Towards Empowerment Gain through Causal Structure Learning in Model-Based RL

Hongye Cao[1], **Fan Feng**[2], **Meng Fang**[3], **Shaokang Dong**[1], **Jing Huo**[1]*, **Yang Gao**[1]
[1] National Key Laboratory for Novel Software Technology, Nanjing University
[2]City University of Hong Kong, [3]University of Liverpool
{hongyecao528,ffeng1017}@gmail.com, Meng.Fang@liverpool.ac.uk
shaokangdong@smail.nju.edu.cn, {huojing,gaoy}@nju.edu.cn

Abstract

Empowerment and causal reasoning are crucial abilities for intelligence. In reinforcement learning (RL), empowerment enhances agents' ability to actively control their environments by maximizing the mutual information between future states and actions. In model-based RL (MBRL), incorporating causal structures into dynamics models provides agents with a structured understanding of the environment to better control outcomes. We posit that learning causal world models can enhance agents' empowerment and, conversely, improved empowerment can facilitate causal reasoning. From this viewpoint, our goal is to enhance agents' empowerment, aiming to improve controllability and learning efficiency, and their ability to learn causal world models. We propose a framework, Empowerment through Causal Learning (ECL), where an agent with the awareness of causal models achieves empowerment-driven exploration and utilize its structured causal perception and control for task learning. Specifically, we first train a causal dynamics model of the environment based on collected data. We then maximize empowerment under the causal structure for policy learning, simultaneously updating the causal model to be more controllable than dynamics model without causal structure. An intrinsic curiosity reward is also included to prevent overfitting in offline model learning. Importantly, our framework is method-agnostic, capable of integrating various causal discovery and policy learning techniques. We evaluate ECL combined with 2 different causal discovery methods in 3 environments, demonstrating its superior performance compared to other causal MBRL methods, in terms of causal discovery, sample efficiency, and episodic rewards.

1 Introduction

Model-based reinforcement learning (MBRL) uses predictive models to enhance decision-making and planning [1]. Recent advances in integrating causal structures into MBRL have provided a more accurate description of systems, aiding adaptation and improving generalization amid environmental changes [2–7], spurious correlations [8–10], and systematic or compositional generalization challenges [11–15]. These studies show that agents with causal world models achieve robustness and adaptability across diverse scenarios. However, these methods often rely on *passively* using the learned or given causal structures to improve RL generalization.

Exploring how agents can *actively* leverage causal structure to better control the environment, aiming to improve controllability and learning efficiency, presents a compelling challenge. To measure the controllability and efficiency, we can employ empowerment gain as the intrinsic motivation for the

[*] Corresponding author.

ICML 2024 Workshop on Foundations of Reinforcement Learning and Control.

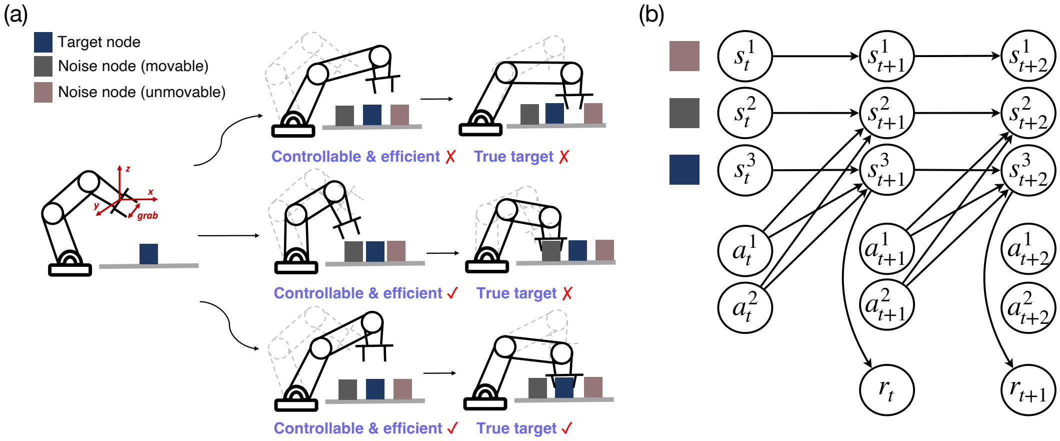


Figure 1: (a). Example of robot manipulation task with three trajectories and three nodes: one target node (movable) and two noisy nodes (one movable, one unmovable). (b). Underlying causal structures of the example.

agents. Empowerment is an information-theoretic framework where agents strive to maximize the mutual information between their actions and future states, conditioned on the initial state [16–20].

