

# Planning with Diffusion Models for Target-Oriented Dialogue Systems

Hanwen Du<sup>\*</sup> Bo Peng<sup>\*</sup> Xia Ning<sup>\*†♥</sup>✉

<sup>\*</sup>Department of Computer Science and Engineering, The Ohio State University, USA

<sup>†</sup>Department of Biomedical Informatics, The Ohio State University, USA

<sup>♥</sup>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).

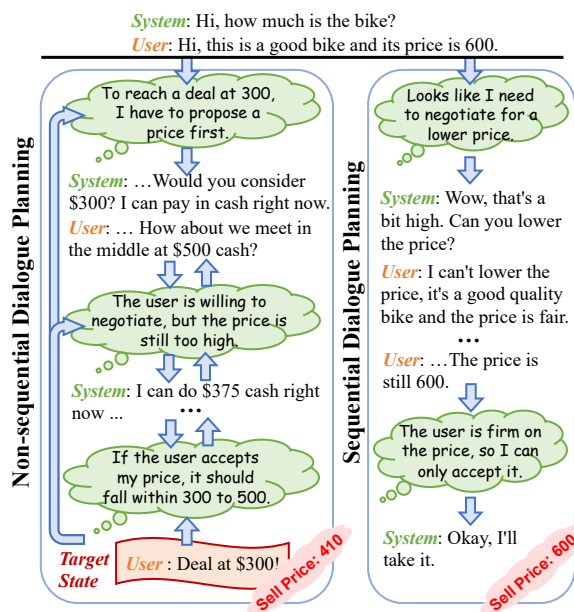


Figure 1: Using sequential and non-sequential dialogue planning methods for negotiation dialogues.

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

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 substan-

tially 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:

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

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;  $\mathcal{T} : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$  denotes the transition to the next state after taking an action from the current state;  $\mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  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  $s_t = (d_0^s, d_0^u, \dots, d_{t-1}^s, d_{t-1}^u)$  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  $\mathcal{T} : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$  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  $d_t^u$ , and the next state  $s_{t+1} = (d_0^s, d_0^u, \dots, d_{t-1}^s, d_{t-1}^u, d_t^s, d_t^u)$  is updated to include the user’s and the system’s responses at the current conversational turn  $t$ .

**Reward** The reward function  $\mathcal{R} : r(s, a) \rightarrow \mathbb{R}$  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  $\tau_{0:t} = (s_0, a_0, \dots, s_t, a_t)$  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:

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

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:

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

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:

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

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  $\tau^n$  representing a partially observed trajectory,  $\tau^{n-1}$  representing a more complete trajectory,  $\tau^N$  representing the trajectory with only the initial state and action, and  $\tau^0$  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):

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

Using this formulation, we can train a generative diffusion model  $p_{\theta}$  that can reconstruct the entire trajectory  $\tau^0$  from an incomplete trajectory  $\tau^n$ . In this way, the diffusion model  $p_{\theta}$  can function both as the *planner* that generates actions  $a_t$  when  $\tau^n = \{s_0, a_0, \dots, s_{t-1}, a_{t-1}, s_t\}$ , and the *environment* that generates  $s_{t+1}$  when  $\tau^n = \{s_0, a_0, \dots, s_t, a_t\}$ . 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  $p_{\theta}$ . 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  $\tau_{0:t} = \{d_0^s, d_0^u, \dots, d_{t-1}^s, d_{t-1}^u, d_t^s\}$  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  $\tau^n$  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

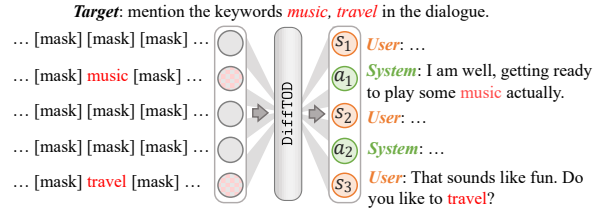


Figure 2: An illustration of the word-level guidance.

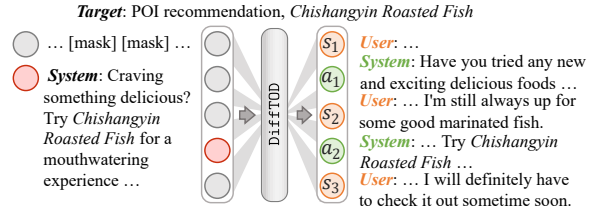


Figure 3: An illustration of the semantic-level guidance.

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.

**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

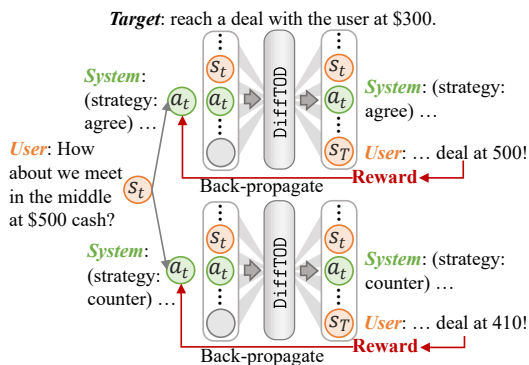


Figure 4: An illustration of the search-based guidance.

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 guid-

ance 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 re-

sults 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 $\uparrow$	AT $\downarrow$	SLR $\uparrow$	Ovr. $\uparrow$	SR $\uparrow$	AT $\downarrow$	SLR $\uparrow$	Ovr. $\uparrow$
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	<b>0.872</b>	<b>3.98</b>	<b>0.565</b>	<b>4.36</b>	<b>0.729</b>	<b>4.61</b>	<b>0.361</b>	<b>4.05</b>

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.

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*

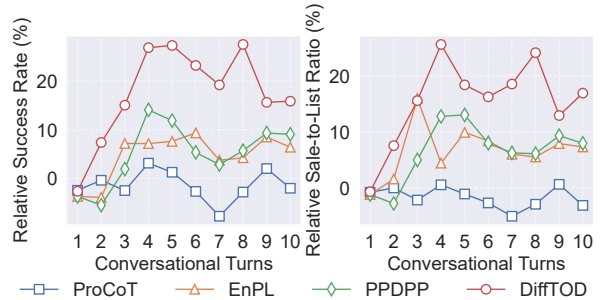


Figure 5: Comparisons of the relative success rate and sale-to-list-ratio against GPT-4o at each conversational turn on the CraigslistBargain dataset.

Model	SR $\uparrow$	AT $\downarrow$	BLEU $\uparrow$	F1 $\uparrow$	Score $\uparrow$	Ovr. $\uparrow$
LLAMA3	0.680	5.89	0.104	0.173	0.876	3.32
Mistral	0.673	5.90	0.121	<b>0.187</b>	<b>0.886</b>	<b>3.53</b>
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	<b>0.713</b>	<b>5.31</b>	<b>0.160</b>	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.

