In the following, we first present the proof of Theorem 1. After that, more experimental details about the synthetic experiments and real-world applications are given.

# A Proof of Theorem 1

Recall that we consider solving a linear regression problem under $m$ strong convex and $M$-smooth. MSE loss function $L_f(\boldsymbol{w})$ in the formulation $\|\boldsymbol{w}^\top \cdot (\boldsymbol{s}_{t-1}, \boldsymbol{a}_{t-1}) - \boldsymbol{s}_t\|_2^2$ with $\boldsymbol{s}_t = \boldsymbol{w}^{\star\top} \cdot (\boldsymbol{s}_{t-1}, \boldsymbol{a}_{t-1})$ is minimized by gradient descent method with

$$\boldsymbol{w}(k) = \boldsymbol{w}(k-1) - \alpha \nabla L_f(\boldsymbol{w}(k-1)) \tag{13}$$

where $\alpha$ is the corresponding step size.

Denote $\boldsymbol{w}^c(k)$ as the network parameters taking causal constraints with respect to $\boldsymbol{w}(k)$ that does not gather causal information. Theorem 1 shows that causal exploration gets a prediction error bound $\delta^k$ times lower at the $k$-th step, where $\delta$ is a density measurement of the causal adjacency matrix $D$.

Below, we present a two-step proof for the convergence of causal exploration. Lemma 2 provides an upper bound for convergence without using any causal information. Lemma 3 demonstrates that utilizing causal structure information results in $\boldsymbol{w}^c(k)$ being closer to the optimal value $\boldsymbol{w}^\star$ compared to $\boldsymbol{w}(k)$ at the same optimization steps. Combining Lemma 2 and Lemma 3, we derive the convergence rate for causal exploration.

**Lemma 2.** *Suppose $L_f(\boldsymbol{w})$ is $m$-strongly convex and $M$-smooth. We have*

$$\|\boldsymbol{w}(k) - \boldsymbol{w}^\star\|^2 \leq (1 - \frac{m}{M})^k \|\boldsymbol{w}_0 - \boldsymbol{w}^\star\|^2. \tag{14}$$

*Proof.* According to gradient descent method (13), we get

$$\begin{aligned}
&\|\boldsymbol{w}(k) - \boldsymbol{w}^\star\|^2 = \|\boldsymbol{w}(k-1) - \alpha \nabla L_f(\boldsymbol{w}(k-1)) - \boldsymbol{w}^\star\|^2 \\
&= \|\boldsymbol{w}(k-1) - \boldsymbol{w}^\star\|^2 + \alpha^2 \|\nabla L_f(\boldsymbol{w}(k-1))\|^2 \\
&\quad - 2\alpha \nabla L_f(\boldsymbol{w}(k-1))(\boldsymbol{w}(k-1) - \boldsymbol{w}^\star).
\end{aligned}$$

By strong convexity

$$\nabla L_f(\boldsymbol{w})(\boldsymbol{w} - \boldsymbol{w}^\star) \geq L_f(\boldsymbol{w}) - L_f(\boldsymbol{w}^\star) + \frac{m}{2}\|\boldsymbol{w} - \boldsymbol{w}^\star\|^2, \tag{15}$$

we further obtain

$$\begin{aligned}
&\|\boldsymbol{w}(k) - \boldsymbol{w}^\star\|^2 \leq \|\boldsymbol{w}(k-1) - \boldsymbol{w}^\star\|^2 - 2\alpha(L_f(\boldsymbol{w}(k-1)) \\
&- L_f(\boldsymbol{w}^\star) + \frac{m}{2}\|\boldsymbol{w}(k-1) - \boldsymbol{w}^\star\|^2) + \alpha^2 \|\nabla L_f(\boldsymbol{w}(k-1))\|^2 \\
&= \|\boldsymbol{w}(k-1) - \boldsymbol{w}^\star\|^2 - 2\alpha(L_f(\boldsymbol{w}(k-1)) - L_f(\boldsymbol{w}^\star)) \\
&\quad - \alpha m \|\boldsymbol{w}(k-1) - \boldsymbol{w}^\star\|^2 + \alpha^2 \|\nabla L_f(\boldsymbol{w}(k-1))\|^2 \\
&\leq \|\boldsymbol{w}(k-1) - \boldsymbol{w}^\star\|^2 - 2\alpha(L_f(\boldsymbol{w}(k-1)) - L_f(\boldsymbol{w}^\star)) \\
&\quad - \alpha m \|\boldsymbol{w}(k-1) - \boldsymbol{w}^\star\|^2 \\
&\quad + 2\alpha^2 M(L_f(\boldsymbol{w}(k-1)) - L_f(\boldsymbol{w}^\star)) \\
&\leq \|\boldsymbol{w}(k-1) - \boldsymbol{w}^\star\|^2 - \alpha m \|\boldsymbol{w}(k-1) - \boldsymbol{w}^\star\|^2 \\
&\quad + 2\alpha(\alpha M - 1)(L_f(\boldsymbol{w}(k-1)) - L_f(\boldsymbol{w}^\star)).
\end{aligned} \tag{16}$$

Consider $\alpha = \frac{1}{M}$, we get

$$\|\boldsymbol{w}(k) - \boldsymbol{w}^\star\|^2 \leq (1 - \frac{m}{M})\|\boldsymbol{w}(k-1) - \boldsymbol{w}^\star\|^2. \tag{17}$$

Using the above equation repeatedly, we obtain

$$\begin{aligned}
\|\boldsymbol{w}(k) - \boldsymbol{w}^\star\|^2 &\leq (1 - \frac{m}{M})\|\boldsymbol{w}(k-1) - \boldsymbol{w}^\star\|^2 \\
&\leq (1 - \frac{m}{M})^2 \|\boldsymbol{w}(k-2) - \boldsymbol{w}^\star\|^2 \\
&\leq \cdots \leq (1 - \frac{m}{M})^k \|\boldsymbol{w}_0 - \boldsymbol{w}^\star\|^2.
\end{aligned} \tag{18}$$

