An Efficient Dialogue Policy Agent with Model-Based Causal Reinforcement Learning

Kai Xu[1], Zhenyu Wang[1,*], Yangyang Zhao[2], Bopeng Fang [3]
[1] School of Software Engineering, South China University of Technology, Guangdong, China
[2] Department of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha, China
[3] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China
***Correspondence:** wangzy@scut.edu.cn

Abstract

Dialogue policy trains an agent to select dialogue actions frequently implemented via deep reinforcement learning (DRL). The model-based reinforcement methods built a world model to generate simulated data to alleviate the sample inefficiency. However, traditional world model methods merely consider one-step dialogues, leading to an inaccurate environmental simulation. Furthermore, different users may have different intention preferences, while most existing studies lack consideration of the intention-preferences causal relationship. This paper proposes a novel framework for dialogue policy learning named MCA, implemented through model-based reinforcement learning with automatically constructed causal chains. The MCA model utilizes an autoregressive Transformer to model dialogue trajectories, enabling a more accurate simulation of the environment. Additionally, it constructs a causal chains module that outputs latent preference distributions for intention-action pairs, thereby elucidating the relationship between user intentions and agent actions. The experimental results show that MCA can achieve state-of-the-art performances on three dialogue datasets over the compared dialogue agents, highlighting its effectiveness and robustness.

1 Introduction

Dialogue policy plays a crucial role in task-oriented dialog systems with a pipeline approach, as it determines the next agent action and drives the dialog generation (Kwan et al., 2023; Zhao et al., 2024). In recent years, the construction of dialogue policy agent has been regarded as a sequential decision-making problem and optimized via deep reinforcement learning (DRL). The optimization process can be summarized: Interaction–Samples–Deep Learning–Action, i.e., the *interaction* between agent and user to generate *samples* and employ *deep learning* to train the samples for predicting the next agent *action*. The traditional dialogue policy methods include model-free and model-based methods, the deep Q-network (DQN)-based methods (Tian et al., 2022; Zhao et al., 2021a; Zhang et al., 2022), the actor-critic(AC) based methods (Malviya et al., 2022; Peng et al., 2018a; Chen et al., 2020) are the model-free methods. However, these methods lack modeling of the environment and don't have the ability of perception, making the dialogue agent inadequately anthropomorphic. Recently, as task-oriented dialogue systems have become more popular and powerful, building a human-like dialogue agent with efficient perception and causal reasoning is a promising research direction.

Compared to the model-free methods, the model-based methods build a world model to facilitate dialogue agents' decisions by simulating human perception (Matsuo et al., 2022). In the dialog policy, the world models help the dialogue agent more efficiently by generating simulated data with learning in imagination, such as DDQ (Peng et al., 2018b), Budget DDQ (Zhang et al., 2020), and DPPO (Huang and Cao, 2023). However, these approaches rely on large amounts of real user-agent interaction data and only consider one-step dialogue transition, ignoring multi-step perception like humans. Research has shown that modeling multi-step state transition has more accurate perceptual capabilities (Benechehab et al., 2023; Micheli et al., 2023). Therefore, designing a new world model that can handle complex sequenced interaction data with fewer samples is a more feasible endeavor similar to human perception.

Furthermore, existing research on dialogue policy also lacks interpretability, establishing state-action mapping relationships through deep neural networks, that hard to uncover the internal relationship between user intentions and actions. Indeed, human beings not only have the ability to perceive the outside environment but infer some

Proceedings of the 31st International Conference on Computational Linguistics, pages 7331–7343
January 19–24, 2025.

latent result based on past interactions, i.e., we can generalize some fundamental causal relationships based on past information to better facilitate future decision-making (Griffiths et al., 2010). The causality strategy from psychological research (Sloman, 2005) can be aware of the causal mechanisms of underlying actions and improve the interpretability (Gao et al., 2024). However, existing dialogue policy methods are not equipped with causal chains. Although some causal reinforcement learning algorithms consider the static prior knowledge (Goyal and Bengio, 2022; Pouncy and Gershman, 2022), which did not improve their reasoning ability during interactions without adaptation. To dynamically acquire user causal chains to explain agents' decisions, we collect success trajectories in interactive dialogues between the agent and the user, constructing causal chains that present potential changes between the user's intentions and the agent's actions. To the best of our knowledge, the proposed model is the first method applying causal chains to learn dialogue policies without requiring any expert experience.

This paper proposes a dialogue policy learning framework with model-based reinforcement learning and causal chains. It implements the perception and reasoning for anthropomorphic dialogue agents. Concretely, we first employ the autoregressive Transformer (Vaswani et al., 2017; Micheli et al., 2023) to build a world model that simulates dialogue trajectories. The constructed world model dynamically learns a sequence modeling problem rather than one-step transitions, making us accurately simulate the dialogue environment. Secondly, we build causal chains via user intentions to form the latent preference distribution and then distill the latent distribution into the policy learning model. Our main contributions are:

- We proposed a sample-efficient dialogue policy learning method that simulates human perception and reasoning, integrating world models and causal chains to facilitate policy learning and enhance interpretability.
- We build an autoregressive Transformer-based world model, which learns dialogue trajectories rather than one-step transition, constructs a causal graph between user intentions and actions during interaction, generates a latent intention-action distribution, and distills the distribution into policy learning.
- We compare plentiful algorithms on three dialogue datasets, and experimental results demonstrate our method has effectiveness, high performance, and robustness.

2 Background and related work

2.1 Dialogue policy

Dialogue policy is one of the core components of pipeline approach dialogue systems, often modeled through deep reinforcement learning (DRL). Dialogue policy learns policy parameters (or Q-value) according to the dialogue state, then selects an action to push natural language generation (Kwan et al., 2023). The DRL-based dialogue policy collects data during the interaction and predicts dialogue action by learning the collected data. It consists of five tuples, $< S, A, P, R, \gamma >$, where S is the dialogue state space, A is the dialogue action space, P is the state transition function of the environment, and R is the reward function. At the moment t, the dialogue agent receives the state s_t, obtains the reward r_t by taking action a_t, and then the environment transfers to the next state s_{t+1}. The optimization objective of the dialogue policy is to maximize the overall cumulative reward.

$$\pi^* = argmax_\pi \mathbb{E}(\sum_{t=0}^{T} \gamma^t r_t) \quad (1)$$

where π^* is the optimal policy, T is the maximum dialogue turns, and γ is the discount factor used to force the dialogue agent to focus on the shot term reward.