In this paper, we explore how to *actively* leverage causal structures to enhance empowerment, thereby improving learning efficiency and whether this, in turn, can improve the learning of the causal structure in RL environments. Conceptually, learning an accurate causal model for environments and improving empowerment are interdependent processes that reinforce each other. Causal models enable agents not only to predict but also to influence future states more effectively by utilizing variables that directly cause important state changes or reward maximization. Consequently, agents with causal world models are better positioned to manipulate state outcomes, resulting in a higher degree of control and efficiency in their actions. At the same time, by improving controllability, agents gain a better understanding of the consequences of their actions, thereby implicitly learning the causal model of their environment.

The given example (Fig.1(a)) discusses robot manipulation, where the goal is to move the target node (movable) while avoiding noisy nodes (some movable and some not). Three possible trajectories (rows 1-3) are shown with different levels of control, efficiency, and success in finding the target. Row 1 represents the least effective trajectory, while rows 2 and 3 indicate that the agent has learned control and efficiency (high empowerment, as these behaviors tend to movable objects). However, row 2 fails to find the target, whereas row 3 successfully identifies it. Assuming the agent has the causal structure between states and actions (Fig.1(b)), it will likely execute actions similar to rows 2-3 since there are causal relationships between actions and movable objects, effectively optimizing empowerment. If the agent also knows the causal relationship between states and rewards, it would further prioritize actions leading to the target object. Conversely, when optimizing empowerment, the agent implicitly learns that action sequences like rows 2 and 3 have a greater impact, facilitating efficient control and implicitly learning the causal state-action relationship.

From this viewpoint, we introduce an Empowerment through Causal Learning (ECL) framework that actively leverages causal structure to maximize empowerment, improving controllability and learning efficiency. The ECL framework consists of three main steps: offline model learning, online model learning, and policy learning. In offline model learning (step 1), we learn the causal dynamics model with causal mask and reward encoder. With the learned causal structure, we then integrate empowerment-driven exploration in online model learning (step 2), to better control the environment, by alternating the updates of the causal structure and policy of empowerment maximization. Finally, the learned causal structure is used to learn policies for down-streaming task with a curiosity reward to maintain robustness and prevent overfitting in model learning (step 3). Importantly, our framework is method-agnostic, able to integrate diverse causal discovery and policy learning techniques.

ECL not only refines policy learning but also ensures that the causal model remains adaptable and accurate, even in the face of novel or shifting environmental conditions. We evaluate ECL with two causal discovery techniques (conditional independence testing and regularization-based) across 3 environments, considering in-distribution and out-of-distribution settings. ECL outperforms other causal MBRL methods, showing remarkable performance with more accurate causal discovery, higher sample efficiency, and improved episodic rewards.

2 Preliminaries

2.1 MDP with Causal Structures

Markov Decision Process In RL, the interaction between the agent and the environment is formalized as an MDP. The standard MDP is defined by the tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, T, \mu_0, r, \gamma \rangle$, where $\mathcal{S}$ denotes the state space, $\mathcal{A}$ represents the action space, $T(s'|s,a)$ is the transition dynamic model, $r(s,a)$ is the reward function, and μ_0 is the distribution of the initial state s_0. The discount factor $\gamma \in [0,1)$ is also included. The objective of RL is to learn a policy $\pi : \mathcal{S} \times \mathcal{A} \to [0,1]$ that maximizes the expected discounted cumulative reward $\eta_{\mathcal{M}}(\pi) := \mathbb{E}_{s_0 \sim \mu_0, s_t \sim T, a_t \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right]$.

Structural Causal Model A *structural causal model* (SCM) [21] is defined by a distribution over random variables $\mathcal{V} = \{s_t^1, \cdots, s_t^d, a_t^1, \cdots, a_t^n, s_{t+1}^1, \cdots, s_{t+1}^d\}$ and a Directed Acyclic Graph (DAG) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with a conditional distribution $P(v_i|\mathrm{PA}(v_i))$ for node $v_i \in \mathcal{V}$. Then the distribution can be specified as:

$$p(v^1, \ldots, v^{|\mathcal{V}|}) = \prod_{i=1}^{|\mathcal{V}|} p(v^i|\mathrm{PA}(v_i)), \tag{1}$$

where $\mathrm{PA}(v_i)$ is the set of parents of the node v_i in the graph $\mathcal{G}$.