□

**Lemma 3.** *Suppose $\boldsymbol{w}^c(k)$ and $\boldsymbol{w}(k)$ are the network parameters with/without causal structure respectively and $\boldsymbol{w}^\star$ is the optimum. It holds that*

$$\|\boldsymbol{w}^c(k) - \boldsymbol{w}^\star\|^2 \leq \delta_k \|\boldsymbol{w}(k) - \boldsymbol{w}^\star\|^2. \tag{19}$$

*Proof.* According to Definition 1, we have

$$\boldsymbol{w}^c(k) = D \odot \boldsymbol{w}(k), \tag{20}$$

where $D$ is the binary causal matrix. Hence, we rewrite $\boldsymbol{w}(k)$ as

$$\boldsymbol{w}(k) = D \odot \boldsymbol{w}(k) + (1 - D) \odot \boldsymbol{w}(k) = \boldsymbol{w}^c(k) + \boldsymbol{w}'_k. \tag{21}$$

Note that we have $\boldsymbol{w}^\star = D \odot \boldsymbol{w}^\star$. Then we obtain

$$\begin{aligned}
&\|\boldsymbol{w}(k) - \boldsymbol{w}^\star\|^2 = \sum_{i,j} (w_{ij}(k) - w^\star_{ij})^2 \\
&= \sum_{i,j} \left[D_{ij} \times w_{ij}(k) - w^\star_{ij} + (1 - D_{ij}) \times w_{ij}(k)\right]^2 \\
&= \sum_{i,j} \left[(D_{ij} \times w_{ij}(k) - w^\star_{ij})^2 + ((1 - D_{ij}) \times w_{ij}(k))^2\right] \\
&= \sum_{i,j} (D_{ij} \times w_{ij}(k) - w^\star_{ij})^2 + \sum_{i,j} ((1 - D_{ij}) \times w_{ij}(k))^2 \\
&= \sum_{i,j} (w^c_{ij}(k) - w^\star_{ij})^2 + \sum_{i,j} (w'_{ij})(k)^2 \\
&= \|\boldsymbol{w}^c(k) - \boldsymbol{w}^\star\|^2 + \|\boldsymbol{w}'_k\|^2 \geq (1 + \rho_k)\|\boldsymbol{w}^c(k) - \boldsymbol{w}^\star\|^2,
\end{aligned} \tag{22}$$

where $\rho_k \in [0, +\infty)$ is the lower bound of the ratio between $\|\boldsymbol{w}'_k\|^2$ and $\|\boldsymbol{w}^c(k) - \boldsymbol{w}^\star\|^2$ whose value is related to the sparsity of causal matrix $D$. By setting $\delta_k = \frac{1}{1+\rho_k}$, we complete the proof of Lemma 3. □

The convexity of $L_f(\boldsymbol{w})$ implies

$$L_f(\boldsymbol{w}^c(k)) - L_f(\boldsymbol{w}^\star) \leq \frac{M}{2}\|\boldsymbol{w}^c(k) - \boldsymbol{w}^\star\|^2. \tag{23}$$

By applying Lemma 2 and Lemma 3 into (23) and denote

$$\delta = \max\{\delta_0, \delta_1, \ldots, \delta_k\}, \tag{24}$$

we can obtain Theorem 1.

# B Synthetic environment

## B.1 More experiment details

**Implementation.** Double DQN [Van Hasselt *et al.*, 2016] is used to train the exploration policy of agents, where both the evaluation network and the target network are three-layer

| Method | No Sampling | Uniform | K-center | Hardest | Coreset |
|---|---|---|---|---|---|
| AUC | $0.907 \pm 0.084$ | $0.892 \pm 0.065$ | $0.891 \pm 0.175$ | $0.816 \pm 0.353$ | $\mathbf{0.969 \pm 0.016}$ |
| Precision | $0.953 \pm 0.074$ | $0.914 \pm 0.079$ | $0.942 \pm 0.089$ | $0.892 \pm 0.081$ | $\mathbf{0.970 \pm 0.064}$ |
| F1-score | $0.831 \pm 0.094$ | $0.842 \pm 0.104$ | $0.890 \pm 0.139$ | $0.852 \pm 0.067$ | $\mathbf{0.928 \pm 0.096}$ |
| Time | $4.360 \pm 2.888$ | $0.880 \pm 0.453$ | $0.909 \pm 0.414$ | $1.389 \pm 0.760$ | $\mathbf{0.379 \pm 0.267}$ |

Table 3: Results of different selection methods on online causal discovery.

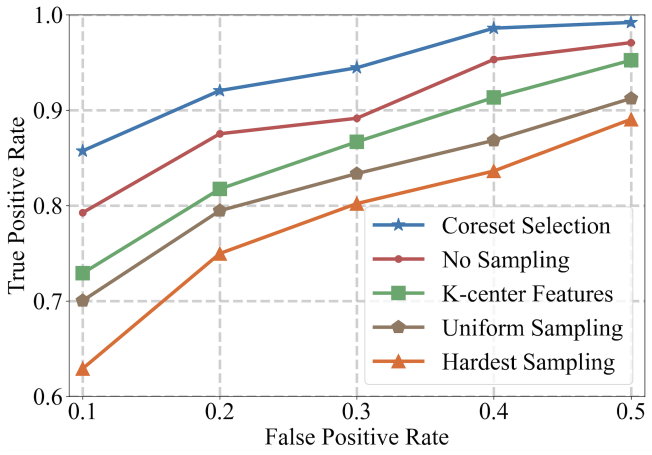



Figure 8: ROC curves of online causal discovery using different sampling methods.

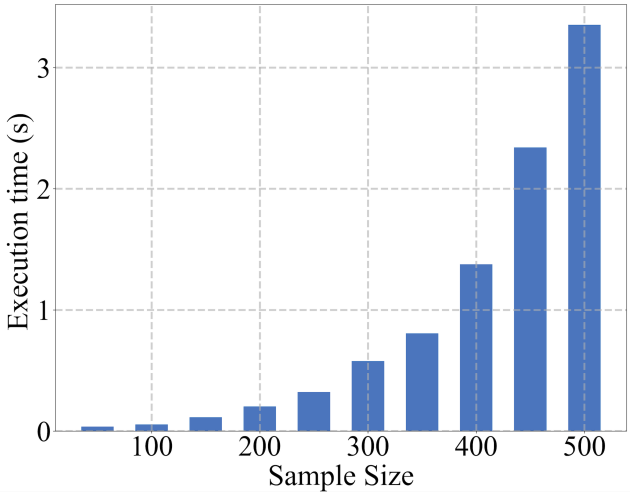



Figure 9: Time cost for the PC algorithm when $|\mathcal{U}| = 12$ and $|\mathcal{V}| = 10$.

fully connected networks with relu activations. The corresponding world models are designed as MLPs without/with activation function in linear/nonlinear module respectively, with 32 and 8 hidden nodes. In all of these settings, we share network parameters except for the last layer which is separately decomposed. We use Adam optimizer to learn the above models with $\eta = 0.1$ and a learning rate of 1e-3.