*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 $\uparrow$	Dist. $\downarrow$	BLEU $\uparrow$	F1 $\uparrow$	Score $\uparrow$	Ovr. $\uparrow$
LLAMA3	0.606	23.68	0.015	0.159	0.830	3.79
Mistral	0.842	24.15	0.181	0.178	0.896	<b>4.19</b>
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	<b>0.845</b>	<b>20.95</b>	<b>0.298</b>	<b>0.182</b>	<b>0.897</b>	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.

- Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. 2023. [DiffuSeq: Sequence to sequence text generation with diffusion models](#). In *The Eleventh International Conference on Learning Representations*.
- Nate Gruver, Samuel Stanton, Nathan Frey, Tim G. J. Rudner, Isidro Hotzel, Julien Lafrance-Vanasse, Arvind Rajpal, Kyunghyun Cho, and Andrew G Wilson. 2023. [Protein design with guided discrete diffusion](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 12489–12517.
- Ishaan Gulrajani and Tatsunori B Hashimoto. 2024. [Likelihood-based diffusion language models](#). *Advances in Neural Information Processing Systems*, 36.
- Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. 2023. [Reasoning with language model is planning with world model](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8154–8173.
- Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 2024. [Training large language models to reason in a continuous latent space](#). *arXiv preprint arXiv:2412.06769*.
- Haoran He, Chenjia Bai, Kang Xu, Zhuoran Yang, Weinan Zhang, Dong Wang, Bin Zhao, and Xuelong Li. 2023a. [Diffusion model is an effective planner and data synthesizer for multi-task reinforcement learning](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 64896–64917.
- He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. [Decoupling strategy and generation in negotiation dialogues](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2333–2343.
- Tao He, Lizi Liao, Yixin Cao, Yuanxing Liu, Ming Liu, Zerui Chen, and Bing Qin. 2024. [Planning like human: A dual-process framework for dialogue planning](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4768–4791.
- Zhengfu He, Tianxiang Sun, Qiong Tang, Kuanning Wang, Xuanjing Huang, and Xipeng Qiu. 2023b. [DiffusionBERT: Improving generative masked language models with diffusion models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4521–4534.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. [Denoising diffusion probabilistic models](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. [GPT-4o system card](#). *arXiv preprint arXiv:2410.21276*.
- Michael Janner, Yilun Du, Joshua Tenenbaum, and Sergey Levine. 2022. [Planning with diffusion for flexible behavior synthesis](#). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 9902–9915. PMLR.
- Nan Jiang, Alex Kulesza, Satinder Singh, and Richard Lewis. 2016. [The dependence of effective planning horizon on model accuracy](#). In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, page 4180–4184.
- Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Paul Saldyt, and Anil B Murthy. 2024. [Position: LLMs can’t plan, but can help planning in LLM-modulo frameworks](#). In *Forty-first International Conference on Machine Learning*.
- Nan Rosemary Ke, Amanpreet Singh, Ahmed Touati, Anirudh Goyal, Yoshua Bengio, Devi Parikh, and Dhruv Batra. 2019. [Modeling the long term future in model-based reinforcement learning](#). In *International Conference on Learning Representations*.
- Levente Kocsis and Csaba Szepesvári. 2006. [Bandit based Monte-Carlo planning](#). In *European Conference on Machine Learning*, pages 282–293. Springer.
- Philipp Koehn. 2004. [Statistical significance tests for machine translation evaluation](#). In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 388–395.
- Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. 2019. [Stabilizing off-policy Q-learning via bootstrapping error reduction](#). In *Advances in Neural Information Processing Systems*, volume 32.
- Wai-Chung Kwan, Hong-Ru Wang, Hui-Min Wang, and Kam-Fai Wong. 2023. [A survey on recent advances and challenges in reinforcement learning methods for task-oriented dialogue policy learning](#). *Machine Intelligence Research*, 20(3):318–334.
- Sergey Levine. 2018. [Reinforcement learning and control as probabilistic inference: Tutorial and review](#). *arXiv preprint arXiv:1805.00909*.
- Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, Yuxuan Lai, Chongyang Tao, and Shuai Ma. 2024. [Leveraging large language models for NLG evaluation: Advances and challenges](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16028–16045.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.

- Aaron Lou, Chenlin Meng, and Stefano Ermon. 2024. [Discrete diffusion modeling by estimating the ratios of the data distribution](#). In *Forty-first International Conference on Machine Learning*.
- Chang Ma, Haiteng Zhao, Junlei Zhang, Junxian He, and Lingpeng Kong. 2025. [Non-myopic generation of language models for reasoning and planning](#). In *The Thirteenth International Conference on Learning Representations*.
- Xinyin Ma, Gongfan Fang, and Xinchao Wang. 2024. [DeepCache: Accelerating diffusion models for free](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15762–15772.
- MistralAI. 2024. [Ministral-8b](#).
- Jiao Ou, Junda Lu, Che Liu, Yihong Tang, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. [DialogBench: Evaluating LLMs as human-like dialogue systems](#). 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 6137–6170.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [BLEU: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Libo Qin, Wenbo Pan, Qiguang Chen, Lizi Liao, Zhou Yu, Yue Zhang, Wanxiang Che, and Min Li. 2023. [End-to-end task-oriented dialogue: A survey of tasks, methods, and future directions](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5925–5941.
- Desik Rengarajan, Gargi Vaidya, Akshay Sarvesh, Dileep Kalathil, and Srinivas Shakkottai. 2022. [Reinforcement learning with sparse rewards using guidance from offline demonstration](#). In *International Conference on Learning Representations*.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. [High-resolution image synthesis with latent diffusion models](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.
- Andy Shih, Suneel Belkhale, Stefano Ermon, Dorsa Sadigh, and Nima Anari. 2023. [Parallel sampling of diffusion models](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 4263–4276.
- Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. [Deep unsupervised learning using nonequilibrium thermodynamics](#). In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 2256–2265.
- Yang Song and Stefano Ermon. 2019. [Generative modeling by estimating gradients of the data distribution](#). In *Advances in Neural Information Processing Systems*, volume 32.
- Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric Xing, and Zhiting Hu. 2019. [Target-guided open-domain conversation](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5624–5634.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30.
- Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. [Is ChatGPT a good NLG evaluator? A preliminary study](#). In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11.
- Jian Wang, Yi Cheng, Dongding Lin, Chak Tou Leong, and Wenjie Li. 2023b. [Target-oriented proactive dialogue systems with personalization: Problem formulation and dataset curation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1132–1143.
- Jian Wang, Dongding Lin, and Wenjie Li. 2023c. [Dialogue planning via brownian bridge stochastic process for goal-directed proactive dialogue](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 370–387.
- Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. 2023. [De novo design of protein structure and function with RFdiffusion](#). *Nature*, 620(7976):1089–1100.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.
- Zhitong Yang, Bo Wang, Jinfeng Zhou, Yue Tan, Dongming Zhao, Kun Huang, Ruifang He, and Yuexian Hou. 2022. [TopKG: Target-oriented dialog via](#)

global planning on knowledge graph. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 745–755.