2.2 World Model

The world model is used to perceive the outside environment Currently, the dialogue policies favor learning in imagination to train dialogue agents, aiming to mitigate sample inefficiencies. Dialogue agents have real experiences and simulated experiences. The real experiences come from the real environment, and simulated experiences come from the world model. The world model simulates the dynamics of the environment, including state transfer $P_w(s'|s,a)$ and dialogue rewards $R_w(r|s,a)$.

$$s' \sim P_w(s'|s,a), \quad r = R_w(r|s,a) \quad (2)$$

The DDQ (Peng et al., 2018b) model is the first study to train a dialogue agent via an imaginary world model. It provides a theoretical foundation for subsequent studies on dialogue policy learning. For example, (Wu et al., 2019) introduces a

switcher to improve DDQ by automatically balancing the usage of simulated experiences and real experiences. Budgeted policy learning (Zhang et al., 2020) leverages active learning and human teaching to guide DDQ in generating more highly effective dialogue experiences. DR-D3Q (Zhao et al., 2020) designs a dynamic reward function based on the user's valid subgoals via the DDQ and replaces the deep Q-network with the Dueling network. Similar to the work of (Zhao et al., 2020), (Huang and Cao, 2023) integrate the proximal policy optimization (PPO) algorithm with a DDQ-based world model, proposed DPPO. However, Both (Zhao et al., 2020) and (Huang and Cao, 2023) do not modify the world model module of DDQ, merely replacing the deep Q-network module in DDQ with different deep reinforcement learning algorithms.

Moreover, the above methods of world model-based dialogue policies only consider one-step transitions. The world model module inputs the current dialogue state and the last agent action and outputs the predicted next user action, reward, and termination signal. The interaction between the agent and user usually involves multiple turns. If the world model can model the whole sequential data, it can make sufficient use of the interaction information and enhance its perceptual ability. Meanwhile, the Transformer has a particular advantage when operating sequential discrete tokens. The Transformer-based modal world model has achieved advanced results in the game tasks (Micheli et al., 2023).

2.3 Causality in Reinforcement Learning

Causality draws some inferences based on existing information, which is combined with reinforcement learning to construct causal reinforcement learning (Wang et al., 2021). Causal reinforcement learning can improve generalization and interpretability. In recent years, there has been a prevalence that integrates causal inference with reinforcement learning. (Ramachandran et al., 2022) proposed a causal aware safe policy improvement method, which learns causal reward with the human demonstrator. However, the collection of expert demonstration data is time-consuming and laborious. (Wen et al., 2024) proposed a diversity-aware causal mode, where they modeled user's feedback through causal inference, combined with offline reinforcement learning, for promoting diversity in interactive recommendations. They lack dynamically expanding causal graphs and are harder to apply to variable scenarios. In addition, some studies integrate causal inference and model-based reinforcement learning (Mutti et al., 2023; Wang et al., 2022; Yu et al., 2023). Based on these studies, the combination of causal reasoning and world modeling can help improve performance.

Our proposed MCA has an independent causal inference module from the world model, where the inference information is the user's intention rather than the complete state, which reduces the size of causal graphs. Furthermore, we continuously optimize the causal information throughout the reinforcement learning interaction without expert experiences. To the best of our knowledge, the proposed MCA model is the first dialogue policy learning framework that mimics human perceptual and causality behaviors. It endows the dialogue agent with adaptive perceptual and memory capabilities, aiming to make the dialogue agent more human-like.

3 Method

3.1 Architecture Overview

We adopt the autoregressive Transformer to build the world model, and the knowledge of causal chains is infused into the direct reinforcement learning for our proposed dialogue policy method. The world model is endowed with the ability to perceive the outside world, while the causal chains are generated via identical user intentions to obtain the latent action distribution. The output layer of the direct deep reinforcement learning module is refashioned for distilling the latent action distribution. The primary motivation of this work is to enhance the sample efficiency by drawing inspiration from human perception (world model) and ratiocination (causal chains). A pictorial description of our framework is presented in Figure 1. It consists of three main modules:

(a) Transformer-based world model: It receives the real squeezed interaction memory and outputs the predicted following user action, current reward, and termination flag. It is based on supervised learning to perceive the outside environment and generate simulated data D^s. The Transformer-based world model has two advantages. (i) We model the entire dialogue sequence without a single step of transfer, which allows us to construct a more accurate world model. (ii) We model the world using only successful interaction trajectories without expert demonstrations.

(b) Direct reinforcement learning: It is imple-

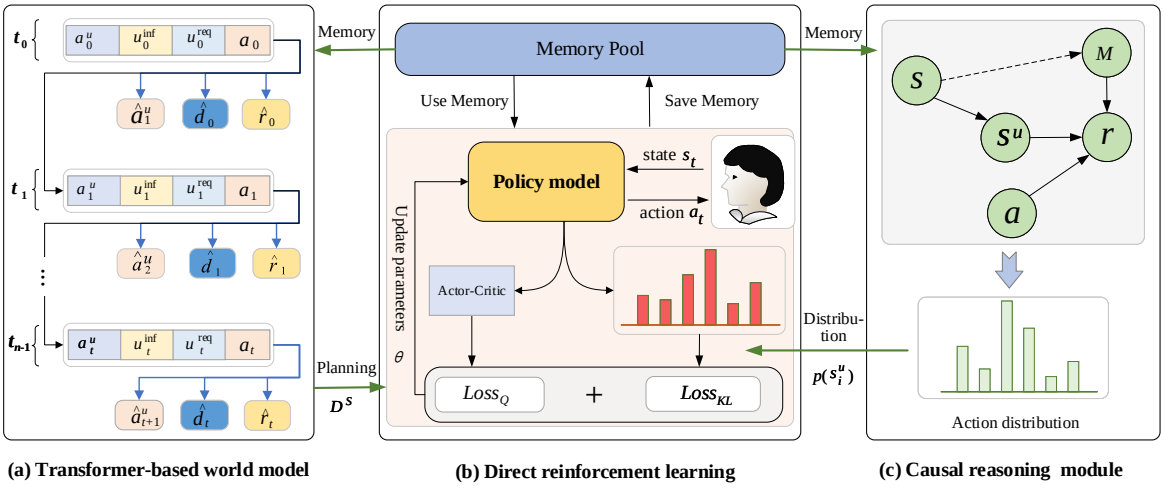


Figure 1: The proposed MCA framework. It contains three modules, (a) Transformer-based world model, (b) Direct reinforcement learning for dialogue interaction, (c) Causal reasoning module.

mented via actor-critic based reinforcement learning. This module is trained with previous interaction memory and the simulated data, distilling the valuable knowledge from the causal reasoning module. The Actor network Q-value distribution will combine with the causal action distribution.