**The $\kappa$ value for online causal discovery.** As can be seen in Figure 9, the execution time of the PC algorithm exhibits a noticeable inflection when the sample size is around 350. This provides a valuable reference for determining an appropriate value for the selection number $\kappa$. As a result, we empirically set $\kappa = 350$ in the synthetic environment and $\kappa = 0.7 \times |\mathcal{B}_t|$ for real-world applications, where $|\mathcal{B}_t|$ represents the number of collected data at time $t$.

**Efficient Causal Discovery.** To speed up PC algorithm with KCI-test, we design an efficient online causal discovery using selection methods. Figure 8 illustrates the corresponding ROC curves of different selection methods including Uniform Sampling, K-center Features [Nguyen *et al.*, 2017], Hardest Sampling [Rahaf and Lucas, 2019] and the Coreset Selection method [Yoon *et al.*, 2021] we use. Tables 3 summarizes the performance of different sampling methods in our synthetic environments. These exciting experimental results demonstrate the superiority of the Coreset Selection method we used in improving the efficiency of causal discovery.

**Causal Exploration Experiments.** The causal matrix in our synthetic environment is set as a lower triangular matrix whose elements are generated from the uniform distribution $[-8, 8]$. After that, for each edge in the graph, we randomly drop it out with an empirically chosen probability $1-p = 0.8$. Besides, the covariance $\Sigma$ is a diagonal matrix whose elements are randomly generated from $[0, 0.1]$. For both linear and nonlinear conditions, we have conducted sufficient experiments. Figure 10 shows the remaining experimental results that are not fully presented in the main body.

In order to enhance the agent's sensitivity to causal informative data, we design a novel form of active reward. Here we investigate the impact of $\beta$ in this new intrinsic reward formulation on causal exploration which is formulated as $r_t = r_t^i + \beta r_t^a$ in equation (7). Figure 11(a) illustrates the evolution of prediction error over training time for different values of $\beta$. We see that incorporating active reward for exploration continuously improves the performance of the world model. We set $\beta = 0.5$ for linear environments and $\beta = 3$ for both nonlinear settings and real-world applications according to the results in Figure 11(a), respectively.

## B.2 Generalization to underestimation scenarios

In some data-hungry scenarios, there may be insufficient data for causal discovery, leading to underestimation of the causal structure, which makes continuous data collection and causal structure correction important components. In other words, the causal structure inferred from causal discovery algorithms may deviate from the ground truth ones, which is particularly prone to occur under conditions of limited sample size or during the initial stages of exploration. Consequently, we conduct an evaluation of our proposed algorithm's performance

