau_{1:t-1}$ with only a subset of states and actions is progressively reconstructed into a more complete trajectory $\tau_{1:t}$ with additional information about the state s_t and the action a_t. More generally, if we allow for non-sequential generation of states and actions, and decompose the whole generation process into N steps, with τ^n representing a partially observed trajectory, τ^{n-1} representing a more complete trajectory, τ^N representing the trajectory with only the initial state and action, and τ^0 representing the complete trajectory, we can see that Equation 4 is actually closely related to the denoising process of diffusion models (Ho et al., 2020):

$$p_\theta(\tau^{0:N}) = p(\tau^N) \prod_{n=1}^{N} p_\theta(\tau^{n-1}|\tau^n). \tag{5}$$

Using this formulation, we can train a generative diffusion model p_θ that can reconstruct the entire trajectory τ^0 from an incomplete trajectory τ^n. In this way, the diffusion model p_θ can function both as the *planner* that generates actions a_t when $\tau^n = \{s_0, a_0, \cdots, s_{t-1}, a_{t-1}, s_t\}$, and the *environment* that generates s_{t+1} when $\tau^n = \{s_0, a_0, \cdots, s_t, a_t\}$. Note that the denoising process of the diffusion model introduces another "step" variable. We use the subscript $t(0{\leq}t{\leq}T)$ to denote the conversational turn, and the superscript $n(0{\leq}n{\leq}N)$ to denote the diffusion step.

While this formulation is general, a key design choice remains undecided: choosing an appropriate space to represent the states and actions and defining how they should be represented in that space. In our implementation, we choose to represent the states and actions in their original natural language forms, and fine-tune a masked diffusion language model (Lou et al., 2024) on the dialogue history from the training dataset to model the likelihood of p_θ. To avoid unnecessary repetition of the same d_t^s, d_t^u across multiple states and actions, we model the likelihood of the trajectory with the equivalent formulation $\tau_{0:t} = \{d_0^s, d_0^u, \cdots, d_{t-1}^s, d_{t-1}^u, d_t^s\}$ by concatenating all the context and history of the dialogue in its natural language form.

While representing states and actions in natural language is simple and intuitive, we also note that our formulation supports other design choices, such as representing states and actions in a unified latent space (Hao et al., 2024). We leave the study of alternative design choices for future work.

4.2 Optimizing Action Strategies

A common characteristic of the TOD datasets is that the action strategies in the conversation history are often suboptimal and not explicitly optimized for target achievement. For example, in the CraigslistBargain dataset (He et al., 2018), some

buyer-seller conversations end without reaching a deal. As a result, the diffusion model trained on these datasets may learn to generate valid but often suboptimal actions. To guide the diffusion model to generate optimal actions, inspired by the control-as-inference graphical model (Levine, 2018; Janner et al., 2022), we introduce a binary variable $\mathcal{O} \in \{0, 1\}$ that indicates whether the dialogue trajectory achieves the target g, and factorize the likelihood of generating a trajectory conditioned on $\mathcal{O} = 1$ as follows:

$$p_\theta(\tau_{0:T}|\mathcal{O}=1) \propto p_\theta(\tau_{0:T}) \cdot p_\theta(\mathcal{O}=1|\tau_{0:T}). \quad (6)$$

This formulation decomposes the trajectory generation process into two parts: sampling the trajectory $\tau_{0:T}$ with the diffusion model, and calculating $p(\mathcal{O}=1|\tau_{0:T})$ as guidance to ensure that the generated trajectory is optimal. In TOD, the reward function $r(s_t, a_t)$ is usually sparse—a reward is assigned only when the dialogue reaches a target state, and most state-action pairs will not receive an intermediate reward (Feng et al., 2023; Kwan et al., 2023). Therefore, optimal trajectory generation can be viewed as generating a feasible dialogue trajectory while ensuring that certain states and actions along the trajectory achieve the target, and the guidance can be formally defined as:

$$p(\mathcal{O}=1|\tau_{0:T}) = \begin{cases} 1 & \exists s_t, a_t \in \tau_{0:T}, g(s_t, a_t)=1 \\ 0 & \text{otherwise,} \end{cases} \quad (7)$$

where $g(s_t, a_t) = 1$ indicates that the action a_t achieves the target g in the given state s_t. This can be implemented as a trajectory inpainting process: the diffusion model performs conditional denoising from an incomplete trajectory τ^n with only the desired states and actions that achieve the target, and then inpaints the rest parts of the trajectory. Since the target can vary across different dialogue scenarios, we design customized guidance mechanisms tailored to different target types. These mechanisms set specific states and actions within the trajectory as conditions according to different target types, and can be used separately or combined to provide effective guidance.

Word-Level Guidance A common type of TOD target is to mention specific keywords (Tang et al., 2019; Zhong et al., 2021) in the conversation. To achieve such targets, we can append the target keyword to the desired place in the dialogue, and then perform denoising using the diffusion model with the target keywords fixed as guidance. The word-level guidance will ensure that the diffusion model can generate a coherent dialogue plan that naturally incorporates the target keywords. An illustration of the word-level guidance is presented in Figure 2.

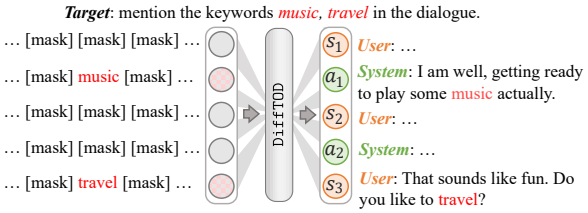

Figure 2: An illustration of the word-level guidance.

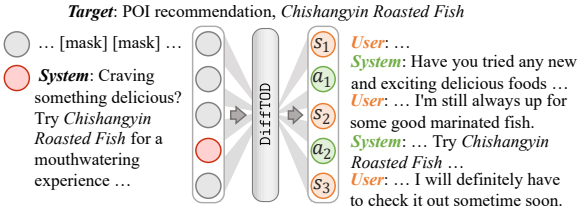

Figure 3: An illustration of the semantic-level guidance.

Semantic-Level Guidance Some TOD targets are not explicitly represented by specific keywords, but instead are defined by the semantic meaning conveyed in the responses (Bai et al., 2021; Yang et al., 2022). For example, in target-driven conversational recommendation (Wang et al., 2023b), the target can be defined semantically as the dialogue reaching a state where the system successfully recommends the specified item. To achieve such targets, the semantic-level guidance performs denoising with the diffusion model by conditioning on the state or action that conveys the desired semantic meaning. Since states and actions described in different natural language forms may share the same semantic meaning, we can sample multiple dialogue plans with paraphrased versions of the same condition and perform Minimum Bayes Risk (MBR) decoding (Koehn, 2004; Gong et al., 2023) to improve the quality of the generated dialogue plans. An illustration of the semantic-level guidance is presented in Figure 3. The prompt template and example outputs for generating the semantic guidance are in Figure A5.

Search-Based Guidance Some TOD settings require strategic planning over a long sequence of states and actions to achieve complex targets (Wang et al., 2023c; He et al., 2024). For example, in the negotiation dialogue setting (He et al., 2018), the system should strategically adjust its bid at each

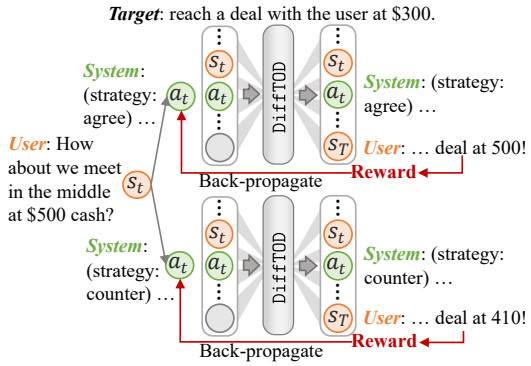

Figure 4: An illustration of the search-based guidance.