(c) Causal reasoning module: It analyzes the user-agent interaction information to obtain successful agent actions with different user intentions, then generates the causal chains to provide knowledge for direct reinforcement learning. It can gradually learn the user's intent preferences, which helps to provide a more personalized response.

The three modules interact with each other to iteratively facilitate efficient dialogue policy learning.

3.2 Transformer-based World model

In dialogue policy learning, the current dialogue state s_t consists of the user intention s_t^u, the last agent response s_{t-1}^a, dialogue history s_t^h, and database query results s_t^d. This combination is more common in recent studies (Peng et al., 2018b; Zhao et al., 2021b; Rohmatillah and Chien, 2023). $s_t = [s_t^u, s_{t-1}^a, s_t^h, s_t^d]$, where [ ] represents the connection operation. The user intention s_t^u includes user action a_t^u, user information slot u_t^{inf}, and user request slot u_t^{req}, $s_t^u = [a_t^u, u_t^{inf}, u_t^{req}]$. t is the number of current dialogue turns.

Existing world model methods (Peng et al., 2018b; Zhao et al., 2020; Zhang et al., 2020) simulate one-step dialogue transfer, lacking consideration of the holistic dialogue interaction. In the real world, the dialog agent usually has to engage in multi-turn conversations with a user to complete a specific task. The users' utterances are drawn into the user's intentions, and dialogue policies select a dialog action to respond to user utterances. This process involves a sequence of user intentions with agent actions. Thus, we use an autoregressive Transformer to model the world model for simulating the sequential interaction data (as shown in Figure 1.a). The discrete user intentions and agent actions are regarded as the tokens to input the autoregressive Transformer. The input sequence is:

$$H_i^t = (s_0^u, a_0, s_1^u, a_1, ..., s_t^u, a_t) \quad (3)$$

which come from the dialogue history s_t^h. The autoregressive Transformer-based world model according to the H_i^t predicts the next user action $\hat{a}_{t+1}^u$, the current termination flag $\hat{d}_t$, and the current reward $\hat{r}_t$. The $\hat{a}_t^u$, $\hat{d}_t$, and $\hat{r}_t$ can be obtained as follows:

$$Transition: \hat{a}_{t+1}^u \leftarrow T(\hat{a}_{t+1}^u \mid s_{\leq t}^u, a_{\leq t} \mid H_i^t) \quad (4)$$

$$Termination: \hat{d}_t \leftarrow T(\hat{a}_{t+1}^u \mid s_{\leq t}^u, a_{\leq t} \mid H_i^t) \quad (5)$$

$$Reward: \hat{r}_t \leftarrow T(\hat{a}_{t+1}^u \mid s_{\leq t}^u, a_{\leq t} \mid H_i^t) \quad (6)$$

where T is the autoregressive Transformer model, $s_{\leq t}^u = \{s_0^u, ..., s_t^u\}$, $a_{\leq t} = \{a_0, ..., a_t\}$. Each s_i^u corresponds to a user action a_i^u, and each pair of (s_i^u, a_i) corresponds to a reward r_i and termination d_i. We train T with planning time steps via successful interaction experiences. The cross-entropy

loss is employed for $\hat{a}_t^u$ and $\hat{d}_t$, and a mean-squared error loss for reward $\hat{r}_t$.

The proposed autoregressive Transformer-based world model is empowered to generate simulated experience D^s with planning G. In more detail, Algorithm 1 (See appendix A) summarizes the proposed world model.

3.3 Causal reasonging Module

The causal reasoning module infers user intention preferences from the interacted data, which considers user preferences and task completion. Figure 1(c) illustrates causal inference information that contains a causal graph G. Traditional online reinforcement learning can be regarded as the causality mapping the relationship from state to action (Schulte and Poupart, 2024). In Figure 1(c), we use s to denote the dialogue state, a denotes the action of the dialogue, and r denotes the immediate reward signal. However, this mapping lacks consideration of the causal relationship between user intent and action. We extend it by adding s^u and M nodes, where s^u is the user's intention and M denotes the causal effect of user intention s^u under dialogue action a.

We consider three factors to estimate the causal effect M. Including the number of intention's actions, the action distribution of current intention, and the action distribution of similar intentions. The number J of user-intention actions reflects the optional action. We use factor C_i to indicate,

$$C_j = exp(-J) \tag{7}$$

The smaller J is, the larger the factor C_j is, and the stronger the causal effect of the intention.

The action distribution of current intention reflects the user's preference probability for different actions, which is denoted by

$$P_j^u = \frac{m_j^{s^u}}{\sum_{j=1}^{|a|} m_j^{s^u}} \tag{8}$$

where $m_j^{s^u}$ is the number of action j correspond to the intention s^u, and $|a|$ is the number of selectable action. The larger P_j^u is, the stronger the user's preference for the current action.

Considering that the causal graph G reflects a limited number of causal relationships between user intentions and actions, for a new user intention, we employ a similar intention to obtain the corresponding action distribution. Hence, we construct the similarity factor for the causal effect.

$$w_j = exp(-(\epsilon + \sum_{(s^{u_i}|s,a)\in\tau} sim(s^{u_i}, s^{u_j}))^{-1}) \tag{9}$$

where τ is the dialog trajectories, ϵ is a hyperparameter that prevents the numerical value from being 0. w_j denotes the weight value of similar intentions. We select the most similar intention in the setting. For a dialog interaction

$$\tau_i = (s_0, a_0, r_0, ..., s_t, a_t, r_t, ..., s_{L-1}, a_{L-1}, r_{L-1}, f) \tag{10}$$

where $f = 1$ indicates a user's goals have been completed. Note that (s_t, a_t) can be attended in either successful or failed dialog interactions. Because an agent completes a user's single-turn requests does not mean it completes the entire user task. Furthermore, the user intention s_i^u is a portion of s_i. A successful agent action means that the user's intention is fulfilled. Therefore, we employ the user intention and the agent action to build the causal chains. The final causal effect is

$$m_j = Softmax(\gamma_1 C_j + \gamma_2 w_j p_j^u) \tag{11}$$

The $Softmax$ function is employed to map the user's preference information within $[0, 1]$. γ_1 and γ_2 are hyperparameters. In this paper, we set them to 1.

From the above analysis, we finally output a prior distribution $p(s^u|s, a_j) \sim m_j$ as the biased knowledge bias, then the distribution $\boldsymbol{p}(s^u|s)$ for current user intention is used to improve policy learning. Moreover, the $\boldsymbol{p}(s^u|s)$ is dynamically shifting during the dialogue interactions. The denser the number of successful dialogs ($f = 1$), the stronger the agent latent action distribution.