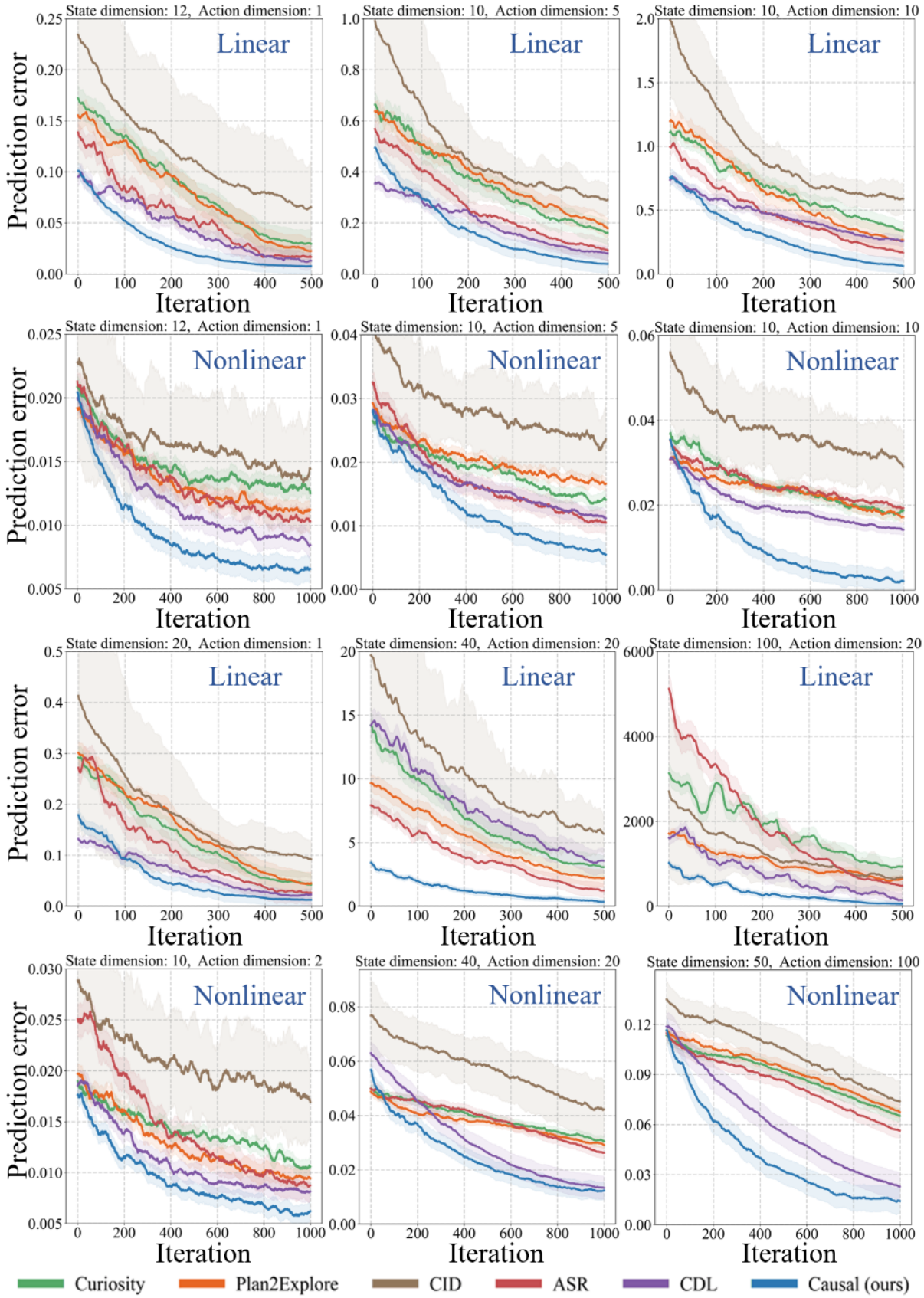
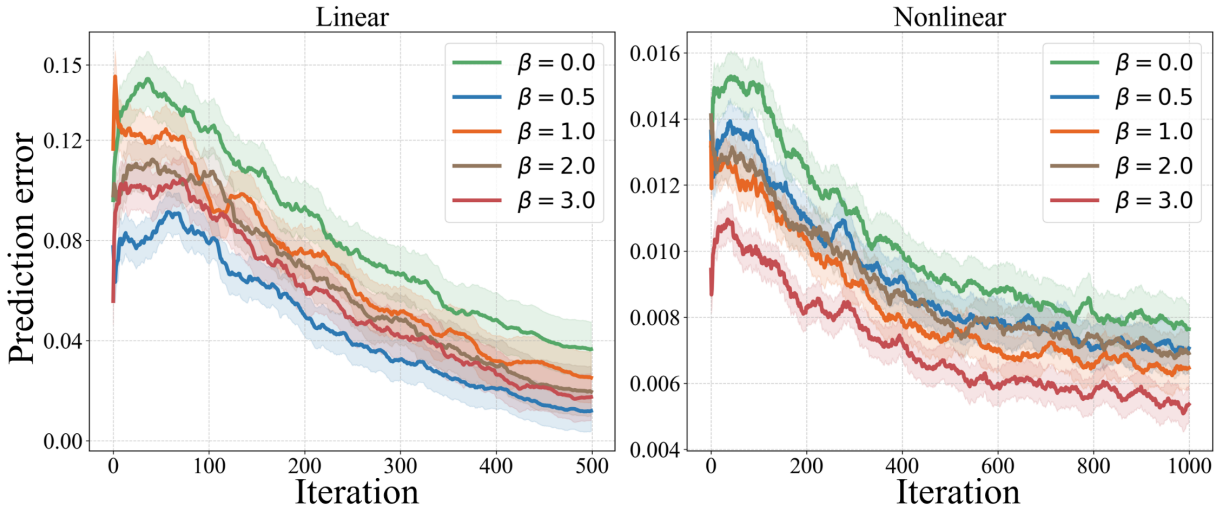
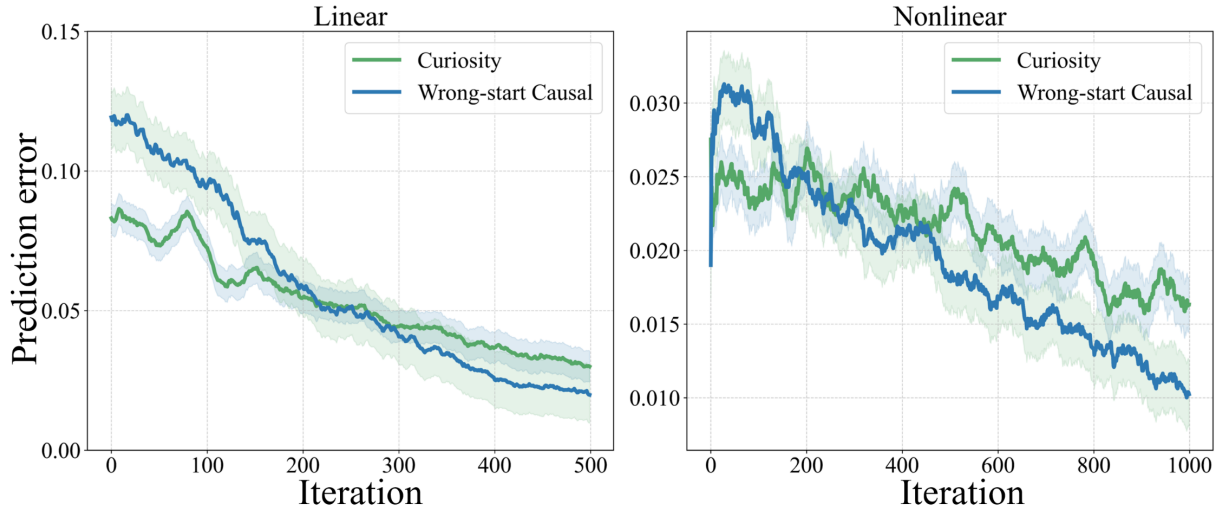


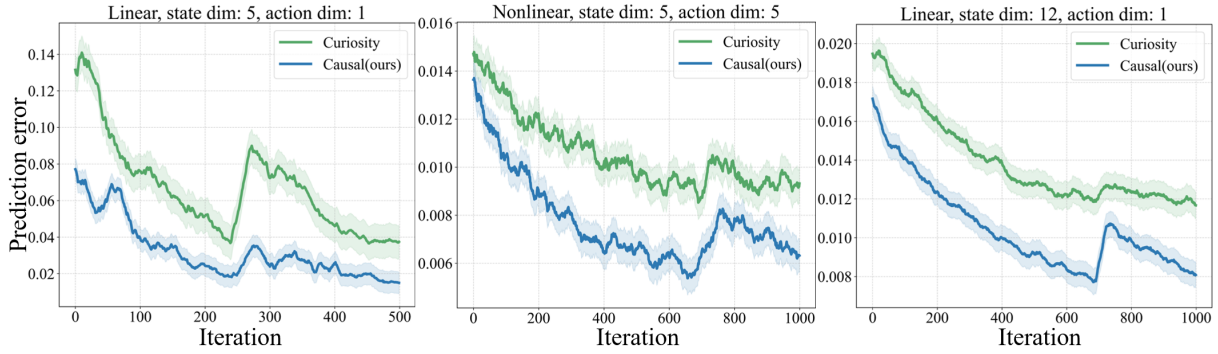

Figure 10: Results on synthetic datasets.