turn in order to reach the best deal with the user. To achieve such targets, we propose a search-based guidance mechanism that enables strategic and non-myopic dialogue planning over a long horizon. Specifically, leveraging the definition of state s_t and action a_t within the conversational MDP framework, we condition the diffusion model to generate different actions a_t at each turn t with either the word-level or the semantic-level guidance. To explore how the future dialogue will unfold if action a_t is taken, we build a conversational search tree where each node represents a possible state in the dialogue, and the tree branches as different actions are explored at each turn. By applying a search and planning algorithm (Kocsis and Szepesvári, 2006; Coulom, 2007) over the tree, search-based guidance ensures that the generated dialogue plans can maximize the cumulative reward and effectively achieve the target. An illustration of the search-based guidance is presented in Figure 4, and the detailed algorithm is in Appendix B.

4.3 Discussion

In this subsection, we summarize the desirable properties of DiffTOD that are helpful for dialogue planning and achieving TOD targets.

Non-Sequential Dialogue Planning DiffTOD enables non-sequential dialogue planning by generating the entire trajectory simultaneously instead of step by step. This allows it to anticipate how future conversations might unfold and strategically plan the dialogue toward the target. As a result, it can optimize action strategies for overall target achievement and maintain global consistency (Section 6.1, 6.2), rather than getting stuck in myopic decisions.

Flexible Guidance By decomposing the trajectory generation process into an unconditional and a conditional part, DiffTOD allows for flexible guidance at test time. This design allows DiffTOD to adapt to different targets without re-training, while policy-based methods have to be re-trained for each new target (Section 6.1).

Tackling Sparse Reward Sequential dialogue planning methods often face challenges in TOD settings with a sparse reward, where feedback is provided only at the end of the conversation (Feng et al., 2023). In contrast, the non-sequential nature of DiffTOD allows for the generation of a globally consistent dialogue plan conditioned on the target state. By formulating dialogue planning as a trajectory inpainting process, DiffTOD ensures that the generated dialogue plan can effectively guide the conversation toward the target (Section 6.2).

Modeling Long-Range Dependency Leveraging the long context length of the diffusion language model, DiffTOD plans actions by considering both past and future dialogue states over a long horizon. This allows it to model complex, long-range dependency and optimize dialogue strategies for overall target achievement (Section 6.3).

5 Experimental Settings

5.1 Datasets

To demonstrate the effectiveness and the flexibility of DiffTOD, we use three datasets with different dialogue targets and settings for evaluation. The first dataset, CraigslistBargain (He et al., 2018), is a dialogue negotiation dataset where buyers and sellers bargain. Since the buyers and sellers pursue different targets in the dialogue, we consider two evaluation settings for two different targets: (1) the system acts as the buyer and negotiates for the lowest possible price with the seller; (2) the system acts as the seller and negotiates for the highest possible price with the buyer. The second dataset, TopDial (Wang et al., 2023b), is a personalized conversational recommendation dataset with dialogues on different topics, such as movie, food and music. The target of the system is to recommend a specified item to the user. The third dataset, PersonaChat (Zhang et al., 2018), is collected from an open-domain chitchat setting. Tang et al. (2019) extracts keywords from each turn in this dataset as targets. In this work, we introduce a simple yet challenging setting where the target of the system is to direct the conversation toward mentioning a list of specified keywords in the exact given order. The statistics of all the datasets are in Appendix C.

5.2 Evaluation Protocols

Our evaluations primarily focus on measuring the successful achievement of TOD targets. For the CraigslistBargain dataset, we follow the literature (Deng et al., 2024a,b) and adopt Success Rate (**SR**) to measure the ratio of successful deals within 10 turns; Average Turn (**AT**) to measure the efficiency of target completion by calculating the average number of turns required to achieve the target; and Sell-to-List Ratio (**SLR**) (Zhou et al., 2019) to measure how much benefit the buyer or seller gets compared with the initial listing price, as is detailed in Appendix D. For the PersonaChat dataset, we adopt Keyword Coverage Ratio (**KCR**) to measure the percentage of the specified keywords that are mentioned in the dialogue; and the edit distance (**Dist.**) between the target keyword list and the sequence of keywords mentioned in the dialogue to measure how well the conversation follows the specified keyword order. For the TopDial dataset, we adopt Success Rate (**SR**) to measure the ratio of successful recommendations within 10 turns, and Average Turn (**AT**) to measure the average number of turns. To provide a dynamic and interactive evaluation environment, we follow the literature (Dao et al., 2024; Deng et al., 2024b) and prompt an LLM as the user simulator. The system will chat interactively with the user simulator for multiple turns until either the target is achieved or a maximum of 10 conversational turns is reached.