Causal Structures in MDP We use a dynamic Bayesian network (DBN)[22] (Fig.1b) denoted by $\mathcal{G}$, to model the MDP and the underlying causal structures between states, actions, and rewards. In the DBN, nodes represent system variables (different dimensions of the state, action, and rewards), while edges denote their relationships within the MDP. This model aligns with the factored MDPs [23, 24], and we employ causal discovery methodologies to learn the structures of $\mathcal{G}$. We have the Markov conditions and faithfulness assumptions and the assumptions on edges in MDP (A1-A4):

Assumption 1 *(Global Markov Condition [25, 21]) The state is fully observable and the dynamics is Markovian. The distribution p over a set of variables $\mathcal{V} = (s_t^1, \cdots, s_t^d, a_t^1, \cdots, a_t^d, r_t)^T$ satisfies the global Markov condition on the graph if for any partition $(\mathcal{S}, \mathcal{A}, \mathcal{R})$ in $\mathcal{V}$ such that if $\mathcal{A}$ d-separates $\mathcal{S}$ from $\mathcal{R}$, then $p(\mathcal{S}, \mathcal{R}|\mathcal{A}) = p(\mathcal{S}|\mathcal{A}) \cdot p(\mathcal{R}|\mathcal{A})$*

Assumption 2 *(Faithfulness Assumption [25, 21]) For a set of variables $\mathcal{V} = (s_t^1, \cdots, s_t^d, a_t^1, \cdots, a_t^d, r_t)^T$, there are no independencies between variables that are not implied by the Markovian Condition.*

Assumption 3 *Under the assumptions that the causal graph is Markov and faithful to the observations, the edge $s_t^i \to s_{t+1}^t$ exists for all state variables s^i.*

Assumption 4 *No simultaneous or backward edges in time.*

Theorem 1 *Assuming A1-A4, we define the conditioning set $\{a_t, s_t \setminus s_t^i\} = \{a_t, s_t^1, \ldots s_t^{i-1}, s_t^{i+1}, \ldots\}$. If $s_t^i \not\perp\!\!\!\perp s_{t+1}^j | \{a_t, s_t \setminus s_t^i\}$, then $s_t^i \to s_{t+1}^j$. Similarly, if* $a_t^i \not\perp\!\!\!\perp s_{t+1}^j | \{a_t \setminus a_t^i, s_t\}$, then $a_t^i \to s_{t+1}^j$.

With Assumptions 1-4 and Theorem 1, we can identify the graph structures in $\mathcal{G}$, which can be represented as the adjacency matrix M. Hence, the dynamic transitions and reward functions in MDP with structures are as follows:

$$\begin{cases} s_{t+1}^i = f\left(M^{s \to s} \odot s_t, M^{a \to s} \odot a_t, \epsilon_{s,i,t}\right) \\ \qquad r_t = R(\psi(s_t), a_t) \end{cases} \tag{2}$$

where s_{t+1}^i represents the next state, $M^{s \to s} \in \{0,1\}^{|s| \times |s|}$ and $M^{a \to s} \in \{0,1\}^{|a| \times |s|}$ are the adjacency matrices indicating the influence of current states and actions on the next state, respectively, $\odot$ denotes the element-wise product, and $\epsilon_{s,i,t}$ represents i.i.d. Gaussian noise. The reward r_t is a function of the current state $\psi(s_t)$, which filters out the state without direct edges to the target, and the action a_t.

2.2 Empowerment in RL

Empowerment is to quantify the influence an agent has over its environment and the extent to which this influence can be perceived by the agent [17, 26, 27]. Within our framework, the empowerment is the channel capacity between the agent actions a_t and its subsequent state s_{t+1} given the causal mask M as follows:

$$\mathcal{E} := \max_{p(a_t)} \mathcal{I}(s_{t+1}; a_t \mid M), \tag{3}$$

where $\mathcal{E}$ is used to represent the channel capacity from the action to state observation. $p(a_t)$ is the distribution of actions.