(a)

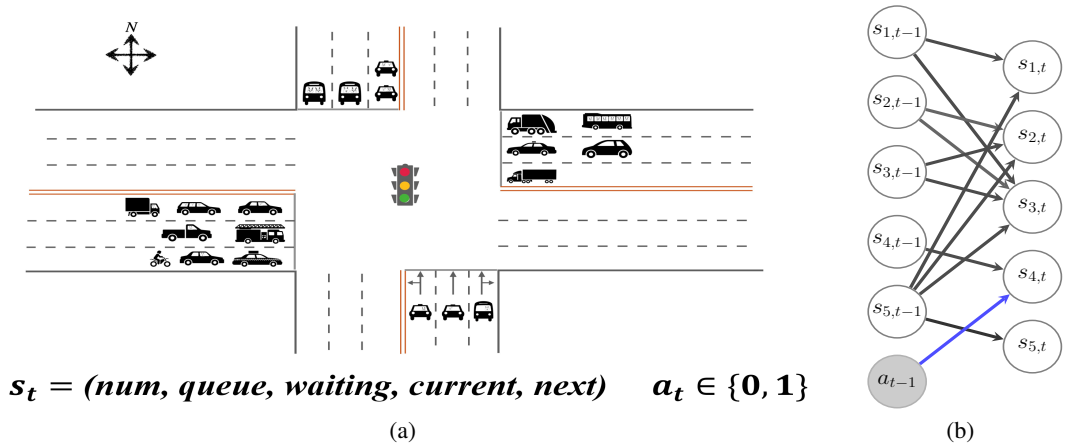



(b)




(c)

Figure 11: (a) Prediction errors of causal exploration for different $\beta$; (b) Performance on underestimation scenarios; (c) Scenarios with structural changes.




(a) (b)

Figure 12: (a) Illustration of the traffic-signal-control environment. (b) Causal graph learned through exploration.

when the estimated causal structure exhibits insufficiencies or redundancies, a scenario termed "underestimation". To highlight the advantages of our sharing-decomposition schema in addressing such problems, we deliberately provide the agent with completely wrong causal information. An erroneous causal graph $\mathcal{G}'$ is supplied to the agent during the initial time steps, namely $t < N$ where $N$ is the period for causal discovery. We introduce perturbations to the true graph $\mathcal{G}$ by randomly eliminating or adding edges with a probability of $p' = 0.8$, generating $\mathcal{G}'$. Figure 11(b) shows the corresponding performance of causal exploration on synthetic data. Indeed, the agent exhibits an impressive ability to correct its causal exploration trajectory in the face of misdirection. Nonetheless, it is worth noting that such a schema only serves as a technique to mitigate the impact of underestimation. The reliability of causal discovery algorithms is the essential guarantee for causal exploration.

### B.3 Generalization to scenarios with causal structural changes

In real-world scenarios, causal structure between variables can often change due to sudden disturbance. For instance, causal relationships between economic variables like stock prices, interest rates, and inflation can be subject to rapid changes caused by market crashes or policy changes.

To evaluate the effectiveness of our approach in handling such mutation, we conduct experiments in a scenario where the causal structure changes randomly once. We use our simulation model to generate the data and compare our method to a non-causal approach. Figure 11(c) illustrates the advantages of our approach in tackling such a challenging task.

Our sharing-decomposition schema enables the agents to quickly adapt to structural changes and make appropriate adjustments. This also demonstrates the robustness of our method, which allows for timely correction of errors in the causal structure. By sharing the same decomposition modules across different time steps and tasks, our method can effectively leverage previous knowledge and transfer it to new situations, while also being flexible enough to accommodate changes in the causal structure. In addition, the ability to adapt to changing causal structures can improve the generalization ability of our method, making it more applicable to a wider range of real-world tasks.

In our future research, we plan to expand our work to situations where changes occur within the model. In these cases, during the model learning phase, it becomes crucial to effectively detect these changes and promptly update the model. Additionally, when it comes to policy learning, a key challenge is determining the most suitable model to utilize. We may encounter entirely new models that have not been encountered before, adding an additional layer of complexity to our research.

## C Traffic Signal Control

Traffic signal control is an important means of mitigating congestion in traffic management. Compared to using fixed-duration traffic signals, an RL agent learns a policy to determine real-time traffic signal states based on current road conditions. The state observed by the agent at each time consists of five dimensions of information, namely the number of vehicles, queue length, average waiting time in each lane plus current and next traffic signal states. Action here is to decide whether to change the traffic signal state or not. For example, suppose the traffic signal is red at time $t$, if the agent takes action 1, then it will change to green at the next time $t + 1$, otherwise, it will remain red. Following the work in IntelliLight [Wei *et al.*, 2018], the traffic environment in our experiment is a three-lane intersection. Table 4 gives a detailed description, and Figure 12(a) provides an illustration. The estimated causal structures are given in Figure 12(b).

**Experiment details and analysis.** We first only use prediction-based causal exploration to learn forward dynamic world models under the same traffic environment in IntelliLight. Then, the agent learns a policy for traffic signal control task in our learned world models, which avoids the high-cost interaction with real traffic environment. For consistency and easy comparison, we use the same DQN network from IntelliLight to train our causal exploration agent.