Besides target achievement, for the Topdial and PersonaChat datasets, we also evaluate the text quality of the dialogue plan by comparing it with the ground-truth dialogue in the test set using reference-based metrics, including **BLEU** (Papineni et al., 2002), word-level F1 (**F1**) and BERT Score (Zhang et al., 2020) (**Score**).

Moreover, as previous researches have demonstrated the effectiveness of LLMs in dialogue evaluation and their strong correlation with human judgments (Zheng et al., 2023; Wang et al., 2023a; Fu et al., 2024), we utilize the state-of-the-art LLM, GPT-4o (Hurst et al., 2024), to provide an overall evaluation score (**Ovr.**) of the dialogue quality on a scale of 1 to 5. We prompt GPT-4o to provide an overall evaluation based on various criteria such as coherence, helpfulness, appropriateness, and target achievement, as detailed in Appendix F.2.

Finally, to ensure the reliability of our evaluation results, we provide human evaluations on the CraigslistBargain dataset in Appendix H. Our results show that both human evaluators and GPT-4o consistently rate `DiffTOD` as superior, and the average disagreement between GPT-4o and human evaluators is low, thus validating the reliability of our LLM-based dialogue evaluation protocol.

5.3 Baselines

We compare our approach against (1) fine-tuning the latest versions of popular open-source LLMs, including LLAMA-3-8B (Dubey et al., 2024) and Mistral-8B (MistralAI, 2024); (2) state-of-the-art closed-source LLMs, including GPT-4o (Hurst et al., 2024) and Claude-3.5 (Anthropic, 2024); and (3) LLM-based dialogue planning methods for TOD, including ProCoT (Deng et al., 2023b) and EnPL (Zheng et al., 2024). Since each response in the CraigslistBargain dataset is annotated with an action type, we also include a task-specific baseline, PPDPP (Deng et al., 2024b), for this dataset, which supports policy learning over action types.

5.4 Implementation Details

For the CraigslistBargain dataset, we adopt the search-based guidance to select the best dialogue plan, and use the word-level guidance to control the types of actions generated at each turn. To demonstrate the flexibility of `DiffTOD`, we fine-tune the diffusion model with the same training data, and then apply different reward functions as guidance that measure the benefit of the buyer and the seller respectively, to achieve different targets in each setting. For the PersonaChat dataset, we adopt the word-level guidance to direct the diffusion model to generate a dialogue plan with the specified keywords. For the TopDial dataset, we adopt the semantic-level guidance by prompting GPT-4o to generate 5 paraphrased versions of the target state (i.e., system recommends the target item) with the same semantic meaning. More implementation details are in Appendix E.

6 Experimental Results

6.1 Negotiation Dialogue

Table 1 presents the results on the CraigslistBargain dataset. We have the following observations:

(1) `DiffTOD` ***achieves consistent improvement over baselines in terms of all the metrics measuring the target achievement success.*** Different from baseline methods that generate states and actions in the dialogue plan sequentially, `DiffTOD` adopts a diffusion model to generate the entire dialogue