3 Empowerment through Causal Learning

3.1 Overview

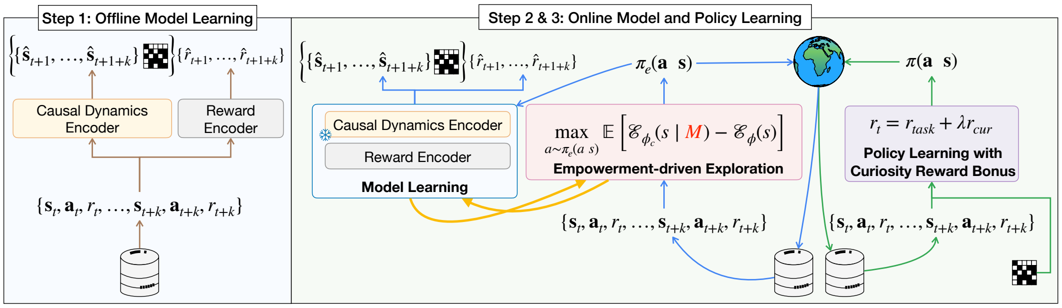

Step 1: Offline Model Learning
$\left\{\{\hat{\mathbf{s}}_{t+1}, \dots, \hat{\mathbf{s}}_{t+1+k}\}\right\}\{\hat{r}_{t+1}, \dots, \hat{r}_{t+1+k}\}$
Causal Dynamics Encoder
Reward Encoder
$\{\mathbf{s}_t, \mathbf{a}_t, r_t, \dots, \mathbf{s}_{t+k}, \mathbf{a}_{t+k}, r_{t+k}\}$
Step 2 & 3: Online Model and Policy Learning
$\left\{\{\hat{\mathbf{s}}_{t+1}, \dots, \hat{\mathbf{s}}_{t+1+k}\}\right\}\{\hat{r}_{t+1}, \dots, \hat{r}_{t+1+k}\}$
$\pi_e(\mathbf{a} \quad \mathbf{s})$
$\pi(\mathbf{a} \quad \mathbf{s})$
Causal Dynamics Encoder
Reward Encoder
Model Learning
$\max_{a \sim \pi_e(a \ s)} \mathbb{E}\left[\mathcal{E}_{\phi_c}(s \mid M) - \mathcal{E}_{\phi}(s)\right]$
Empowerment-driven Exploration
$r_t = r_{task} + \lambda r_{cur}$
Policy Learning with Curiosity Reward Bonus
$\{\mathbf{s}_t, \mathbf{a}_t, r_t, \dots, \mathbf{s}_{t+k}, \mathbf{a}_{t+k}, r_{t+k}\}$
$\{\mathbf{s}_t, \mathbf{a}_t, r_t, \dots, \mathbf{s}_{t+k}, \mathbf{a}_{t+k}, r_{t+k}\}$

Figure 2: Framework of ECL. Gold lines: offline model learning. Blue lines: online model learning alternating with empowerment-driven exploration (yellow lines). Green lines: policy learning.

This framework (Fig. 2) consists of three main steps: offline model learning, online model learning, and policy learning. In offline model learning **(step 1)**, we learn the causal structures of the environment dynamics, capturing the causal dynamics and reward structures. This causal model is trained using offline collected data to identify the causal structures (i.e., causal masks), dynamics and reward models by maximizing the likelihood of observed trajectories. With the learned structured causal model in place, we then integrate empowerment-driven exploration in online model learning **(step 2)**, to learn policies that enhance the agent's ability to control and influence its environment effectively. By alternating the updates of the causal structure and policy to achieve empowerment maximization, the overall optimization objective is to learn the policy that maximizes empowerment with the causal structure. Finally, in **step 3**, the learned causal structure is used as a model to learn policies for down-streaming task policy. In addition to the task reward, to maintain robustness and prevent overfitting in model learning, the curiosity reward is also incorporated.

3.2 Step 1: Offline Model Learning with Causal Structures

We learn the causal and model structures from the offline dataset $\mathcal{D}$. Specifically, we employ a causal dynamic encoder and a reward encoder to maximize the likelihood of observed trajectories.

Causal Dynamics Encoder The causal dynamics encoder consists of two parts: the dynamics encoder P_{ϕ_c} and causal mask M. The dynamics encoder P_{ϕ_c} maximizes the likelihood of observed states:

$$\mathcal{L}_{\mathrm{dyn}} = \mathbb{E}_{(s_t, a_t, s_{t+1}) \sim \mathcal{D}} \left[\sum_{i=1}^{d_S} \log P_{\phi_c}(s_{t+1}^i | s_t, a_t; \phi_c) \right], \tag{4}$$

where d_S is the dimension of the state space, and ϕ_c denotes the parameters of the dynamics encoder.

Causal Discovery For causal discovery, with the learned dynamics model P_{ϕ_c}, we further embed the causal structure into the objective function. To learn the causal structure, we employ two off-the-shelf causal discovery methods: conditional independence testing in [2] and regularization by sparse filters [3]. We also maximize the likelihood of states by updating the dynamics encoder and masks. Thus, the total objective for the causal dynamics encoder is:

$$\mathcal{L}_{\mathrm{c-dyn}} = \mathbb{E}_{(s_t,a_t,s_{t+1})\sim\mathcal{D}}\left[\sum_{i=1}^{d_S}\log P_{\phi_{\mathrm{c}}}(s_{t+1}^{i}|M^{s\to s^j}\odot s_t, M^{a\to s^j}\odot a_t;\phi_{\mathrm{c}}) + \mathcal{L}_{\mathrm{causal}}\right], \tag{5}$$

where $\mathcal{L}_{\mathrm{causal}}$ represents the objective term associated with learning the causal structure[2].