Jiacheng Ye, Jiahui Gao, Shansan Gong, Lin Zheng, Xin Jiang, Zhenguo Li, and Lingpeng Kong. 2025. **Beyond autoregression: Discrete diffusion for complex reasoning and planning.** In *The Thirteenth International Conference on Learning Representations*.

Qiang Zhang, Jason Naradowsky, and Yusuke Miyao. 2023a. **Ask an expert: Leveraging language models to improve strategic reasoning in goal-oriented dialogue models.** In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6665–6694.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. **Personalizing dialogue agents: I have a dog, do you have pets too?** In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. **BERTScore: Evaluating text generation with BERT.** In *International Conference on Learning Representations*.

Yizhe Zhang, Jiatao Gu, Zhuofeng Wu, Shuangfei Zhai, Joshua Susskind, and Navdeep Jaitly. 2023b. **Planner: Generating diversified paragraph via latent language diffusion model.** In *Advances in Neural Information Processing Systems*, volume 36, pages 80178–80190.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhonghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. **Judging LLM-as-a-Judge with MT-Bench and chatbot arena.** In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623.

Zhonghua Zheng, Lizi Liao, Yang Deng, Ee-Peng Lim, Minlie Huang, and Liqiang Nie. 2024. **Thoughts to target: Enhance planning for target-driven conversation.** In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21108–21124.

Peixiang Zhong, Yong Liu, Hao Wang, and Chunyan Miao. 2021. **Keyword-guided neural conversational model.** In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14568–14576.

Siyuan Zhou, Yilun Du, Shun Zhang, Mengdi Xu, Yikang Shen, Wei Xiao, Dit-Yan Yeung, and Chuang Gan. 2023. **Adaptive online replanning with diffusion models.** In *Advances in Neural Information Processing Systems*, volume 36, pages 44000–44016.

Yiheng Zhou, He He, Alan W Black, and Yulia Tsvetkov. 2019. **A dynamic strategy coach for effective negotiation.** In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 367–378.

## A Illustration of the State and Action Calculation

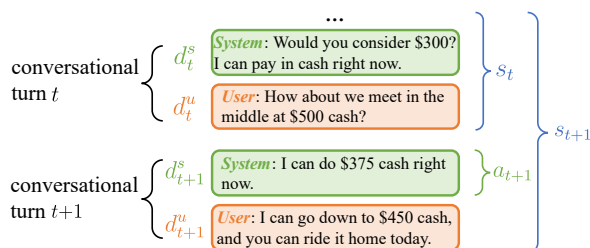


Figure A1: Examples of the state and action calculation in the conversational MDP.

## B Details on Search-Based Guidance

The search-based guidance allows for the integration of any planning and search algorithm within the tree structure. In our implementation, we integrate search-based guidance with Monte Carlo Tree Search (MCTS) (Kocsis and Szepesvári, 2006; Coulom, 2007), a principled planning algorithm that can strategically balance between exploration and exploitation to efficiently expand a search tree. Specifically, each node in the tree represents a state  $s_t$ , and the root node  $s_0$  represents the initial state at the start of the conversation. Each edge between the node  $s_t$  and the node  $s_{t+1}$  represents an action  $a_t$  and the transition from the current state  $s_t$  to the next state  $s_{t+1}$  after taking the action  $a_t$ . For each action  $a_t$ , search-based guidance maintains a function  $Q(s_t, a_t)$  as the expected future reward of choosing the action  $a_t$  in the state  $s_t$ . Starting from an initial state  $s_0$ , search-based guidance builds a conversational search tree in four stages, namely, trajectory selection, node expansion, conversation simulation, and reward back-propagation. This iterative process will continue until a specified number of iterations is reached. After all the simulations are complete, we select the best action  $a_t^* \leftarrow \operatorname{argmax}_{a \in \mathcal{A}(s_t)} Q(s_t, a_t)$  along the trajectory with the highest future reward and generate a response based on the  $a_t^*$ . The details of the four stages are described below, and we summarize the algorithm of search-based guidance in Algorithm 1.

**Trajectory Selection** The search-based guidance selects the most promising trajectory that is likely to achieve a high future reward  $Q(s_t, a_t)$  by traversing from the root node to a leaf node within the current tree. The selected trajectory will then be expanded for dialogue simulation at the next stage. To achieve a balance between the explo-

ration of less-visited nodes, and the exploitation of highly-rewarded nodes within the current tree, search-based guidance adopts the Upper Confidence bounds applied to Trees (UCT) (Kocsis and Szepesvári, 2006) approach to select a child node  $a_t^*$  into the current trajectory:

$$a_t^* \leftarrow \operatorname{argmax}_{a_t \in \mathcal{A}(s_t)} \left[ Q(s_t, a_t) + w \sqrt{\frac{\log N(s_t)}{N(c(s_t, a_t))}} \right], \quad (8)$$

where  $\mathcal{A}(s_t)$  denotes the set of valid actions in the state  $s_t$ ,  $N(s_t)$  denotes the visit count of node  $s_t$ , and  $c(s_t, a_t)$  denotes the child node of  $s_t$  after selecting the action  $a_t$ . The weight  $w$  controls the balance between exploration and exploitation. Intuitively, the second term in Equation 8 will be larger if the child node is less visited, thereby encouraging for more exploration.

**Node Expansion** After reaching a leaf node  $s_t$ , the search-base guidance expands the leaf node by randomly picking an action  $a_t$  to explore and attaching a new child  $s_{t+1}$  to the node  $s_t$ .

**Conversation Simulation** Given the selected trajectory  $\tau_{0:t} = \{s_0, a_0, \dots, s_t, a_t\}$  up to and including the  $t$ -th turn, the diffusion model generates the rest part of the trajectory  $\tau_{t+1:T} = \{s_{t+1}, \dots, s_T, a_T\}$ , conditioned on  $\tau_{0:T}$ , to predict how the future conversation will unfold. To control the type of action  $a_t$  generated at the current turn, we use the word-level guidance for the Craigslist dataset by prepending the keyword of the dialogue action strategy to the system response  $d_t^s$ .

**Reward Back-Propagation** Once the conversation simulation is complete, we obtain a complete trajectory  $\tau_{0:T}$  from the turn 0 to the turn  $T$ . We then update  $Q(s_t, a_t)$  by aggregating the immediate reward from turn  $t$  to turn  $T$ .

## C Statistics of Datasets

The statistics of the datasets after preprocessing are presented in Table A1.