Model	As Buyer				As Seller			
	SR↑	AT↓	SLR↑	Ovr.↑	SR↑	AT↓	SLR↑	Ovr.↑
LLAMA3	0.426	7.49	0.156	3.33	0.516	6.61	0.134	3.70
Mistral	0.561	6.49	0.347	3.59	0.589	6.21	0.188	3.82
GPT-4o	0.479	6.95	0.292	3.73	0.409	7.66	0.174	3.85
Claude-3.5	0.798	4.75	0.254	3.50	0.582	6.41	0.260	3.82
ProCoT	0.456	6.95	0.228	3.42	0.689	6.38	0.298	4.01
EnPL	0.644	6.23	0.382	4.21	0.627	5.99	0.228	3.62
PPDPP	0.665	5.57	0.338	3.59	0.633	4.99	0.347	3.61
DiffTOD	**0.872**	**3.98**	**0.565**	**4.36**	**0.729**	**4.61**	**0.361**	**4.05**

Table 1: Performance comparison between our approach and baselines on the CraigslistBargains dataset. SR, AT, SLR and Ovr. represent success rate, average turn, sale-to-list ratio, and overall dialogue quality, respectively.

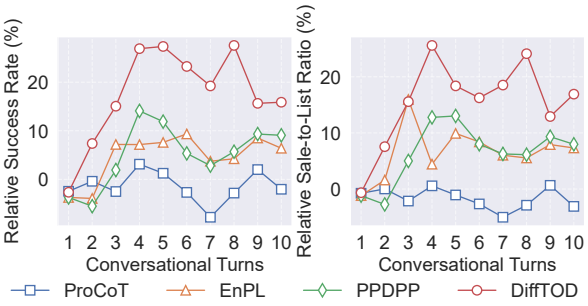

Figure 5: Comparisons of the relative success rate and sale-to-list-ratio against GPT-4o at each conversational turn on the CraigslistBargain dataset.

Model	SR↑	AT↓	BLEU↑	F1↑	Score↑	Ovr.↑
LLAMA3	0.680	5.89	0.104	0.173	0.876	3.32
Mistral	0.673	5.90	0.121	**0.187**	**0.886**	**3.53**
GPT-4o	0.640	6.01	0.046	0.140	0.819	3.39
Claude-3.5	0.633	5.73	0.029	0.165	0.845	3.52
ProCoT	0.688	5.57	0.030	0.153	0.834	3.35
EnPL	0.659	5.81	0.018	0.169	0.826	3.51
DiffTOD	**0.713**	**5.31**	**0.160**	0.168	0.870	3.38

Table 2: Performance comparison between our approach and baselines on TopDial dataset. SR, AT, F1, Score and Ovr. represent success rate, average turn, word-level F1, BERT Score and overall dialogue quality, respectively.

plan simultaneously. This non-sequential approach allows DiffTOD to plan dialogues for overall target achievement and global consistency, rather than being constrained by locally optimal actions that maximize immediate rewards but may undermine overall target achievement (Janner et al., 2022; He et al., 2023a). Using the search-based guidance, DiffTOD can effectively plan dialogue actions that successfully achieve the target in the fewest possible conversational turns.

(2) DiffTOD ***demonstrates strong flexibility and outperforms baselines in both buyer and seller settings.*** DiffTOD allows for flexible guidance that can be tailored to achieve different targets at test time. By applying customized guidance strategies to maximize buyer or seller benefits respectively, DiffTOD fine-tuned on the same dialogue history data can achieve consistent improvement in both settings, even compared with baselines that are individually tuned and optimized for each setting.

To further validate the effectiveness of DiffTOD, we visualize the relative Success Rate (SR) and Sale-to-List-Ratio (SLR) of different dialogue planning methods against GPT-4o at each turn. The experimental results in Figure 5 show that DiffTOD consistently outperforms other methods at almost every turn. Notably, DiffTOD consistently achieves a higher SR and SLR than other methods as the conversational turn increases. This demonstrates the effectiveness of DiffTOD in dialogue planning, particularly in complex situations that require lengthy, multi-turn negotiations.

6.2 Conversational Recommendation

Table 2 presents the results on the TopDial dataset. We have the following observations:

(1) DiffTOD ***outperforms baselines in terms of success rate and average turn.*** The non-sequential nature of DiffTOD allows generating a globally consistent dialogue plan conditioned on the target state, ensuring that the dialogue plan can successfully achieve the target by the end of the conversation. In the sparse reward setting where only a final reward is provided upon successful recommendation of the target item, sequential planning methods often struggle due to the absence of intermediate reward signals (Andrychowicz et al., 2017; Rengarajan et al., 2022).