3.4 Dialogue Policy Learning

The simulation experience D^s obtained from the world model is directly used to improve dialogue policy learning. The distribution $p(s_i^u|s_i, a_i)$ obtained from the causal reasoning module is used as a priori knowledge to facilitate the efficiency of the policy model. We employ the PPO as a direct reinforcement learning to train the dialogue agent, the Actor network as a policy model, and the Critic network as an evaluating model. The Actor network outputs action distribution $q(a_i|s_i)$ with dialogue state s_t. The distribution $p(s_i^u|s_i^u, a_i)$ is integrated into the Actor network. We draw inspiration from the knowledge distillation (Bachem

and Geist, 2024; He et al., 2022) to learn the latent distribution. Then the Kullback-Leibler (KL) divergence is used between the $q(s_i, a_i)$ and $p(s_i^u|s_i, a_i)$. The loss function is defined as follows:

$$L(\theta) = L_{PPO}(\theta) + \alpha H(log_{\pi_\theta} a) + \beta KL(\boldsymbol{q}|\boldsymbol{p}) \tag{12}$$

where $L_{PPO}(\theta)$ is the PPO loss, $H(log_{\pi_\theta} a)$ is the entropy regularization, $KL(\boldsymbol{q}|\boldsymbol{p})$ is the distill the ratiocination causal action distribution, α and β is the hyper-parameters, $\boldsymbol{q}$ and $\boldsymbol{p}$ are the Actor network action distribution and the ratiocination causal action distribution, respectively. We use the reverse KL because it has stronger policy improvement guarantees (Chan et al., 2022; Bachem and Geist, 2024). Algorithm 2 (see appendix A) summarizes the whole procedure of the proposed MCA framework. Note that our method can be implemented using different reinforcement learning algorithms. For example, value-based reinforcement learning methods, such as DQN, Dueling DQN. When using a value-based reinforcement learning algorithm, its current network structure needs to be modified to enable it to output the latent agent action distribution, and the simulated data are directly incorporated into the interaction experiences.

4 Experiment

4.1 Datasets and Baseline models

We implement experiments on three task-oriented dialogue datasets: movie-ticket booking, taxi booking, and automatic diagnosis. The movie and taxi domain datasets are provided by the Microsoft Dialogue Challenge (Budzianowski et al., 2018), and the diagnosis domain dataset comes from (Liu et al., 2018). These datasets have been commonly used in dialogue policy learning (Xu et al., 2024; Huang and Cao, 2023; Qiu et al., 2023; Zhao et al., 2024). The movie ticket booking task includes 128 user goals and 2890 dialogs. The taxi calling task includes 3094 dialogs with 158 user goals for experiments. The automatic diagnosis dataset is collected from the pediatric department in a Chinese online healthcare community, which includes 67 symptoms and more than 700 user goals. We compare our proposed MCA framework with some baselines:

- **DQN** has successfully learned policies from high-dimensional state inputs. Most of the dialogue policy learning methods are implemented based on DQN. Such as DDQ (Peng et al., 2018b) and DPPO (Huang and Cao, 2023), etc.
- **Dueling** modifies the structure of DQN to alleviate the over-optimization of q-values. It decomposes a single stream of fully connected layers into the Q-value and the advantage value function.
- **DRQN** modifies the portion of deep learning in DQN, replacing MLP with the LSTM network, which achieved advanced results in dialogue policy learning (Wang et al., 2016).
- **MAXMIN** (Lan et al., 2020) employs multiple DQNs and selects the minimum Q-value among the maximum Q-value of DQNs to alleviate the overestimation bias of Q-learning.
- **DPAV** utilizes a weight between the maximum q-value and the minimum q-value to estimate the ground truth q-value. The weight can be obtained via the heuristics algorithm (Tian et al., 2022).
- **DDQ** is the first study to apply a world for modeling dialogue policy learning. Its world model only considers one-step transfer (Peng et al., 2018b).
- **DR-D3Q** integrates an adaptive reward to the loss function and replaces the reinforcement learning portion of DDQ with the Dueling network (Zhao et al., 2020).
- **LKTD** quantifies the uncertainty in the environmental dynamics through Kalman Temporal-Difference and can supervise the uncertainty during policy updating, enhancing the robustness of policy learning (Shih and Faming Liang, 2024).

We use dialogue success rate (***suc.***) and average reward (***rew.***) to evaluate the experimental performances. The ***suc.*** indicates the completion of tasks within the maximum number of turns. The ***rew.*** is the average of the cumulative rewards by the reinforcement learning agent for completing dialogues. These metrics are frequently employed to evaluate the dialog policy learning (Lu et al., 2023; Zhao et al., 2024).

To comprehensively evaluate the overall performance of each model, we added new evaluation metrics, AUS and AUR, which reflect the Area Under the Success rate curve and the Area Under

Reward curve, respectively. These two metrics are used because they can cover both efficiency and performance. The higher the learning efficiency of the model, the faster the curve rises and the larger the area is, and the higher the performance of a model, the higher the curve peaks and the larger the area is. The AUS is calculated using the trapezoidal rule and is normalized. $AUS = \frac{\sum_{k=1}^{K-1}(suc._{k+1} - suc._i)}{2K(MAX_{suc.} - MIN_{suc.})}$, where $MAX_{suc.}$ is the maximum, and $MIN_{suc.}$ is the minimum of dialogue success rate. k is the number of iterations. The AUR is calculated in the same way as AUS. Implementation details are described in Appendix B.

4.2 Evaluation Results

In this section, we display the performance of the proposed MCA with baseline algorithms. Fig.2 presents the experimental results. In the movie domain, the proposed MCA achieved the highest dialogue success rates and the largest dialogue average rewards. Moreover, the MCA has faster convergence in three dialog domains, owing to the incorporation of the causal reasoning module and the autoregressive Transformer-based world model. In Fig.2(a), the DDQ model has a sub-optimal learning efficiency only than MCA, attributed to the available world model. The MAXMIN has a relatively good dialog success rate at the 400th epoch, but the training speed is relatively inferior. The main reason is that MAXMIN introduces multiple agents that do not balance the Q-values well in the early stages of training. The LKTD has a relatively fast training speed during the early epoch stages. But after the 150th epoch, its results begin to fluctuate. DPAV and Dueling agents have similar dialogue success in the 400th epoch, but DPAV has a slight performance advantage over Dueling.