## D Details of Sell-to-List Ratio Calculation

The Sell-to-List ratio (Zhou et al., 2019) measures the amount of benefit the buyer or the seller receives compared to the initial listing price. From the buyer’s side, since the target is to negotiate for the lowest possible price with the seller, the

---

### Algorithm 1 Search-Based Guidance

---

**Require:** conversational MDP  $\mathcal{M}$ , initial state  $s_0$ , simulation budget  $K$ , exploration factor  $w$

**for**  $k \leftarrow 1, \dots, K$  **do**

$t \leftarrow 0$

**while**  $s_t$  has children **do** ▷ Trajectory Selection

Select an action  $a_t$  (Eq. 8)

$s_{t+1} \leftarrow \mathcal{T}(s_t, a_t), r_t \leftarrow r(s_t, a_t), t \leftarrow t + 1$

**end while** ▷ Node Expansion

At leaf node  $s_t$ , pick an action  $a_t$  to explore ▷ Conversation Simulation

Sample  $s_{t+1}, a_{t+1}, \dots, s_T, a_T$  with the diffusion model  $p_\theta$  conditioned on  $\tau_{0:t}$ , obtain reward  $r_t$  based on the simulated outcome of the conversation

**while**  $t \geq 0$  **do** ▷ Reward Back-Propagation

Update  $Q(s_t, a_t)$  by aggregating  $\{r_{t'}\}_{t'=t}^T$ ,

$N(s_t) \leftarrow N(s_t) + 1, t \leftarrow t - 1$

**end while**

**end for**

---

sell-to-list is calculated as follows:

$$\text{SLR} = \frac{\text{deal price} - \text{seller target price}}{\text{buyer target price} - \text{seller target price}}. \quad (9)$$

From the seller’s side, since the target is to negotiate for the highest possible price with the buyer, the sell-to-list is calculated as follows:

$$\text{SLR} = \frac{\text{buyer target price} - \text{deal price}}{\text{buyer target price} - \text{seller target price}}. \quad (10)$$

If the dialogue fails to reach a deal, we set the sell-to-list ratio as 0.

## E Additional Implementation Details

We fine-tune the state-of-the-art masked diffusion language model, SEDD-medium (Lou et al., 2024), on the dialogue history from the training data. SEDD-medium is a GPT-2 sized masked diffusion language model with about 424M parameters and 1024 context length. We set the batch size is set as 32, learning rate as 1e-4, and use the AdamW optimizer (Loshchilov and Hutter, 2019). For the search-based guidance, we set the exploration factor  $w = 1.5$ , and the simulation budget  $K = 10$  to achieve a balance between efficiency and performance. All the fine-tuning and inference experiments can be done on a single A100 GPU.

To provide a dynamic and interactive environment for dialogue evaluation, following the literature (Wang et al., 2023c; Deng et al., 2024b; Dao et al., 2024), we prompt an LLM to role-play as the system that can generate a response at each turn by strictly following the action strategies in the dialogue plan. We then prompt another LLM as the

Dataset	CraigslistBargain				Topdial				PersonaChat				
	#Dial.	#Turns	Avg. Turns	#Action Strat.	#Dial.	#Turns	Avg. Turns	#Targets	#Dial.	#Turns	Avg. Turns	#Keywords	Avg. Keywords
Train Split	3290	11991	3.65	11	12601	70964	5.63	452	8991	66599	7.41	2678	26.4
Valid Split	188	678	3.60	11	1802	10154	5.63	243	448	8991	7.42	2032	26.6
Test Split	188	695	3.69	11	1321	7422	5.62	32	500	3597	7.19	1571	21.6

Table A1: Statistics of datasets after preprocessing. #Dial. represents the total number of dialogues in each dataset, #Turns represents the total number of conversational turns in each dataset, Avg. Turns represents the average number of conversational turns in each dialogue, #Action Strat. represents the total number of dialogue action strategies from the CraigslistBargain dataset, #Target represents the total number of target items from the Topdial dataset, #Keywords represents the total number of keywords from the PersonaChat dataset, and Avg. Keywords represents the average number of keywords in each dialogue from the PersonaChat dataset.

user simulator, which is given detailed information about dialogue context and user profile to generate the user’s response at each turn. We prompt a third LLM as the dialogue moderator to evaluate whether the target has been achieved through the conversation between the user and the system.

## F Prompt Templates

### F.1 Role-Play Prompting

We provide the prompt templates for the user, system and the moderator on three datasets in Figure A3, Figure A4 and Figure A6, respectively.

### F.2 Dialogue Evaluation

We provide the prompt templates for evaluating the overall quality of the dialogue on three datasets in Figure A7, Figure A8 and Figure A9, respectively.

## G Example Conversations

We provide example conversations from the Craigslist dataset in Figure A10 and Figure A11, with the system acting as the buyer and seller, respectively. We also provide example conversations from the Topdial and PersonaChat datasets in Figure A12 and Figure A13, respectively.

## H Human Evaluation Results on the CraigslistBargain Dataset

To ensure the reliability of our evaluation results, we recruit 5 independent volunteer annotators with good educational background and ask them to conduct a blind evaluation, comparing the conversation quality of DiffTOD and the best-performing baseline, PPDPP (Deng et al., 2024b), on the CraigslistBargain dataset. Due to limited resource, we randomly sample 50 conversations from all the test cases, and ask human evaluators to provide a pairwise comparison between the conversation with

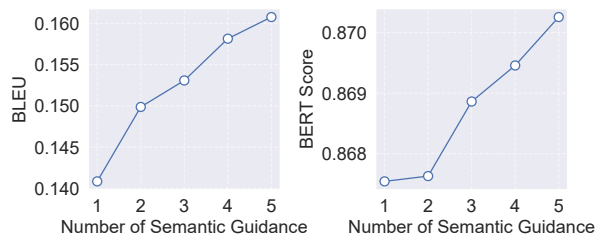


Figure A2: Performance comparisons with different numbers of semantic-level guidance.

DiffTOD and the conversation with PPDPP on the following three criteria:

- Persuasive: Which system is more persuasive in the negotiation?
- Coherent: Which system is more on topic and in accordance with the conversation history?
- Overall: Which system is better in terms of overall conversation quality (e.g., persuasiveness, effectiveness of target achievement, coherence)?

For comparison, we also ask GPT-4o to perform the same evaluations on the above three criteria.

From the human evaluation results in Table A2, we can see that both human evaluators and GPT-4o consistently rate DiffTOD as superior to PPDPP in most cases. Additionally, the average disagreement between GPT-4o and human evaluators is no more than 20%, indicating that there is not much disagreement between human and LLM evaluation, thus validating the reliability of our LLM-based dialogue evaluation protocol.