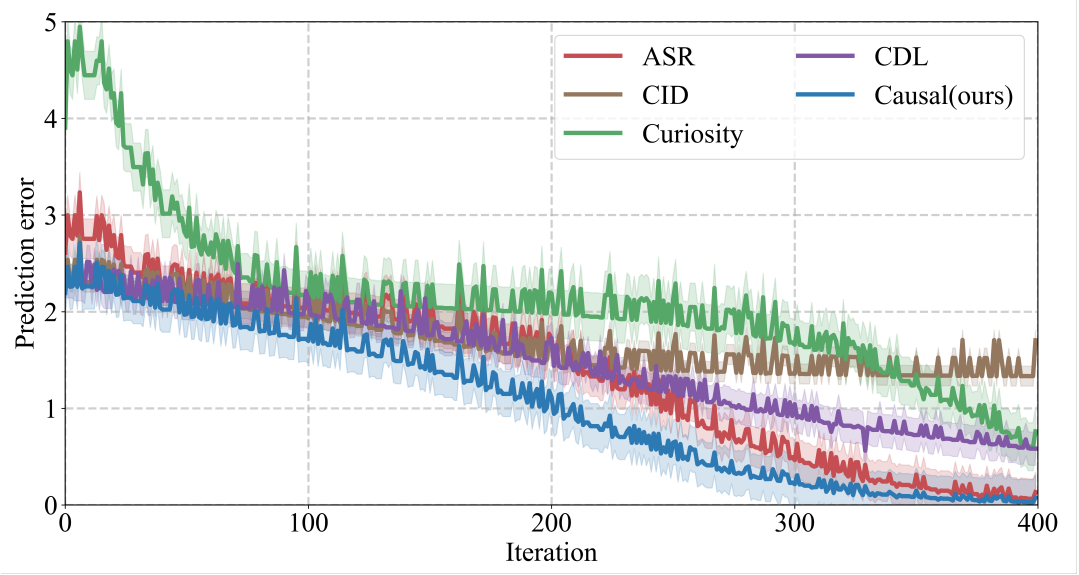



Figure 13: Prediction errors in traffic light control.

| Traffic flow setting | Directions | Arrival Rate (cars/s) | | Duration ($\times 10^3$ s) |
|---|---|---|---|---|
| | | Mean | Std | |
| Complex traffic | East-West | 0.211 | 0.023 | 216 |
| | South-North | 0.155 | 0.030 | |

Table 4: Traffic Dataset Description

After that, we solve the traffic signal control task in our world model with no environment interaction in a zero-shot manner. Figure 13 visualizes the prediction errors of different models during the exploration process. The corresponding causal graph is shown in Figure 12(b). As is illustrated, the state of the traffic signal at the next time step $s_{4,t}$ is causally linked to the previous state $s_{4,t-1}$ and action $a_{t-1}$, which is in line with the definition of traffic signal control tasks. The queue length $s_{2,t}$ is determined by previous queue length $s_{2,t-1}$ and waiting time $s_{3,t-1}$ plus the traffic state $s_{5,t-1}$. Factors influencing the waiting time $s_{3,t}$ include the number of vehicles $s_{1,t-1}$ and the queue length $s_{2,t-1}$. These results align well with the common-sense reasoning.

## D More Results of Mujoco tasks

We use PPO algorithm [Schulman *et al.*, 2017] for optimization during both the task-agnostic exploration and policy learning stages and adopt the hyperparameters from Table 3 of PPO with a trajectory length of 2048, an Adam stepsize of 3e-4, a minibatch size of 64, a discount factor ($\gamma = 0.99$), a GAE parameter ($\lambda = 0.95$), and a clipping parameter ($\epsilon = 0.2$). Both the actor-critic network and the world model are 2-(hidden)-layer neural networks, consisting of 256 and 64 hidden nodes respectively. Activation functions are Tanh and ReLU here.

Performance of causal exploration on some other MuJoCo tasks are provided in Figure 15. Predictions given by world models under causal structural constraints are more accurate and stable than those of other methods. The learned world model of causal exploration provides the agent with more information in the following policy learning stage, resulting in higher scores achieved in a shorter time. Figure 14 illustrates the identified causal structures during exploration, which explains for the performance gain.

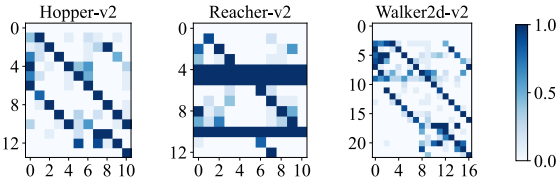
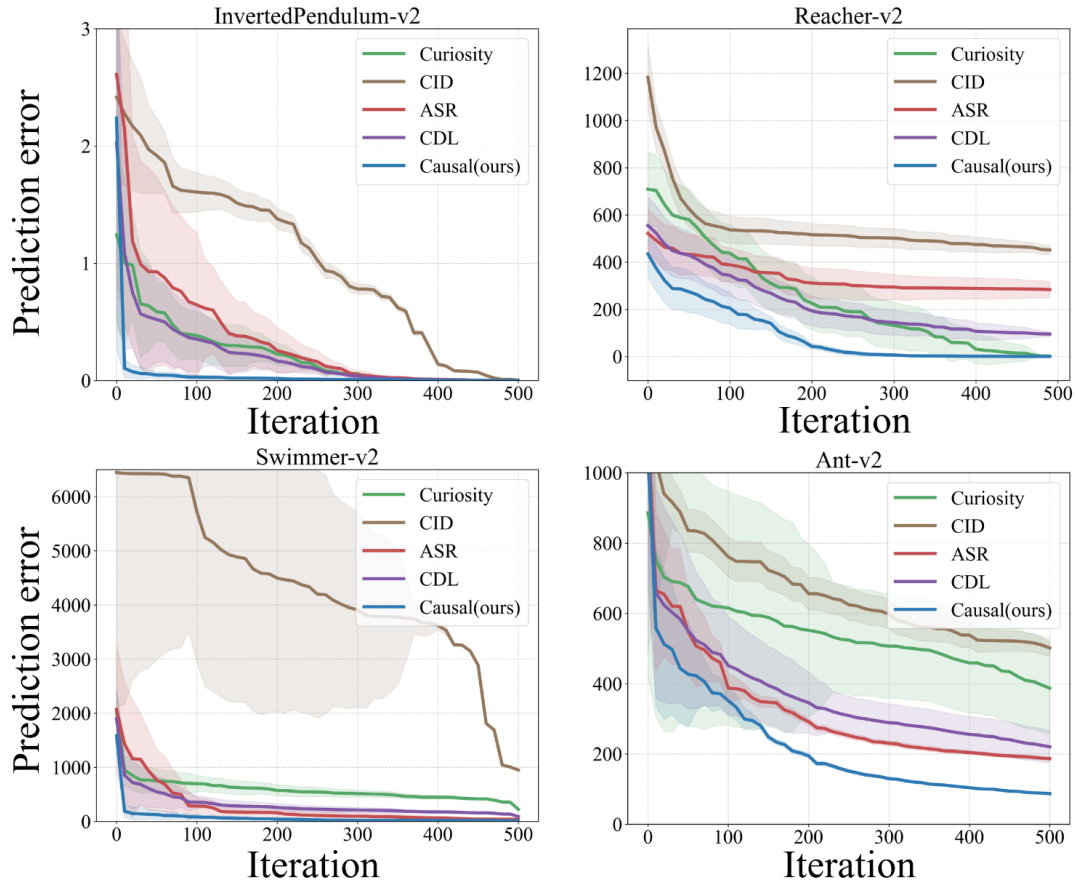



Figure 14: Identified causal structures for MuJoCo tasks.