(2) DiffTOD ***performs on par with the baselines in terms of text generation quality.*** When evaluating the generated dialogue plan against the ground-truth dialogue plan with reference-based metrics, DiffTOD performs on par with the baselines. A similar trend is observed in the assessment of overall dialogue quality. This suggests that the diffusion language model is capable of generating dialogue plans that can achieve the target without significantly compromising text generation quality.

6.3 Open-Domain Chitchat

Table 3 summarizes the experimental results on the PersonaChat dataset. We observe that DiffTOD ***achieves a consistent improvement over baselines***

Model	KCR↑	Dist.↓	BLEU↑	F1↑	Score↑	Ovr.↑
LLAMA3	0.606	23.68	0.015	0.159	0.830	3.79
Mistral	0.842	24.15	0.181	0.178	0.896	**4.19**
GPT-4o	0.685	23.50	0.010	0.133	0.815	3.94
Claude-3.5	0.776	23.35	0.015	0.156	0.829	3.97
ProCoT	0.634	21.99	0.038	0.170	0.826	3.64
EnPL	0.706	23.19	0.038	0.161	0.828	3.99
DiffTOD	**0.845**	**20.95**	**0.298**	**0.182**	**0.897**	4.05

Table 3: Performance comparison on PersonaChat dataset. KCR, Dist., F1, Score and Ovr. represent keyword coverage ratio, edit distance, word-level F1, BERT Score and overall dialogue quality, respectively.

in keyword coverage ratio and more faithfully maintains the specified order of keywords. The target for the PersonaChat dataset requires the system to not only incorporate all specified keywords in the dialogue plan, but also to decide the appropriate ordering of keyword mentions between the user and the system over multiple turns. This creates a complex, long-range dependency where keyword transitions depend on each other, and earlier mention of later keywords can disrupt the order. Sequential dialogue planning methods often struggle with such targets due to cumulative errors and lack of foresight (Ke et al., 2019; Kumar et al., 2019). In our experiments, we observe that they tend to forget keywords and violate the constraint of order as the number of turns increases. By contrast, DiffTOD can effectively handle these complex dependencies by leveraging the word-level guidance to enforce both keyword coverage and mentioning order. This enables DiffTOD to effectively achieve the target and demonstrate superior planning capabilities in dialogue settings with complex targets.

7 Conclusion

We present DiffTOD, a novel non-sequential dialogue planning framework that enables non-myopic lookahead exploration and optimizes action strategies for overall target achievement. DiffTOD models the likelihood of the dialogue trajectory with a diffusion language model. To optimize the action strategies, DiffTOD decomposes the trajectory generation process into an unconditional and a conditional part and introduces three guidance mechanisms tailored to different target types for flexible test-time guidance. Extensive experiments demonstrate that DiffTOD outperforms baselines on target achievement success and shows strong flexibility across complex and diverse dialogue scenarios.

8 Limitations

Dynamic Adjustment of Dialogue Plans Our framework leverages a diffusion model to simulate the transitions between the states and actions. As a model-based planning approach, there is a possibility that the simulated environment may not perfectly align with real-world conversations, leading to discrepancies between the generated dialogue plans and the actual conversations. To address this issue, future work may introduce replanning techniques (Zhou et al., 2023) that can dynamically adjust the dialogue plan when the actual conversation diverges from the original dialogue plan.

Inference Cost Since diffusion models require iterative denoising over multiple diffusion steps, they may incur higher computational costs compared to standard autoregressive decoding. To better understand and quantify the inference cost, we present additional analysis on the inference cost of the diffusion models from both an empirical and a theoretical perspective in Appendix J, and our results show that future work may introduce acceleration sampling techniques for diffusion models (Shih et al., 2023; Ma et al., 2024) into our framework to reduce the inference cost.