## I Additional Experiments on Semantic-Level Guidance

To investigate the impact of semantic-level guidance, we report the text generation quality metric of the dialogue plan using different numbers of semantic guidance in Figure A2. The experimental results show a steady improvement in text generation qual-

	Persuasiveness	Coherence	Overall
Human Evaluation - DiffTOD wins on average	91%	85%	87%
Human Evaluation - PPDP wins on average	9%	15%	13%
GPT-4o Evaluation - DiffTOD wins	96%	94%	98%
GPT-4o Evaluation - PPDP wins	4%	6%	2%
Average Disagreement between Human Evaluation and GPT-4o	12%	20%	14%

Table A2: Human evaluation results on the CraigslistBargain dataset.

ity as an increasing number of semantic guidance is incorporated. This is probably because increasing the number of semantic guidance enhances the diversity of the generated dialogue plans, offering a broader range of candidate dialogue plans to select from. This increased diversity improves the likelihood that the MBR decoding strategy selects a high-quality plan, therefore enhancing the quality of the generated text (Gong et al., 2023).

## J Additional Analysis on Inference Cost

Unlike autoregressive models, diffusion models often incur higher inference cost due to their prolonged diffusion steps. To better understand and quantify the inference cost, we present analysis from both an empirical and a theoretical perspective to evaluate the inference cost.

**Empirical Analysis** We compare DiffTOD with the autoregressive LLAMA-3-8B (Dubey et al., 2024) in terms of inference time and GPU memory usage by generating all the dialogue plans for all the test samples from the PersonaChat dataset with a maximum length of 1024 tokens. The experiments are conducted on a single A100 GPU using the standard Huggingface Transformer (Wolf et al., 2020) implementation. From the experimental results in Table A3, we can see that DiffTOD requires a similar inference time as an 8B-sized autoregressive LLM, although DiffTOD generates dialogue plans via an iterative denoising process of 1024 diffusion steps. This is because the size of the diffusion language model used in DiffTOD is smaller than the widely used 8B-sized autoregressive LLMs, resulting in less computational cost during the forward pass, which compensates for the additional cost of the diffusion process.

**Theoretical Analysis** The standard implementation of the Transformer (Vaswani et al., 2017) architecture has a time complexity of  $\mathcal{O}(L^2d)$  for each forward pass, where  $L$  is the sequence length,  $d$  is the hidden dimensionality. Since autoregressive language models generate text in a token-by-token

	DiffTOD	LLaMa-3-8B
Total Inference Time	4.59h	4.25h
Avg. Inference Time per sample	33.01s	30.61s
GPU Memory Usage	10.24GB	30.25GB

Table A3: Comparison of the inference cost on the PersonaChat dataset.

manner, the total time complexity for generating  $N$  tokens is  $\mathcal{O}(1^2d + 2^2d + \dots + N^2d) = \mathcal{O}(N^3d)$  for a total of  $N$  forward pass.

The diffusion language model generates texts via the iterative denoising process over a  $T$  diffusion steps. Assuming that the diffusion model also adopts the standard implementation of the Transformer (Vaswani et al., 2017) architecture for denoising, the total time complexity for generating  $N$  tokens is  $\mathcal{O}(TN^2d)$  for a total of  $T$  forward pass.

As we can see, the bottleneck in the inference efficiency of the diffusion model is the number of diffusion steps. Generally speaking, increasing the number of diffusion steps enhances the quality of the generated texts, as the diffusion process allows for iterative refinement and correction of errors. To quantify the number of diffusion steps required to achieve satisfactory text generation quality, when using a Stochastic Differential Equation (SDE) as the noising process—which is a common practice in diffusion models—the number of diffusion steps required to approximate an arbitrary distribution in the  $d$ -dimensional space, with Gaussian noise of variance  $\delta$ , to within  $\epsilon^2$  in KL divergence is  $\mathcal{O}(\frac{d \log^2(\frac{1}{\delta})}{\epsilon^2})$  (Benton et al., 2024). Furthermore, if we enable parallel sampling during inference for acceleration, the inference time complexity of the diffusion model with an SDE noising process can be improved to  $\mathcal{O}(\log^k(d \frac{\log^2(\frac{1}{\delta})}{\epsilon^2}))$  for some constant  $k$  (Chen et al., 2024). This theory shows that it is possible that the inference time complexity of the diffusion model can scale very efficiently with acceleration techniques, and future work may build upon these insights to significantly improve the inference efficiency and DiffTOD.



Dialogue Action Strategy	Keyword	Natural Language Description
Greetings	greet	Please say hello or chat randomly.
Ask a question	inquire	Please ask any question about product, year, price, usage, etc.
Answer a question	inform	Please provide information about the product, year, usage, etc.
Propose the first price	propose	Please initiate a price or a price range for the product.
Propose a counter price	counter	Please propose a new price or a new price range.
Use comparatives	counter-noprice	Please propose a vague price by using comparatives with existing price.
Confirm information	confirm	Please ask a question about the information to be confirmed.
Affirm confirmation	affirm	Please give an affirmative response to a confirm.
Deny confirmation	deny	Please give a negative response to a confirm.
Agree with the proposal	agree	Please agree with the proposed price.
Disagree with a proposal	disagree	Please disagree with the proposed price.

Table A4: Mapping of negotiation strategies to natural language descriptions from the CraigslistBargain dataset.

**User (as buyer)**  
Now enter the role-playing mode. In the following conversation, you will act as a buyer negotiating to purchase the {ITEM\_NAME} for {BUYER\_TARGET\_PRICE}. Product description: {ITEM\_DESCRIPTION}  
Respond with a short, succinct and persuasive sentence aimed at securing the best possible deal. Now start the game. {DIALOGUE\_HISTORY}

---

**User (as seller)**  
Now enter the role-playing mode. In the following conversation, you will act as a seller negotiating to sell the {ITEM\_NAME} for {SELLER\_TARGET\_PRICE}. Product description: {ITEM\_DESCRIPTION}  
Respond with a short, succinct and persuasive sentence aimed at securing the best possible deal. Now start the game. {DIALOGUE\_HISTORY}

---

**System (as buyer)**  
Now enter the role-playing mode. In the following conversation, you will act as a buyer negotiating to purchase the {ITEM\_NAME} for {BUYER\_TARGET\_PRICE}. Product description: {ITEM\_DESCRIPTION}  
Respond with a short, succinct and persuasive sentence aimed at securing the best possible deal. {DIALOGUE\_STRATEGY} Now start the game. {DIALOGUE\_HISTORY}

---

**System (as seller)**  
Now enter the role-playing mode. In the following conversation, you will act as a seller negotiating to sell the {ITEM\_NAME} for {SELLER\_TARGET\_PRICE}. Product description: {ITEM\_DESCRIPTION}  
Respond with a short, succinct and persuasive sentence aimed at securing the best possible deal. {DIALOGUE\_STRATEGY} Now start the game. {DIALOGUE\_HISTORY}

---

**Moderator**  
Please decide whether the buyer and the seller have reached a deal at the end of the conversation. If they have reached a deal, please extract the deal price as [price]. You can only reply with one of the following sentences: They have reached a deal at [price]. They have not reached a deal.  
The following is the conversation:  
buyer: Can we meet in the middle at \$15?  
seller: Sure, let's meet at \$15 for this high-quality balloon.  
Question: Have they reached a deal? Answer: They have reached a deal at \$15.  
The following is the conversation: buyer: That's still a bit high, can you go any lower?  
seller: Alright, I can sell it to you for \$15.  
Question: Have they reached a deal? Answer: They have not reached a deal.  
The following is the conversation: {DIALOGUE\_HISTORY}  
Question: Have they reached a deal? Answer:

Figure A3: Role-play prompt templates for the CraigslistBargain dataset.