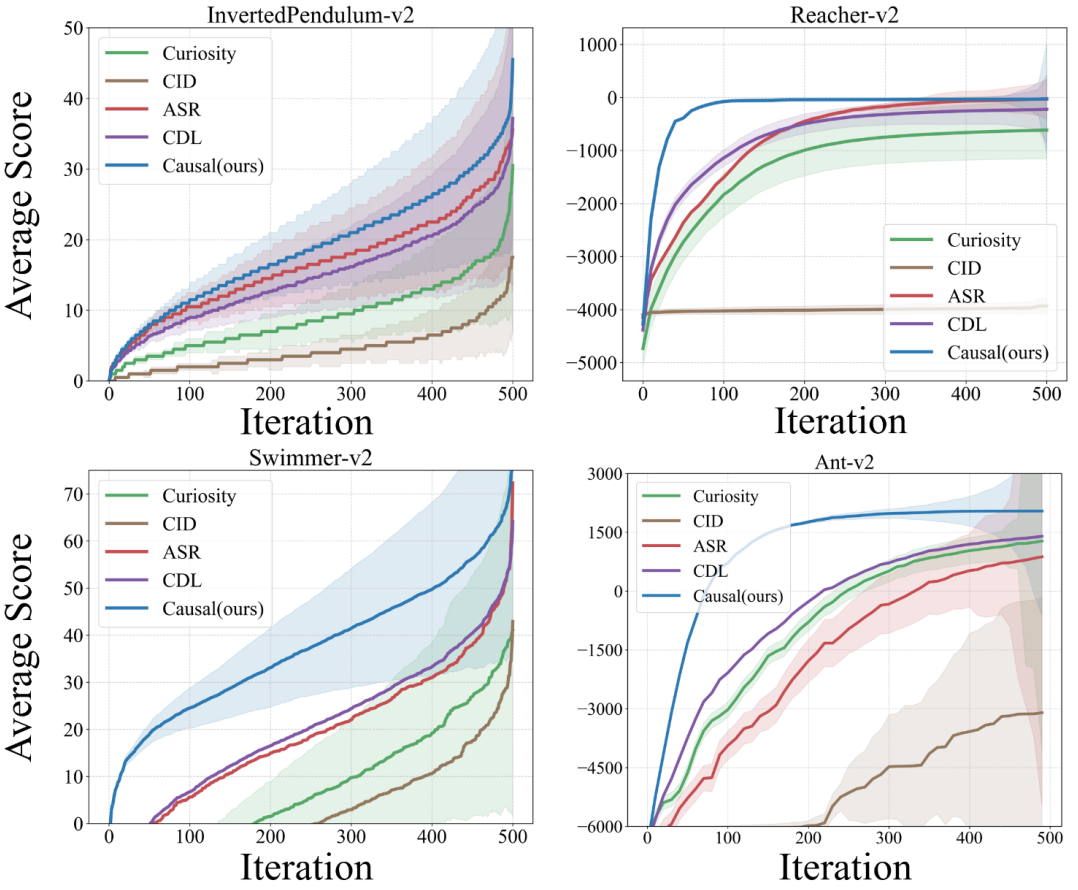
We also conduct several experiments to test the performance of causal exploration with a different form of intrinsic reward. To be specific, we formulate our world model as $\mu_{w^c}(\boldsymbol{s}_t, \boldsymbol{a}_t), \sigma^2_{w^c}(\boldsymbol{s}_t, \boldsymbol{a}_t)$ to model the transition probability as $p(\boldsymbol{s}_{t+1} \mid \boldsymbol{s}_t, \boldsymbol{a}_t) \sim \mathcal{N}(\boldsymbol{s}_{t+1}; \mu_{w^c}, \sigma^2_{w^c})$. Then, the negative log-likelihood is used both for the world model learning and causal exploration, which is a replacement for equation (3) and (5), and is formulated as:

$$L_{(\mu_{w^c}, \sigma^2_{w^c})} = \frac{(\boldsymbol{s}_{t+1} - \mu_{w^c}(\boldsymbol{s}_t, \boldsymbol{a}_t))^2}{2\sigma^2_{w^c}(\boldsymbol{s}_t, \boldsymbol{a}_t)} + \frac{1}{2} \log \sigma^2_{w^c}(\boldsymbol{s}_t, \boldsymbol{a}_t),$$
$$r^i_t = \frac{\eta}{2} L_{(\mu_{w^c}, \sigma^2_{w^c})}. \tag{25}$$

Corresponding results are shown in Figure 16. However, various forms of intrinsic rewards don't exhibit significant differences in performance. In some tasks, the introduction of an additional covariance network even lead to performance not as favorable as when directly using regression loss.