Evaluation Quality Our evaluation protocol is based on simulated conversations between LLMs of different roles. Although such an LLM-based evaluation protocol has been widely adopted in dialogue systems (Deng et al., 2024b; He et al., 2024; Li et al., 2024) and demonstrates a strong correlation with human judgments (Zheng et al., 2023; Wang et al., 2023a; Fu et al., 2024), engaging in conversations with real users can provide a more accurate and trustworthy evaluation of dialogue quality. To ensure the reliability of our evaluation results, we provide human evaluations on sampled test cases from the CraigslistBargain dataset in Appendix H, but due to limited resources, we are unable to perform human evaluation at a larger scale.

9 Ethics Statement

Our work aims to improve the planning capability of TOD systems to better assist users in achieving a variety of targets in the dialogue. While our framework is not designed for unethical usage, there is often a potential risk of the misuse of such systems by modifying the target for unintended or unethical purposes. We strongly oppose any unlawful or unjust usage of our framework.

All the datasets used in this research are from public open-access datasets, which do not contain sensitive or private information.

10 Acknowledgement

The authors thank Reza Averly, Frazier N. Baker, Vishal Dey, Ruoxi Gao and Xiao Hu for their valuable assistance with data annotation, and Xinyi Ling for refining the figures in this paper. The authors also thank the anonymous reviewers for their insightful comments and constructive feedback.

References

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. 2017. Hindsight experience replay. In *Advances in Neural Information Processing Systems*, volume 30.

Anthropic. 2024. Claude.ai.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, volume 34, pages 17981–17993.

Xuefeng Bai, Yulong Chen, Linfeng Song, and Yue Zhang. 2021. Semantic representation for dialogue modeling. 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 4430–4445.

Richard Bellman. 1957. A Markovian decision process. *Journal of Mathematics and Mechanics*, pages 679–684.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. 2024. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *The Twelfth International Conference on Learning Representations*.

Haoxuan Chen, Yinuo Ren, Lexing Ying, and Grant M. Rotskoff. 2024. Accelerating diffusion models with parallel sampling: Inference at sub-linear time complexity. In *Advances in Neural Information Processing Systems*, volume 37, pages 133661–133709.

Ziqi Chen, Bo Peng, Tianhua Zhai, Daniel Adu-Ampratwum, and Xia Ning. 2025. Generating 3D small binding molecules using shape-conditioned diffusion models with guidance. *Nature Machine Intelligence*, pages 1–13.

Rémi Coulom. 2007. Efficient selectivity and backup operators in Monte-Carlo tree search. In *International Conference on Computers and Games*, pages 72–83. Springer.

Huy Quang Dao, Yang Deng, Khanh-Huyen Bui, Dung D. Le, and Lizi Liao. 2024. Experience as source for anticipation and planning: Experiential policy learning for target-driven recommendation dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14179–14198.

Yang Deng, Wenqiang Lei, Wai Lam, and Tat-Seng Chua. 2023a. A survey on proactive dialogue systems: problems, methods, and prospects. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*.

Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023b. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10602–10621.

Yang Deng, Lizi Liao, Wenqiang Lei, Grace Yang, Wai Lam, and Tat-Seng Chua. 2025. Proactive conversational AI: A comprehensive survey of advancements and opportunities. *ACM Trans. Inf. Syst.*

Yang Deng, Lizi Liao, Zhonghua Zheng, Grace Hui Yang, and Tat-Seng Chua. 2024a. Towards human-centered proactive conversational agents. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 807–818.

Yang Deng, Wenxuan Zhang, Wai Lam, See-Kiong Ng, and Tat-Seng Chua. 2024b. Plug-and-play policy planner for large language model powered dialogue agents. In *The Twelfth International Conference on Learning Representations*.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yihao Feng, Shentao Yang, Shujian Zhang, Jianguo Zhang, Caiming Xiong, Mingyuan Zhou, and Huan Wang. 2023. Fantastic rewards and how to tame them: A case study on reward learning for task-oriented dialogue systems. In *The Eleventh International Conference on Learning Representations*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576.