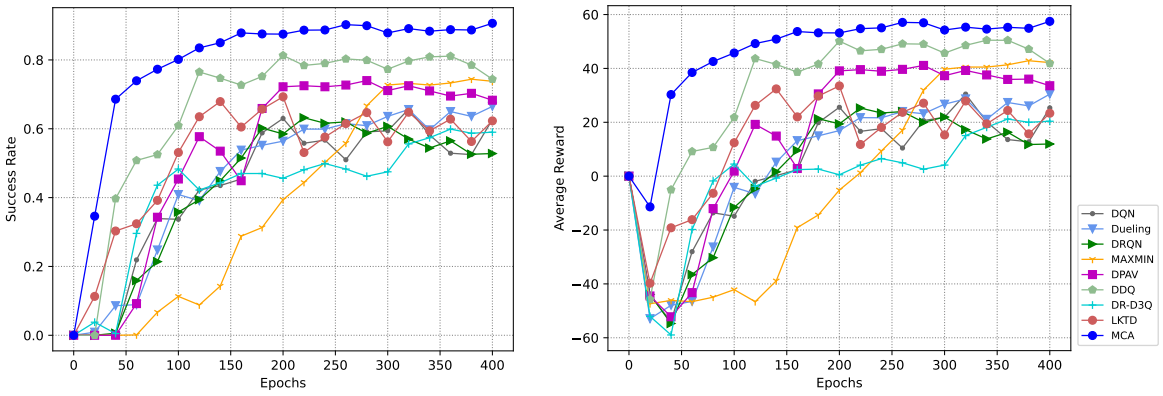
In the dialogue average reward for the movie domain, MCA has the highest average reward, and DDQ has the second-best results. The DDQ model shows a decrease in average dialogue rewards in the later stages of its training. Owing to the poorer simulated experience hinders the learning performance of the dialogue agent. DRQN has the worst average dialog reward because the LSTM-based state encoding does not outperform the state encoding of multilayer perceptron machines. The dialogue average reward falls before it rises, reflecting that the dialog agent cannot make good decisions in the early training.

In the taxi domain, MCA maintains an advanced learning rate with the highest dialog success rate and achieves better results before 100 epochs. DDQ achieves better results in the early stages of training, but the experimental results fluctuated widely in the later stages. Its performance is comparable to its performance in the movie domain. MAXMIN has the slowest convergence rate compared to the other agents, but after the 150th epoch, its dialogue success rate start to increase. The poor results in the early stages of training because the MAXMIN does not explore a better Q value. In Fig.2(d), the average dialog reward obtained by the proposed MCA is at the leading level. The advantages of the other algorithms are similar to the results of their dialog success rates. In the diagnosis domain (see Fig.2(e) and Fig.2(f)), MCA has a clear advantage in terms of dialogue success rate and dialogue average reward. The final results for DQN and DDQ are comparable, but the performance advantage of DDQ over DQN is in general. LKTD has the worst results in the diagnosis domain, showing a weaker adaptation.

The above experimental results show that the proposed MCA framework achieves advanced results in terms of dialog success rate and average dialog rewards, highlighting the performance and effectiveness of our proposed methodology. Table 6 displays the detailed performance of three task-oriented dialogue domains (see Appendix C for details). Both AUS and AUR demonstrate that our proposed MCA achieves the best results in three domains. The overall performance is optimal. The experimental results demonstrate that our proposed method can balance convergence and performance.

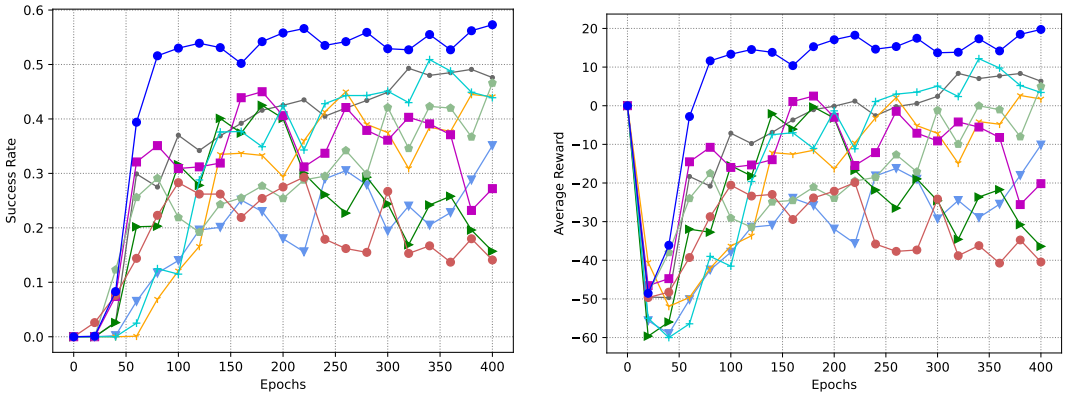
4.3 Human Evaluation

We recruited 20 volunteers to perform the human evaluation, each rated on an integer scale of 0-5. For the sake of fairness, each volunteer does not know the agent that makes, and they only evaluate whether the agent's policy actions can complete a specific task. Table 1 shows the results of human evaluation. It can be seen that the proposed MCA obtains high human scores in the three different domains. The DDQ model achieves the second-best human scores in the movie domain. The DQN receives better human scores in the taxi domain. The DR-D3Q achieves human scores that are just lower than the MCA on the diagnosis domain, and the results of the DR-D3Q are closer to the results



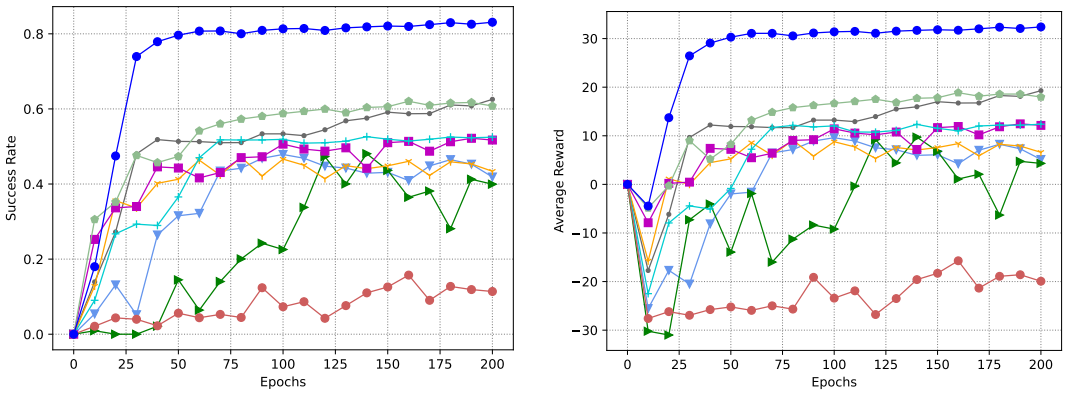
(a) Movie domain success rate

(b) Movie domain average reward



(c) Taxi domain success rate

(d) Taxi domain average reward



(e) Diagnosis domain success rate

(f) Diagnosis domain average reward

Figure 2: The results of proposed MCA with different baselines.

of the DPAV. The human evaluation demonstrates the superiority of the MCA method. It offers a performance advantage over the second-best method.