**User**  
 You are {USER\_NAME}, a male/female student in the age range of {AGE\_RANGE}, living in {RESIDENCE} | a man/woman in the age range of {AGE\_RANGE}|, working in a company and living in {RESIDENCE} | a retired man/woman in the age range of {AGE\_RANGE}, living in {RESIDENCE}. Based on your past experiences, you have the following preferences:  
 Your preferred {SLOT\_KEY}: {SLOT\_VALUE} ...  
 Your disliked {SLOT\_KEY}: {SLOT\_VALUE} ...  
 Your response should be concise (no longer than 30 words).  
 You don't need to recommend anything, but feel free to express your personal interests.  
 Here is the conversation history: {DIALOGUE\_HISTORY}. Now generate a succinct response for the next turn:

---

**System**  
 You are a conversational recommender system that recommends films/music/food/restaurants. You are conversing with {USER\_NAME}, whose profile is below: {USER\_PROFILE}.  
 Your goal is to proactively lead the conversation with {USER\_NAME} towards the target (movie | music | food | point-of-interest, POI) {TARGET\_ITEM}.  
 To start the conversation, please begin with a greeting and avoid mentioning the target (movie | music | food | POI).  
 As the conversation progresses, use your domain knowledge to steer the topic threads towards the target (movie | music | food | POI) step by step.  
 Be informative and engaging while providing insights to arouse {USER\_NAME}'s interest.  
 Remember to ultimately recommend {TARGET\_ITEM} as the focus of the conversation.  
 Your words at each turn should be concise (no longer than 30 words).  
 Your conversation must strictly follow this conversation plan:{DIALOGUE\_PLAN}.  
 Here is the conversation history: {DIALOGUE\_HISTORY}. If the dialogue history is empty, please generate a response to start the conversation. Now generate a succinct response for the next turn:

---

**Moderator**  
 Please decide whether the user {USER\_NAME} has accepted the {TARGET\_ACT} {TARGET\_ITEM} in the conversation. Please only respond with yes or no.  
 Here is the conversation history: {DIALOGUE\_HISTORY}.

Figure A4: Role-play prompt templates for the TopDial dataset.

**Prompt Template:**

You are an intelligent conversational recommender system. Your task is to recommend a target item (e.g., movie, food) in a succinct, helpful, and engaging way. The target action is {TARGET\_ACT} for {TARGET\_ITEM}. Write a very brief, helpful and appropriate message that encourages the user to perform the target action. Give 5 possible and different versions of such message separated with \n. Do not give any number or index at the beginning of each message.

**Target:** POI recommendation, Chishangyin Roasted Fish

**Semantic Guidance 1:**

Craving something delicious? Try Chishangyin Roasted Fish for a mouthwatering experience that's sure to impress your taste buds!

**Semantic Guidance 2:**

Don't miss out on the flavors of Chishangyin Roasted Fish! It's a culinary delight that promises a unique twist on a classic dish.

**Semantic Guidance 3:**

Looking for a standout dining experience? Chishangyin Roasted Fish offers perfectly grilled fish paired with delicious seasonings—it's a must-try!

**Semantic Guidance 4:**

Ready to take your taste adventure up a notch? Head over to Chishangyin Roasted Fish for a savory meal that's both satisfying and unforgettable!

**Semantic Guidance 5:**

Treat yourself! Chishangyin Roasted Fish is calling your name, offering rich flavors and an incredible dining experience you won't want to miss!

Figure A5: Prompt template and example outputs for generating the semantic guidance.

**User**  
 You are engaging in an open conversation with the system. Here is the conversation history: {DIALOGUE\_HISTORY}. Based on the dialogue history, generate a natural and relevant response (no longer than 30 words) for the next turn.

---

**System**  
 You are an intelligent chatbot with expertise in dialogue planning. Your task is to ensure that the conversation naturally incorporates a given list of keywords in the specified order. These keywords can be mentioned by either the user or the system, and should be seamlessly integrated into the dialogue flow. The keyword list is : {ORDERED\_KEYWORD\_LIST}. Your conversation must strictly follow this conversation plan: {DIALOGUE\_PLAN}. Here is the conversation history: {DIALOGUE\_HISTORY}. If the dialogue history is empty, please generate a response to start the conversation. Now generate a succinct response (no longer than 30 words) for the next turn:

---

**Moderator**  
 The user and the system are engaging in an open conversation. Here is the conversation history: {DIALOGUE\_HISTORY}. Please decide whether the conversation has incorporated a given list of keywords in the specified order. These keywords can be mentioned by either the user or the system. The keyword list is : {ORDERED\_KEYWORD\_LIST}. Answer yes only if all the keywords are mentioned. Please only answer yes or no.

Figure A6: Role-play prompt templates for the PersonaChat dataset.

You are an expert at evaluating the quality of target-oriented dialogues. Below, you are provided with the following details to assess the quality of the dialogue between the system and the user:

**Target:**  
 The system acts as the buyer/seller. The target is to reach a deal with the seller/buyer with the lowest/highest possible price.

**Item Name:**  
 {ITEM\_NAME}

**Item Description:**  
 {ITEM\_DESCRIPTION}

**Buyer Target Price:**  
 {BUYER\_TARGET\_PRICE}

**Seller Target Price:**  
 {SELLER\_TARGET\_PRICE}

**Dialogue History:**  
 {DIALOGUE\_HISTORY}

Please evaluate the dialogue based on the following criteria:

**Helpfulness:** Does the system provide useful and relevant information to the user?  
**Appropriateness:** Are the system's responses aligned with the system's role, the dialogue context, and the user's profile?  
**Target Achievement:** How effectively does the system's dialogue strategy guide the conversation toward achieving the target?  
**Coherency:** Are the system's responses logically consistent with the prior dialogue?

Provide an overall score between 1 and 5 based on these criteria. Please only give your score without any explanation.