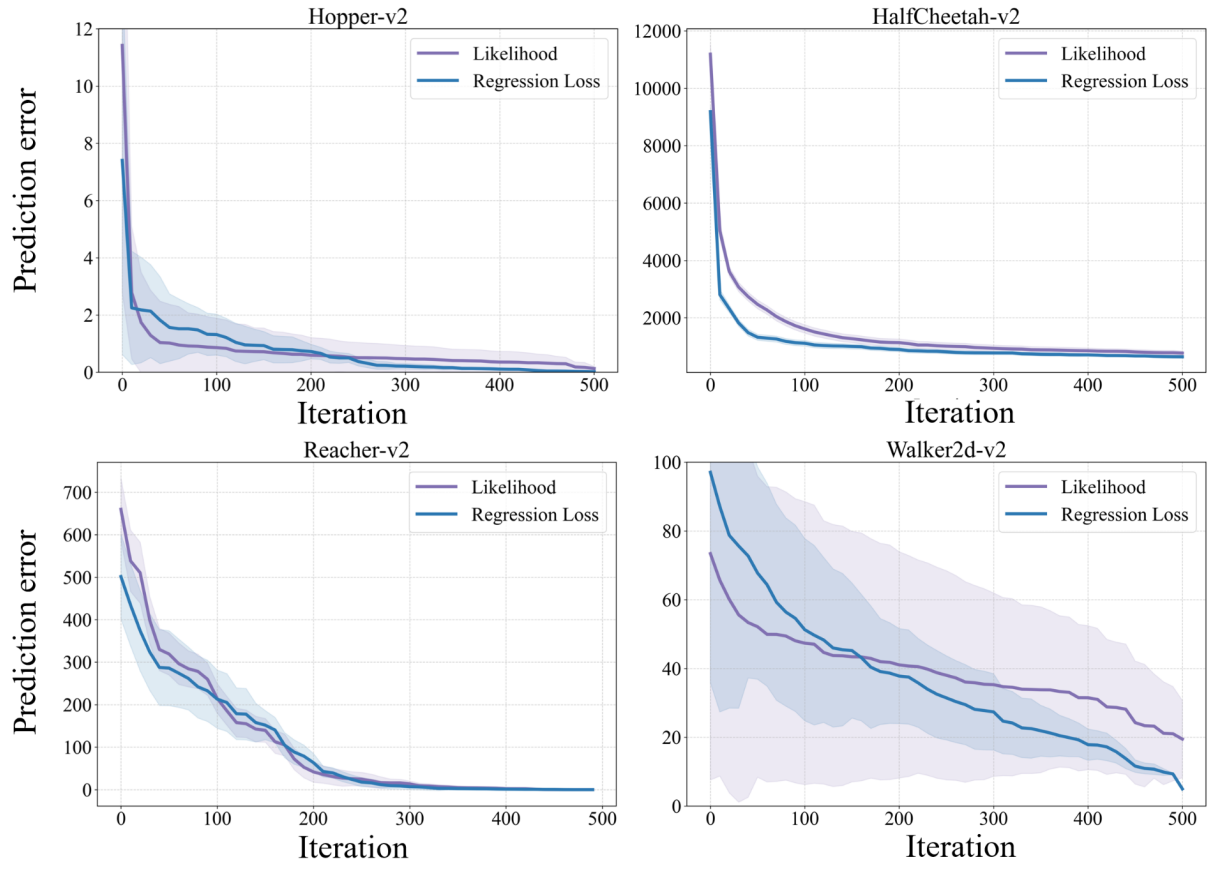



(a)

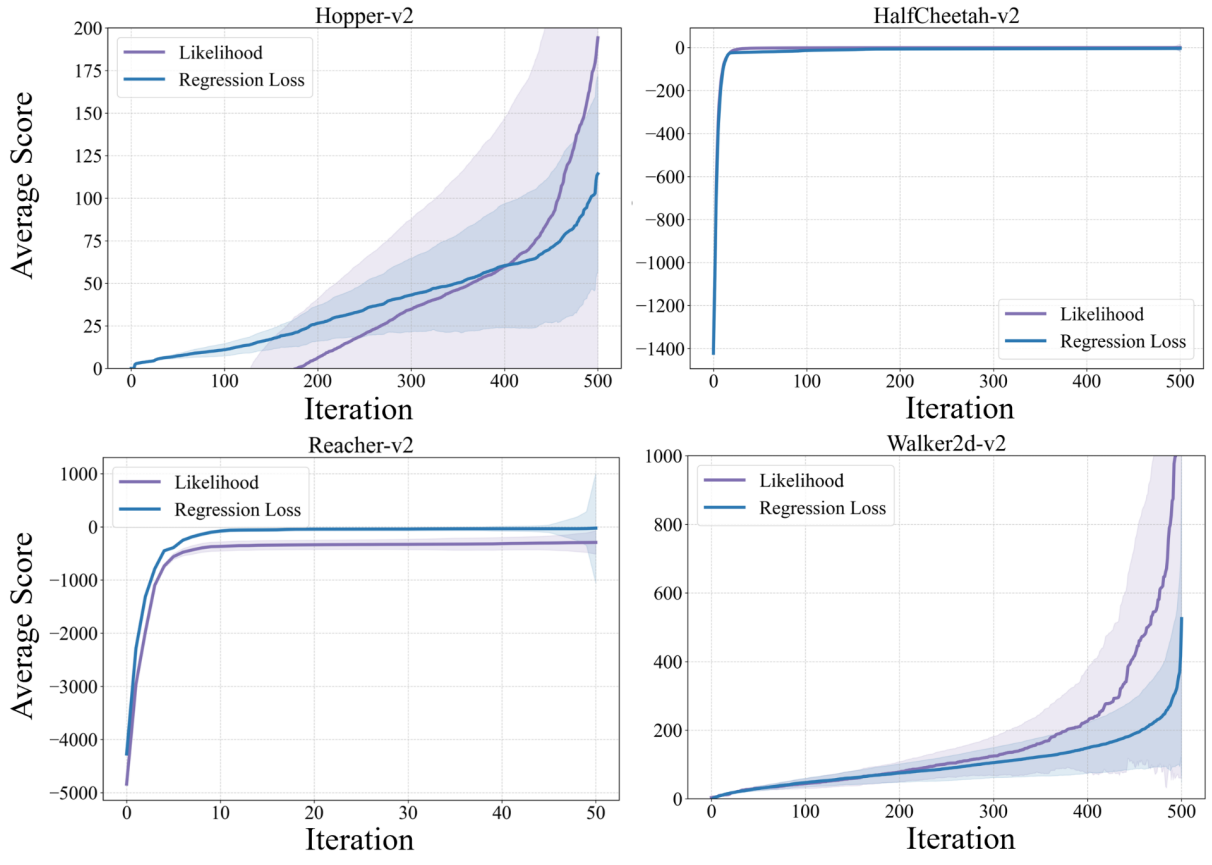



(b)

Figure 15: Application to some other MuJoCo tasks.




(a)




(b)

Figure 16: Performance of causal exploration with different forms of intrinsic rewards.