4.4 Ablation study

The reinforcement learning (RL) component of our proposed method is implemented through the PPO. The Transformer-based world model and causal ratiocination are constructed as different modules to facilitate policy learning. To assess the effectiveness of various modules, we conducted ablation experiments on each of the three domain datasets. Tables 2 to 4 display the experimental results. Where RL is policy learning using only the PPO algorithm, 'W' is an abbreviation to denote the Transformer-based world model, and 'C' denotes the causal reasoning module. Table 2 shows the experimental results in the movie domain. The reinforcement

	DQN	Dueling	DRQN	MAXMIN	DPAV	DDQ	DR-D3Q	LKTD	MCA
movie	2.65	2.30	1.90	1.80	3.25	3.45	3.35	3.40	**3.95**
taxi	2.60	1.85	1.95	2.00	2.15	2.10	2.25	1.60	**2.95**
diagnosis	2.30	2.25	1.75	1.70	3.25	3.05	3.30	1.65	**3.85**

Table 1: Experimental results via human evaluation.

learning baseline module has better $suc.$, $rew.$ than the RL+W module, but the metrics AUS, AUR and $Overall$ are inferior to RL+W. This indicates that RL+W has faster convergence. RL+C has a higher $suc.$, AUS, AUR and $Overall$ than RL. This indicates better performance and faster convergence. The proposed method is RL+W+C, which achieves more advanced results in the movie domain and demonstrates the superior performance of the model. The results with integrated world model (+W) and causal inference modules (+C) are consistently optimal in taxi (Table 3) and diagnosis (Table 4) domains.

	movie				
	suc.	*rew.*	*AUS*	*AUR*	*Overall*
RL	0.8700	58.41	0.6744	0.7224	0.6984
RL+W	0.8490	50.68	0.7522	0.7563	0.75425
RL+C	0.8805	54.89	0.8300	0.8390	0.8345
RL+W+C	**0.9065**	**58.47**	**0.8619**	**0.8663**	**0.8641**

Table 2: Ablation study in movie domain.

	taxi				
	suc.	*rew.*	*AUS*	*AUR*	*Overall*
RL	0.5491	15.80	0.5105	0.4724	0.4915
RL+W	0.5313	8.21	**0.6662**	0.6761	**0.6713**
RL+C	0.5614	18.78	0.5913	0.6332	0.6121
RL+W+C	**0.5733**	**19.74**	0.6526	**0.6821**	0.6674

Table 3: Ablation study in taxi domain.

	diagnosis				
	suc.	*rew.*	*AUS*	*AUR*	*Overall*
RL	0.8075	31.61	0.8016	0.9079	0.8548
RL+W	0.8105	31.42	0.8412	0.9210	0.8811
RL+C	0.8182	32.37	0.8151	0.9139	0.8645
RL+W+C	**0.8309**	**32.40**	**0.8924**	**0.9402**	**0.9163**

Table 4: Ablation study in diagnosis domain.

4.5 Forward and inverse KL divergence

Our MCA model uses the inverse KL divergence to fuse the causal reasoning modules. We implemented an experiment to verify the effect of temperature coefficients under the forward and inverse KL divergence. Table 5 illustrates the experimental $Overall$ results in different domains. We discovered that smaller inverse KL divergence achieves better results when tested on three dialogue tasks. This implies that the MCA model focuses more on the causal reasoning module (Smaller temperature values correspond to the amplification of the causal reasoning module). In general, the inverse KL divergence is more effective than the forward KL divergence, especially in the taxi domain, where the difference between inverse and forward KL divergence is more distinct. This is mainly due to the distribution of the original datasets, while the inverse KL divergence with a smaller temperature is more likely to find the optimal action.

Temperature		0.1	0.3	0.5	0.7	0.9
movie	*Forword*	0.8282	0.8511	0.8302	0.8649	0.8379
	Reverse	0.8641	0.8567	0.8651	**0.8655**	0.8316
taxi	*Forword*	0.1795	0.5263	0.5912	0.6101	0.6027
	Reverse	**0.6674**	0.6432	0.6439	0.6587	0.6344
Diagnosis	*Forword*	0.6054	0.8007	0.8383	0.8625	0.8575
	Reverse	**0.9163**	0.9070	0.8873	0.8689	0.8651

Table 5: Experimental results on forward and reverse KL on different domains.

5 Conclusion

In this paper, we proposed a Transformer-based world model and constructed a causal reasoning module. We endow the dialog agent with a world model and causality, making its decisions more human-like with perception and reasoning. We do four main works for the policy learning of dialog agents: (i) Constructed an autoregressive Transformer-based world model for processing dialogue sequences and generating simulation data. (ii) Constructed a causal reasoning module to generate latent dialogue agent distribution of user intention interaction successfully experiences. (iii) Modified the network structure of direct deep reinforcement learning methods, empowering it to use both latent distributions and simulated data. (vi) Experiments demonstrate the effectiveness, high performance, and robustness, of both individual and integrated modules. In the future, improving the causal reasoning module via a nonlinear network is a valuable direction.

References

Olivier Bachem and Matthieu Geist. 2024. On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes. *arXiv:2306.13649*.

Abdelhakim Benechehab, Giuseppe Paolo, Albert Thomas, Maurizio Filippone, and Balázs Kégl. 2023. Multi-timestep models for model-based reinforcement learning. *arXiv preprint arXiv:2310.05672*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasi. 2018. Microsoft Dialogue Challenge: Building End-to-End Task-Completion Dialogue Systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026.

Alan Chan, Hugo Silva, Sungsu Lim, Tadashi Kozuno, A. Rupam Mahmood, and Martha White. 2022. Greedification Operators for Policy Optimization: Investigating Forward and Reverse KL Divergences. *Journal of Machine Learning Research*, 23:1–79.

Zhi Chen, Lu Chen, Xiaoyuan Liu, and Kai Yu. 2020. Distributed Structured Actor-Critic Reinforcement Learning for Universal Dialogue Management. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 28:2400–2411.

Chen Gao, Yu Zheng, Wenjie Wang, Fuli Feng, Xiangnan He, and Yong Li. 2024. Causal Inference in Recommender Systems: A Survey and Future Directions. *ACM Transactions on Information Systems*, 42(4):Article No. 88, Pages 1 – 32.

Anirudh Goyal and Yoshua Bengio. 2022. Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 478:2266.

Thomas L. Griffiths, Nick Chater, Charles Kemp, Amy Perfors, and Joshua B. Tenenbaum. 2010. Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14(8):357–364.