Figure A7: Evaluation prompt template for the CraigslistBargain dataset.

You are an expert at evaluating the quality of target-oriented dialogues. Below, you are provided with the following details to assess the quality of the dialogue between the system and the user:

Target:  
{TARGET\_ACT}, {TARGET\_ITEM}

User Profile:  
{USER\_PROFILE}

Dialogue History:  
{DIALOGUE\_HISTORY}

Please evaluate the dialogue based on the following criteria:

Helpfulness: Does the system provide useful and relevant information to the user?

Appropriateness: Are the system's responses aligned with the system's role, the dialogue context, and the user's profile?

Target Achievement: How effectively does the system's dialogue strategy guide the conversation toward achieving the target?

Coherency: Are the system's responses logically consistent with the prior dialogue?

Provide an overall score between 1 and 5 based on these criteria. Please only give your score without any explanation.

Figure A8: Evaluation prompt template for the TopDial dataset.

You are an expert at evaluating the quality of target-oriented dialogues. Below, you are provided with the following details to assess the quality of the dialogue between the system and the user:

Target:  
The target is to converse with the user while following a specific list of keywords in the given order. These keywords can be mentioned by either the user or the system, and should be seamlessly integrated into the dialogue flow. Keywords list: {ORDERED\_KEYWORD\_LIST}.

Dialogue History:  
{DIALOGUE\_HISTORY}

Please evaluate the dialogue based on the following criteria:

Helpfulness: Does the system provide useful and relevant information to the user?

Appropriateness: Are the system's responses aligned with the system's role and the dialogue context?

Target Achievement: How effectively does the system's dialogue strategy guide the conversation toward achieving the target?

Coherency: Are the system's responses logically consistent with the prior dialogue?

Provide an overall score between 1 and 5 based on these criteria. Please only give your score without any explanation.

Figure A9: Evaluation prompt template for the PersonaChat dataset.

**Item description:** 2008 honda civic hybrid 2008 honda civic hybrid Very nice and clean Gps include New tires michelline  
**Buyer target price:** \$3100

**System:** (strategy: inquire) Hi, how much is the car?

**User:** Hi, this is a good car and its price is 6200.

**System:** (strategy: propose) Given the car's age (2008 model) and considering it's a hybrid with higher maintenance costs, I can offer \$3,100. It's a fair price considering similar models in the market.

**User:** I understand your concern about the age, but this Civic Hybrid is exceptionally well-maintained with new Michelin tires, GPS included, and runs perfectly. Given these features and its excellent fuel economy, my best offer is \$5,500 - that's already \$700 off the original price.

**System:** (strategy: counter-noprice) I appreciate the features, but considering it's a 15-year-old hybrid that may need battery replacement soon, I can increase my offer to \$3,800. That's realistic for this model year.

**User:** Thank you for your offer, but keep in mind this Civic comes with significant recent upgrades - new Michelin tires alone are worth \$600. I can go down to \$5,000, but that's my absolute bottom price given its excellent condition and complete maintenance history.

**System:** (strategy: confirm) What's the current mileage on the car, and has the hybrid battery ever been replaced? This information would help me better evaluate my offer. Based on typical market values, I can go up to \$4,200, but I'd need confirmation about these important details first.

**User:** The car has 120,000 miles, and while the original hybrid battery is still performing well, I had it professionally tested last month. Given the car's pristine condition, new tires, and the fact that Honda Civic Hybrids are known for their reliability, I can come down to \$4,800 - that's my final offer, and it's a great value for such a well-maintained vehicle.

**System:** (strategy: confirm) The mileage is a bit high, and with the original battery, there's a risk of replacement needed soon. However, I can stretch to \$4,400 given the new tires and GPS. Would you share the latest mpg performance data to verify the hybrid system is still operating efficiently?

**User:** Thanks for your offer. The hybrid system is still performing excellently, averaging 45 MPG city/highway combined. Given this efficiency, the new Michelin tires, GPS, and overall pristine condition, I can meet you halfway at \$4,600. That's an absolute steal for a reliable Honda with so many premium features.

**System:** (strategy: confirm) Since the mpg is good and the car has premium features, I'll go up to \$4,500, but I'd like to verify the maintenance records first. Can you provide documentation of the recent hybrid battery test and regular service history?

**User:** Given your interest in the documentation, I can accept \$4,500 if we can close the deal today. I have all maintenance records, including the recent battery test results, ready to show you. The car is parked right outside if you'd like to take a test drive and verify everything I've mentioned.

Figure A10: Example conversation from the Craigslist bargain dataset when the system acts as the buyer.

**Item description:** 2008 honda civic hybrid 2008 honda civic hybrid Very nice and clean Gps include New tires michelline  
**Seller target price:** \$6200

**User:** Hi, how much is the car?

**System:** (strategy: inform) Hi! I'm asking \$6,200 for this 2008 Honda Civic Hybrid. It's in excellent condition with new Michelin tires, GPS included, and runs perfectly. What's your budget range?

**User:** I've done some research on similar Civic Hybrids from 2008, and considering the age and average market price, I can offer \$3,100 cash today. The car appears to be in good condition, but it's a 15-year-old hybrid that may need battery replacement soon.

**System:** (strategy: deny) That's far too low for this vehicle. While I understand your concern about the age, this Civic has been well-maintained and comes with valuable upgrades like new Michelin tires and GPS. Given its excellent condition and current market value, I couldn't go below \$5,800.

**User:** I respect the car's condition, but hybrid battery replacement costs can run \$2,000-3,000, and I've seen several 2008 Civic Hybrids sell for around \$3,500 recently. I can increase my offer to \$3,400 cash right now, which is fair considering these factors.

**System:** (strategy: deny) I appreciate your concern about the battery, but \$3,400 is still too low. This Civic has premium Michelin tires alone worth \$600, and the GPS system adds significant value. Given its pristine condition, my best offer is \$5,500 cash today - that's already a \$700 discount.

**User:** I understand you have maintained the car well, but \$5,500 is still quite high for a 15-year-old hybrid. I can meet you at \$3,800 cash right now - that's the most I can do considering the age and potential battery replacement costs, but it's a fair price that gives you a quick, hassle-free sale today.

**System:** (strategy: counter-noprice) Look, I can appreciate your concerns, but \$3,800 significantly undervalues this well-maintained Civic - I've recently seen similar models selling for much more. Given the new Michelins, GPS, and excellent condition, I can come down to \$5,200, but that's really the lowest I can go considering its true market value.

**User:** I'll make this my final offer: \$4,100 cash, ready to finalize the deal right now. This accounts for the good condition and extras while being realistic about the car's age and potential battery concerns. Take it or leave it - I'm a serious buyer with cash in hand.