Reward Encoder Similarly, the reward encoder $P_{\varphi_{\mathrm{r}}}$ aims to maximize the likelihood of the rewards:

$$\mathcal{L}_{\mathrm{rew}} = \mathbb{E}_{(s_t,a_t,r_t\sim\mathcal{D})}\left[\log P_{\varphi_r}\left(r_t|\psi(s_t),a_t\right)\right], \tag{6}$$

where $\psi(\cdot)$ is the operation to filter out the irrelevant states with causal dynamics model. Hence, the overall objective of the offline model learning with causal structures is to maximize $\mathcal{L} = \mathcal{L}_{\mathrm{dyn}} + \mathcal{L}_{\mathrm{c-dyn}} + \mathcal{L}_{\mathrm{rew}}$.

3.3 Step 2: Online Model Learning with Empowerment-driven Exploration

In Step 2, we aim to simultaneously optimize the learning of the causal structure and empowerment. Specifically, as illustrated in Fig. 2, we alternately optimize the empowerment-driven exploration policy π_e, the causal mask M, and the reward encoder φ_r. To ensure stable learning, we keep the dynamic encoder ϕ_c learned in Step 1 fixed, focusing solely on the alternating optimization of the causal structure and empowerment.

Empowerment-driven Exploration To enhance the agent's control and efficiency given the causal structure, instead of maximizing $\mathcal{I}\left(s_{t+1},a_t|s_t\right)$ at every environment step, we consider a baseline that uses a dense dynamics model ϕ without causal structures. We then optimize the difference between the empowerment gain of the causal dynamics model and the baseline dense dynamics model.

We first denote the empowerment gain of the causal dynamic model and dynamic model as $\mathcal{E}_{\phi_c}(s) = \max_a \mathcal{I}\left(s_{t+1};a_t \mid s_t;\phi_c,M\right)$ and $\mathcal{E}_{\phi}(s) = \max_a \mathcal{I}\left(s_{t+1};a_t \mid s_t;\phi\right)$, respectively. Here, ϕ corresponds to the dynamic model without considering causal structures. For this purpose, we separately train a well-tuned ϕ on offline data to serve as a baseline for optimization.

Then, we have the following objective function:

$$\max_{a\sim\pi_e(a|s)} \mathbb{E}_{s_t,a_t,s_{t+1}\sim\mathcal{D}}\left[\mathcal{E}_{\phi_c}(s) - \mathcal{E}_{\phi}(s)\right]. \tag{7}$$

In practice, we employ the estimated $\hat{\mathcal{E}}_{\phi_c}(s)$ and $\hat{\mathcal{E}}_{\phi}(s)$, specifically

$$\hat{\mathcal{E}}_{\phi_c}(s) = \max_{a_t\sim\pi_e(a|s)} \mathbb{E}_{\pi_e(a_t|s_t)p_{\phi_c}(s_{t+1}|s_t,a_t)}\left[\log P_{\phi_c}(s_{t+1}\mid s_t,a_t;M,\phi_c) - \log P(s_{t+1}|s)\right], \tag{8}$$

and

$$\hat{\mathcal{E}}_{\phi}(s) = \max_{a_t\sim\pi_e(a|s)} \mathbb{E}_{\pi_e(a_t|s_t)p_{\phi}(s_{t+1}|s_t,a_t)}\left[\log P_{\phi}(s_{t+1}\mid s_t,a_t;\phi) - \log P(s_{t+1}|s)\right], \tag{9}$$

where $P(s_{t+1}|s)$ is the marginal distribution of the future state s_{t+1}. Hence, the objective function Eq. 7 is derived as:

$$\max_{a\sim\pi_e(a|s)} \mathcal{H}(s_{t+1}^{\phi_c}\mid s_t) - \mathcal{H}(s_{t+1}^{\phi}\mid s_t) + \mathbb{E}_{a\sim\pi_e(a|s)}\left[\mathbb{KL}\left(P_{\phi_c}(s_{t+1}\mid s_t,a_t;M)\|P_{\phi}(s_{t+1}\mid s_t,a_t)\right)\right], \tag{10}$$

where $s_{t+1}^{\phi_c}$ and s_{t+1}^{ϕ} denote the state at time $t+1$ under the causal dynamics and dynamics model, respectively. Since Computing $\mathcal{H}(s_{t+1}^{\phi_c}\mid s_t) - \mathcal{H}(s_{t+1}^{\phi}\mid s_t)$ requires integrating over actions. So for simplicity, we update π_e by only optimizing the KL term.

Online Model Learning In Step 2, we fix the dynamics encoder ϕ_c and further fine-tune the causal mask M and the reward encoder φ_r. We adopt an alternating optimization with the policy π_e to learn the model. Specifically, given M, we first optimize π_e. Then, using the actions from π_e, we collect new trajectories online to update M and φ_r.