Xingwei He, Yeyun Gong, A. Long Jin, Weizhen Qi, Hang Zhang, Jian Jiao, Bartuer Zhou, Biao Cheng, Siu Ming Yiu, and Nan Duan. 2022. Metric-guided Distillation: Distilling Knowledge from the Metric to Ranker and Retriever for Generative Commonsense Reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 839–852.

Chenping Huang and Bin Cao. 2023. Learning dialogue policy efficiently through dyna proximal policy optimization. In *International Conference on Collaborative Computing: Networking, Applications and Worksharing*, pages 396–414.

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*, 20:318–334.

Qingfeng Lan, Yangchen Pan, Alona Fyshe, and Martha White. 2020. Maxmin Q-learning: Controlling the estimation bias of Q-learning. In *Proceedings of the 37th International Conference on Learning Representations*.

Qianlong Liu, Zhongyu Wei, Baolin Peng, Xiangying Dai, Huaixiao Tou, Ting Chen, Xuanjing Huang, and Kam-fai Wong. 2018. Task-oriented Dialogue System for Automatic Diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 201–207.

Keting Lu, Yan Cao, Xiaoping Chen, and Shiqi Zhang. 2023. Efficient Dialog Policy Learning With Hindsight , User Modeling , and Adaptation. *IEEE Transactions on Cognitive and Developmental Systems*, 15(2):395–408.

Shrikant Malviya, Piyush Kumar, Suyel Namasudra, and Uma Shanker Tiwary. 2022. Experience replay-based deep reinforcement learning for dialogue management optimisation. *ACM Transactions on Asian and Low-Resource Language Information Processing*, page Just Accepted.

Yutaka Matsuo, Yann LeCun, Maneesh Sahani, Doina Precup, David Silver, Masashi Sugiyama, Eiji Uchibe, and Jun Morimoto. 2022. Deep learning, reinforcement learning, and world models. *Neural Networks*, 152:267–275.

Vincent Micheli, Eloi Alonso, and François Fleuret. 2023. Transformer-based World Models Are Happy With 100k Interactions. In *The Eleventh International Conference on Learning Representations*.

Mirco Mutti, Riccardo De Santi, Emanuele Rossi, Juan Felipe Calderon, Michael Bronstein, and Marcello Restelli. 2023. Provably efficient causal model-based reinforcement learning for systematic generalization. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, volume 37, pages 9251–9259.

Baolin Peng, Xiujun Li, Jianfeng Gao, Jingjing Liu, Yun-nung Chen, and Kam-fai Wong. 2018a. Adversarial advantage actor-critic model for task-completion dialogue policy learning. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6149–6153.

Baolin Peng, Xiujun Li, Jianfeng Gao, Jingjing Liu, Kam-Fai Wong, and Shang-Yu Su. 2018b. Deep dyna-q: Integrating planning for task-completion dialogue policy learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2182–2192.

Thomas Pouncy and Samuel J. Gershman. 2022. Inductive biases in theory-based reinforcement learning. *Cognitive Psychology*, 138:101509.

Junyan Qiu, Haidong Zhang, and Yiping Yang. 2023. Reward estimation with scheduled knowledge distillation for dialogue policy learning. *Connection Science*, 35(1):2174078.

Govardana Sachithanandam Ramachandran, Kazuma Hashimoto, and Caiming Xiong. 2022. Causal-aware Safe Policy Improvement for Task-oriented dialogue. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 92–102.

Mahdin Rohmatillah and Jen-Tzung Chien. 2023. Hierarchical Reinforcement Learning With Guidance for Multi-Domain Dialogue Policy. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:748–761.

Oliver Schulte and Pascal Poupart. 2024. Why online reinforcement learning is causal. *arXiv preprint arXiv:2403.04221*.

Frank Shih and Faming Liang. 2024. Fast Value Tracking for Deep Reinforcement Learning. *arXiv:2403.13178*.

Steven Sloman. 2005. *Causal Models: How People Think about the World and Its Alternatives*. Oxford University Press.

Chang Tian, Wenpeng Yin, and Marie Francine Moens. 2022. Anti-overestimation dialogue policy learning for task-completion dialogue system. In *Proceedings of the Association for Computational Linguistics: NAACL Findings*, pages 565–577.

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. Curran Associates, Inc.

Lingxiao Wang, Zhuoran Yang, and Zhaoran Wang. 2021. Provably efficient causal reinforcement learning with confounded observational data. In *Advances in Neural Information Processing Systems*, volume 34, pages 21164–21175.

Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Van Hasselt, Marc Lanctot, and Nando De Frcitas. 2016. Dueling network architectures for deep reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pages 1995–2003.

Zizhao Wang, Xuesu Xiao, Zifan Xu, Yuke Zhu, and Peter Stone. 2022. Causal dynamics learning for task-independent state abstraction. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 23151–23180.

Xin Wen, Weizhi Nie, Jing Liu, Yuting Su, Yongdong Zhang, and An An Liu. 2024. CDCM: ChatGPT-Aided Diversity-Aware Causal Model for Interactive Recommendation. *IEEE Transactions on Multimedia*, 26:6488–6500.

Yuexin Wu, Xiujun Li, Jingjing Liu, Jianfeng Gao, and Yiming Yang. 2019. Switch-Based Active Deep Dyna-Q: Efficient Adaptive Planning for Task-Completion Dialogue Policy Learning. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, pages 7289–7296.

Kai Xu, Zhengyu Wang, Yuxuan Long, and Qiaona Zhao. 2024. Deep reinforcement learning-based dialogue policy with graph convolutional Q-network. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4555–4565.

Zhongwei Yu, Jingqing Ruan, and Dengpeng Xing. 2023. Explainable reinforcement Learning via a causal world model. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence*, pages 4540–4548.

Haodi Zhang, Zhichao Zeng, Keting Lu, Kaishun Wu, and Shiqi Zhang. 2022. Efficient Dialog Policy Learning by Reasoning with Contextual Knowledge. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, pages 11667–11675.

Zhirui Zhang, Xiujun Li, Jianfeng Gao, and Enhong Chen. 2020. Budgeted policy learning for task-oriented dialogue systems. In *Proceedings of the Conference 57th Annual Meeting of the Association for Computational Linguistics*, pages 3742–3751.

Yangyang Zhao, Zhenyu Wang, and Zhenhua Huang. 2021a. Automatic curriculum learning with over-repetition penalty for dialogue policy learning. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, pages 14540–14548.

Yangyang Zhao, Zhenyu Wang, Kai Yin, Rui Zhang, Zhenhua Huang, and Pei Wang. 2020. Dynamic reward-based dueling deep dyna-Q : Robust policy learning in noisy environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9676–9684.