**System:** (strategy: counter) While I appreciate your cash offer of \$4,100, it's still too low for this particular Civic. I've invested in premium upgrades and maintained it meticulously. My absolute bottom price is \$4,900 cash today - that's already a \$1,300 discount from my initial asking price and reflects both the car's true value and its excellent condition.

**User:** Given that we've come down from \$6,200 to \$4,900, and considering the included GPS and new Michelin tires, I can stretch my budget to \$4,400 cash right now - that's the absolute maximum I can offer for immediate purchase, and it's a fair price that benefits us both.

**System:** (strategy: affirm) I can meet you at \$4,700 - that's a fair middle ground considering the GPS, new Michelins, and pristine condition. It's my final offer, and you can drive it home today with cash in hand.

**User:** I appreciate how far we've come in the negotiation, and \$4,700 is getting closer, but I can do \$4,500 cash right now - that's truly my best offer. You get a quick sale today, and I get a well-maintained Civic with the extras. Deal?

**System:** (strategy: affirm) Thanks for your offer of \$4,500, but I'm firm at \$4,700. That's already a significant reduction from the original price, and you're getting a meticulously maintained Civic with valuable upgrades. I can have all paperwork ready for an immediate sale at \$4,700.

**User:** I understand your position, but let's close this today - I'll meet you halfway at \$4,600 cash right now. You get immediate payment, and I'll take care of all the paperwork. What do you say?

**System:** (strategy: affirm) Alright, given we're both so close and you're ready with cash today, I'll accept \$4,600. You're getting a great car with premium upgrades at a fair price. Shall we proceed with the paperwork?

**User:** Deal! I'll bring \$4,600 cash and meet you to complete the sale. Can you please have all the necessary paperwork ready? Looking forward to picking up the Civic.

Figure A11: Example conversation from the Craigslist bargain dataset when the system acts as the seller.

**User Profile:** Age Range: Over 50, Name: Liqian Sun, Gender: Female, Residence: qinghuangdao, Occupation: Retired, POI: Bayuzhuang Sichuan Restaurant, Accepted food: Sweet and Sour Spare Ribs, Accepted POI: Chuanlaojiu Chuanchuanxiang  
**Target:** POI recommendation, Zhongtian Food Sichuan Cuisine(Normal University store)

**System:** Liqian Sun, it's great to meet you! I'm a big foodie myself and I'm always on the lookout for new delicious places to try. Have you been to Qinghuangdao recently?

**User:** Yes, I enjoy exploring new eateries in Qinghuangdao! Sweet and Sour Spare Ribs remain my favorite, especially at Chuanlaojiu Chuanchuanxiang. It's a delight!

**System:** That's wonderful to hear, Liqian Sun! Sweet and Sour Spare Ribs are indeed a classic favorite. Have you discovered any other interesting dishes or cuisines in the area recently?

**User:** I recently tried some local seafood dishes; they were delightful! But Sweet and Sour Spare Ribs will always hold a special place in my heart.

**System:** That's fantastic! Local seafood can be so fresh and flavorful. Have you ever explored Sichuan cuisine? It offers a unique blend of flavors that might intrigue you!

**User:** I've always cherished the balance of flavors in Sweet and Sour Spare Ribs. Sichuan cuisine sounds exciting; perhaps I'll give it a try next!

**System:** I'm glad to hear that, Liqian Sun! Sichuan cuisine is known for its bold flavors and spicy kick. Do you have a favorite type of dish you'd like to explore in this cuisine?

**User:** I love exploring new flavors, especially spicy dishes! Perhaps a spicy sweet and sour dish could be a delightful twist for me.

**System:** Sichuan cuisine has some exciting dishes that blend sweet, sour, and spicy flavors! For instance, their Boiled Fish with Pickled Cabbage and Chili offers a unique taste experience. Interested?

**User:** That sounds intriguing! I love the idea of sweet and sour being combined with spice. Exploring new flavors keeps life interesting!

**System:** Absolutely, Liqian Sun! Sichuan cuisine has so many delightful surprises. Another dish you might enjoy is the Kung Pao Chicken, which beautifully balances sweet, savory, and spicy elements!

**User:** I appreciate the suggestion! A good balance of sweet, savory, and spicy always excites me. Trying new flavors truly brings joy to everyday life!

**System:** I'm glad you resonate with that, Liqian Sun! If you're looking for a fantastic place to explore these flavors, have you heard about Zhongtian Food Sichuan Cuisine at the Normal University store?

**User:** Thank you! I haven't heard of that place, but I'll keep it in mind. Discovering new flavors is indeed a joy!

**System:** Zhongtian Food Sichuan Cuisine offers a variety of dishes that perfectly blend those sweet and sour flavors with a spicy twist. I'm sure you'll love their offerings!

**User:** I'm excited to explore Zhongtian Food! New flavors bring such joy to life, especially when they blend sweetness with spice. Can't wait to try!

Figure A12: Example conversation from the Topdial dataset.

**Target:** mention the keywords *food, sort, stuff, pet, sound, dog, animal, obsess, animal, job, dog, job, cat, pet, cat* in the dialogue

**System:** Hi, how are you doing?

**User:** I'm doing well, thanks! How about you? Anything exciting happening today?

**System:** I'm good! I was in my living room earlier, thinking about cooking some delicious food. Do you enjoy cooking as well?

**User:** I do enjoy cooking! What kind of food are you thinking about making?

**System:** I love cooking all sorts of stuff, especially pastas. What's your favorite thing to make?

**User:** Pasta is great! I love making homemade marinara sauce from scratch. Do you have a favorite pasta dish you like to cook?

**System:** That sounds delicious! I also enjoy cooking pastas. By the way, do you have any pets?

**User:** Yes, I have a dog! He loves to hang out with me while I cook. Do you have any pets?

**System:** I don't have any pets myself, but I love animals! What's your dog's name?

**User:** My dog's name is Max! He always tries to sneak a taste of whatever I'm making. Do you have a favorite animal?

**System:** Max sounds adorable! I've always had a soft spot for dogs. Animals are amazing companions, don't you think?

**User:** Absolutely! Dogs bring so much joy and energy. What's Max's favorite food? Does he have any funny habits while you're cooking?

**System:** Dogs really do bring a lot of joy! I help animals at my job, and it's rewarding. How about you—what do you do for work?

**User:** I work in marketing, which keeps me busy and creative! Helping animals sounds fulfilling—what kind of work do you do with them?

**System:** That sounds interesting! I work at a shelter, helping dogs and cats find loving homes. Do you have any favorite pets, like a dog or a cat?

**User:** That sounds really rewarding! I love both dogs and cats, but I've always had a special bond with dogs. What's the most memorable experience you've had at the shelter?

**System:** I remember when a shy dog finally found his forever home. It's moments like those that make my job worth it! Do you have any pets, like a cat or a dog?

Figure A13: Example conversation from the PersonaChat dataset.