[2] Detailed loss functions are given in Appendix D.2

3.4 Step 3: Policy Learning with Curiosity Reward

We learn the downstream policy for the task given the causal structures. To mitigate the potential overfitting of the causal model learned in Steps 1&2, we use a curiosity reward (CUR) to serve as an intrinsic motivation objective or exploration bonus, in conjunction with a task-specific reward, to prevent overfitting in model learning.

$$r_{\mathrm{cur}} = \mathbb{E}_{(s_t, a_t, s_{t+1} \sim \mathcal{D})} \left[\mathbb{KL} \left(P_{\mathrm{true}} || P_{\phi_c} \right) - \mathbb{KL} \left(P_{\mathrm{true}} || P_{\phi} \right) \right], \tag{11}$$

where P_{true} is the ground truth dynamics of the system. By taking account of r_{cur}, we encourage the agent to explore states that the causal dynamics cannot capture but the dense dynamics can, thus preventing the policy from being overly conservative due to offline model learning. Hence The shaped reward is shown as follows:

$$r(s, a) = r_{\mathrm{task}}(s, a) + \lambda r_{\mathrm{cur}}(s, a), \tag{12}$$

where $r_{\mathrm{task}}(s, a)$ is the task reward, λ is a balancing hyperparameter.

4 Practical Implementation

We introduce the practical implementation of ECL for casual dynamics learning with empowerment-driven exploration and task learning. The proposed framework for the entire learning process is illustrated in Figure 2, comprising three steps that step 2 and 3 are executed cyclically over time.

Step 1: Offline Model Learning Initially, following [2], we establish a transition collection policy π_{collect} by formulating a reward function that incentivizes selecting transitions that cover more state action pairs to expose causal relationships thoroughly. We train the dynamics model ϕ_c by maximizing the log-likelihood $\mathcal{L}_{\mathrm{dyn}}$, following Eq. 4. Then, we employ causal discovery approach for learning causal mask by maximizing the log-likelihood $\mathcal{L}_{\mathrm{c-dyn}}$ followed Eq. 5. Subsequently, we train the reward predictor φ_{r} by maximizing the likelihood in accordance with Eq. 6.

Step 2: Online Model Learning We execute empowerment-driven exploration by maximizing $\mathcal{I}(s_{t+1}; a_t | M) - \mathcal{I}(s_{t+1}; a_t)$ followed Eq. 7 with causal dynamics model and dynamics model without causal mask for policy π_e learning. Furthermore, the learned policy π_e is used to sample transition for casual mask M fine-tuning with fixed ϕ. We alternately perform empowerment-driven exploration for policy learning and causal model learning for causal structure optimization.

Step 3: Policy Learning During downstream task learning, we incorporate the causal effects of different actions as curiosity rewards combined with the task reward, following Eq. 12. The causality introduced by CUR in task learning maintains essential exploration, thereby facilitating the learning of an optimal policy to maintain robustness and prevent overfitting in model learning. We maximize the discounted cumulative reward $\eta_{\hat{\mathcal{M}}}(\pi_\theta)$ to learn the policy by the cross entropy entropy (CEM) [28].

5 Experiments

We aim to answer the following questions in the evaluation: (i) How does the performance of ECL compared to other causal and dense causal models across different environments for tasks and dynamics learning? (ii) Whether different causal discovery methods in step 1 and 2, impact policy performance? (iii) Does ECL improve the causal discovery and learning efficiency with the empowerment gain? (iv) What are the effects of the hyperparameters in ECL?

5.1 Setup

Environments. We select 3 different environments for experimental evaluation. **Chemical [29]:** The task is to discover the causal relationship (Chain, Collider & Full) of chemical items which proves the learned dynamics and explains the behavior without spurious correlations. **Manipulation [2]:** The task is to prove the learned dynamics and policy for difficult settings with spurious correlations and multi-dimension action causal influence. **Physical [29]:** We also evaluate our method in the dense mode environment Physical. For the details of environment setup, please refer to Appendix D.2.

Baselines. We compare ECL with 3 causal and 2 dense dynamics methods. CDL [2]: infers causal relationships between the variables for dynamics learning with Conditional Independence Test (CIT). ASR [3]: causal structure learning based on regularization, where the causal mask is learned as trainable parameters. GRADER [8]: generalizing goal-conditioned RL with CIT by variational causal reasoning. GNN [29]: a graph neural network with dense dependence for each state variable.

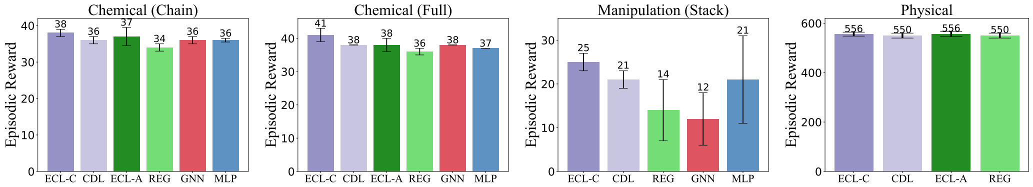


Figure 3: The task learning of episodic reward in three environments.