Yangyang Zhao, Zhenyu Wang, Changxi Zhu, and Shihan Wang. 2021b. Efficient dialogue complementary policy learning via deep Q-network policy and episodic memory policy. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4311–4323.

Yangyang Zhao, Kai Yin, Zhenyu Wang, Mehdi Dastani, and Shihan Wang. 2024. Decomposed deep q-network for coherent task-oriented dialogue policy learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:1380–1391.

A Pseudo-code for proposed Model

The Transformer-based world model for generate simulated samples is described in Algorithm 1. We proposed the model-based causal dialog agent described in Algorithm 2.

Algorithm 1 Transformer-based world model generate simulated samples

initialize the planning step G and the maximum dialogue turns L, batch size bs
randomly sample batch sizes bs from successful interaction experiences, obtaining a set $H \in (H_0, H_1, ..., H_{bs})$
update world model parameters via Z-step minibatch SGD of multi-task learning
for $g = 1$ to G **do**
  termination flag d_t= 0, dialogue turn t=0
  sample a user goal, obtain user action a^u and generate an initial dialogue state s_t
  while not d_t and $t < L$ **do**
    with probability ϵ select a random action a_t, otherwise select $a_t = arg_{a'} max Q_\theta(s, a')$
    generate sequential H_i^t according to Eq. (1)
    world model responds with a_{t+1}^u, r_t and d_t
    update dialogue state to s_{t+1} according to a_{t+1}^u
    store $(s_t, a_t, r_t, s_{t+1}, d_t)$ to D^s
    set $t = t + 1$, $s_t = s_{t+1}$
  end while
end for

B Implementation details

The hidden layer size in deep reinforcement learning is 80 and an activation function of $tanh$. The discount factor for the reinforcement learning reward is $\gamma = 0.9$. On the movie ticket booking dataset, the size of the memory pool D^u is 5,000, and the size of the simulated experience pool D^s is 5,000. On the taxi calling dataset, the size of D^u is 10,000, and the size of D^s is 10,000. The experience pool is $D^u = 10,000$, and the simulated experience pool is $D^s = 10,000$ in the diagnosis domain. The model optimizer for value-based reinforcement learning uses $RMSprop$, the learning rate is 0.001, and the batch size is 16. The proposed method is implemented via proximal policy optimization (PPO) methods, and the learning rates are 0.0003 and 0.001 for the Actor and Critic network, respectively. The clipping value is 0.2 in the movie domain, 0.6 in the taxi domain, the α parameter of entropy regularization is 0.01, and β for distilling the episodic memory distribution is 0.9. The optimizer for updating the Actor and

Algorithm 2 Proposed Model-based Causal dialogue Agent

for each episode **do**
  initialize raw state s_1
  for $t = 1$ to L **do**
    select an action a_t according to the probability of the actor network.
    execute action a_t, receive environment rewards r_t and come into next state s_{t+1}, observe user response a_t^u, update task completed signal f
    store (s_t, a_t, r_t, s_{t+1}) to D^u
    set $s_t = s_{t+1}$
  end for
  obtaining τ based on D^u and calculating the $p(s_t^u)$ according to Eq. 11.
  obtain D^s by executing Algorithm 1
  sample random minibatch of (s_t, a_t, r_t, s_{t+1}) from D^s and D^u, respectively
  execute a gradient descent step via Eq. 12 with D^s and D^u, and update network parameter θ
end for

Critic network is $Adam$. The maximum dialogue length L is 40 in the movie and taxi domains. In the diagnosis domain, $L = 26$. All user requests completed within L turns are considered dialogue successful. The agent receives a reward of $2 * L$ when they complete all of the user's requests within L. If the agent fails to complete the requests or the dialogue turns over L, it receives a reward of $-L$. Additionally, the agent obtains a reward of -1 for each time step to incentivize the task completed. All world models use a planning step 5. The number of DQNs in MAXMIN is 5. The balance weight of the DPAV algorithm is 0.75 in the movie domain and 0.2 in the taxi domain, 0.6 for the automatic diagnosis domain The adaptive reward parameter of DR-D3Q is 0.05 on the movie domain, 0.9 on the taxi domain, and 0.4 on the disease diagnosis domain. The proposed autoregressive Transformer-based world model has a timestep parameter of $L/2$, the embedding dimension is 256, the layers are 10, the attention heads are 4, the weight decay is 0.01, the embedding dropout is 0.1, the attention dropout is 0.1, and the residual dropout is 0.1. All experiment results are averaged in 5 tests. The user simulator used in our experiments is the same as (Zhao et al., 2024) and (Xu et al., 2024).

	movie			taxi			diagnosis		
	AUS	*AUR*	*Overall*	*AUS*	*AUR*	*Overall*	*AUS*	*AUR*	*Overall*
DQN	0.4765	0.5132	0.4949	<u>0.4149</u>	<u>0.4625</u>	<u>0.4387</u>	0.6070	0.7633	0.6852
Dueling	0.5004	0.5276	0.5140	0.2173	0.2525	0.2349	0.4324	0.6506	0.5415
DRQN	0.4719	0.5031	0.4875	0.2761	0.3104	0.2933	0.2941	0.5860	0.4401
MAXMIN	0.4145	0.4625	0.4385	0.3106	0.3670	0.3388	0.4876	0.6939	0.5908
DPAV	0.5706	0.6153	0.5930	0.3585	0.4014	0.3800	0.5335	0.7219	0.6277
DDQ	<u>0.7144</u>	<u>0.7543</u>	<u>0.7344</u>	0.3202	0.3596	0.3399	<u>0.6432</u>	<u>0.7829</u>	<u>0.7131</u>
DR-D3Q	0.4590	0.4769	0.4680	0.3607	0.4040	0.3824	0.5298	0.7072	0.6185
LKTD	0.568	0.5939	0.5810	0.2142	0.2358	0.2250	0.0921	0.3955	0.2438
MCA	**0.8619**	**0.8663**	**0.8641**	**0.6526**	**0.6821**	**0.6674**	**0.8924**	**0.9402**	**0.9163**

Table 6: The experiment results in three task-oriented domains. The best results are indicated in bold, and the second-best results are underlined. $Overall$ is the average of the two metrics AUS and AUR.

C Detailed Experimental Results

The detailed performance of three task-oriented dialogue domains, see table 6. The proposed MCA algorithm achieves the best results in all three domains.DDQ overall achieves the second-best results in the domains of movie and diagnosis, and DQN maintains the second-best results in the domain of taxi. The experimental results demonstrate the superiority of the MCA algorithm.