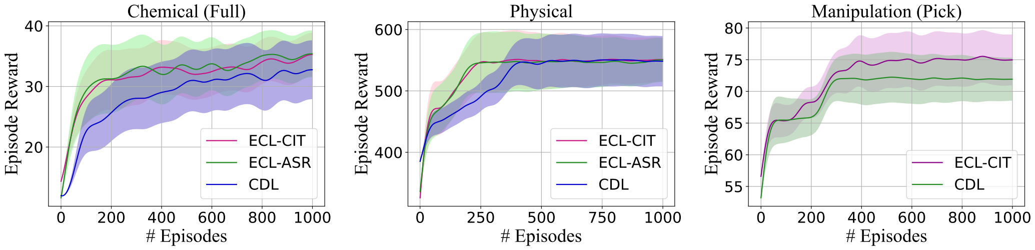


Figure 4: The learning curves of episodic reward in three different environments and the shadow is the standard error.

Monolithic [2]: a multi-layer perceptron (MLP) network that takes all state variables and actions for prediction. For ECL, we employ both the conditional independence testing (same setup in CDL [2]) and sparse regularization (same setup in ASR [3]) as the causal discovery modules.

Evaluation Metric. In tasks learning, we utilize episodic reward and the task success as evaluation criteria for downstream tasks. For causal dynamics learning, we employ five metrics to evaluate the learned causal graph and assess the mean accuracy for dynamics predictions of future states both In-Distribution (ID) and Out-Of-Distribution (OOD).

5.2 Results

5.2.1 Task Learning

We evaluate each method with the following downstream tasks in the chemical (C), physical (P) and the manipulation (M) environments. **Match** (C): match the object colors with goal colors individually. **Push** (P): use the heavier object to push the lighter object to goal position. **Reach** (M): move the end-effector to the goal position. **Pick** (M): pick the movable object to the goal position. **Stack** (M): stack the movable object on the top of the unmovable object.

As shown in Fig. 3, compared to dense models GNN and MLP, as well as the causal approaches CDL and REG, ECL-CIT attains the highest reward across 3 environments. Notably, ECL-CIT outperforms other methods in the intricate manipulation tasks. Furthermore, ECL-ASR surpasses REG, elevating model performance and achieving a reward comparable to CDL. The proposed curiosity reward encourages exploration and avoids local optimality during the policy learning process. Additionally, Figure 4 depicts the learning curves across three environments. Across these diverse settings, ECL exhibits elevated sample efficiency compared to CDL and higher reward attainment.

Sample Efficiency Analysis. After validating the effectiveness of ECL in reward learning, we further substantiate the improvements in sample efficiency of ECL during task learning. As depicted in Figure 5, we illustrate task success in both collider and manipulation reach tasks. The compared experimental results underscore the efficiency of ECL, demonstrating enhanced sample efficiency across different environments.

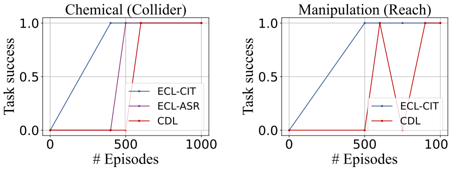


Figure 5: Task success in the collider and manipulation environments.

5.2.2 Causal Dynamics learning

Causal Graph Learning. To evaluate the efficacy of ECL for learning causal relationships, we first conduct experimental analyses across three chemical environments, employing five evaluation metrics. We conduct causal learning based on the causal discovery with CIT and ASR respectively. The comparative results using the same causal discovery methods are presented in Table 1, with

Table 1: Compared results of causal graph learning on three chemical environments.

Metrics	Methods	Chain	Collider	Full
Accuracy	ECL/CDL	1.00±0.00/1.00±0.00	1.00±0.00/1.00±0.00	**1.00±0.00**/0.99±0.00
	ECL/REG	0.99±0.00/0.99±0.00	0.99±0.00/0.99±0.00	**0.99±0.01**/0.98±0.00
Recall	ECL/CDL	**1.00±0.00**/0.99±0.01	1.00±0.00/1.00±0.00	**0.97±0.01**/0.92±0.02
	ECL/REG	**1.00±0.00**/0.94±0.01	**0.99±0.01**/0.89±0.09	**0.90±0.02**/0.79±0.01
Precision	ECL/CDL	1.00±0.00/1.00±0.00	1.00±0.00/1.00±0.00	0.96±0.02/ **0.97±0.02**
	ECL/REG	0.99±0.01/0.99±0.01	0.99±0.01/0.99±0.01	**0.97±0.03**/0.92±0.05
F1 Score	ECL/CDL	**1.00±0.00**/0.99±0.01	1.00±0.00/1.00±0.00	**0.97±0.01**/0.94±0.01
	ECL/REG	**0.99±0.00**/0.96±0.01	**0.99±0.00**/0.94±0.05	**0.93±0.02**/0.85±0.02
ROC AUC	ECL/CDL	**1.00±0.00**/0.99±0.01	1.00±0.00/1.00±0.00	**0.98±0.01**/0.96±0.01
	ECL/REG	0.99±0.01/0.99±0.01	**0.99±0.01**/0.93±0.04	0.95±0.01/0.95±0.01

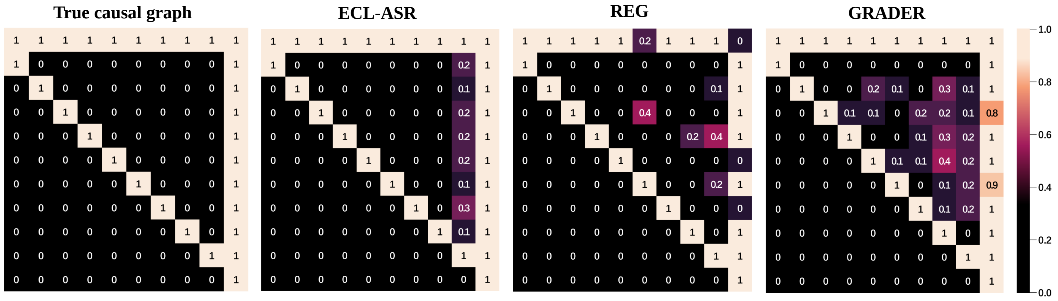


Figure 6: The compared causal gragh in the chemical collider environment.

each cell containing the comparative results for that method across different scenarios. These results demonstrate the superior performance of our approach in causal inference, exhibiting both efficiency and robustness as evinced by the evaluation metrics of F1 score and ROC AUC. All results exceed 0.90. Notably, our approach exhibits exceptional learning capabilities in chain and collider environments.

Visualization. Moreover, we visually compare the inferred causal graph with the ground truth graph in terms of edge accuracy. The results depicted in Figure 6 illustrate the causal graphs of ECL-ASR compared to REG and GRADER in the collider environment. For nodes exhibiting strong causality, ECL-ASR achieves fully accurate learning and substantial accuracy enhancements compared to REG. Concurrently, ECL-ASR elucidates the causality between action and state more effectively. Furthermore, ECL-ASR mitigates interference from irrelevant causal nodes more proficiently than GRADER. These findings substantiate that ECL attains superior performance compared to other causal discovery methods in causal learning. For full experimental results, please refer to Appendix D.

6 Related Work

Empowerment is an intrinsic motivation to improve the controllability over the environment [17, 26]. This concept is from the information-theoretic framework, wherein actions and future states are viewed as channels for information transmission. In RL, empowerment is applied to uncover more controllable associations between states and actions or skills [30, 19, 31, 20]. By quantifying the influence of different behaviors or skills on state transitions, empowerment encourages the agent to explore further to enhance its controllability over the system [16, 32]. Maximizing empowerment $\max_{\pi} I$ can be used as the learning objective functions, empowering agents to demonstrate intelligent behavior without requiring predefined external goals and model reconstruction.

7 Conclusion

This study propose an method-agnostic framework of empowerment through causal structure learning in MBRL to improve controllability and learning efficiency within environments. We maximize empowerment under causal structure to prioritize controllable information and optimize causal world models to guide downstream task learning. Further, we propose an intrinsic-motivated curiosity reward during task learning to prevent overfitting. Extensive experiments across 3 environments substantiate the remarkable performance. For our future work, we will concentrate on extending this framework to disentangle directable behaviors.

References

[1] Thomas M Moerland, Joost Broekens, Aske Plaat, Catholijn M Jonker, et al. Model-based reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 16(1):1–118, 2023.

[2] Zizhao Wang, Xuesu Xiao, Zifan Xu, Yuke Zhu, and Peter Stone. Causal dynamics learning for task-independent state abstraction. *arXiv preprint arXiv:2206.13452*, 2022.

[3] Biwei Huang, Chaochao Lu, Liu Leqi, José Miguel Hernández-Lobato, Clark Glymour, Bernhard Schölkopf, and Kun Zhang. Action-sufficient state representation learning for control with structural constraints. In *International Conference on Machine Learning*, pages 9260–9279. PMLR, 2022.

[4] Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. *arXiv preprint arXiv:2006.10742*, 2020.

[5] Biwei Huang, Fan Feng, Chaochao Lu, Sara Magliacane, and Kun Zhang. Adarl: What, where, and how to adapt in transfer reinforcement learning. *arXiv preprint arXiv:2107.02729*, 2021.

[6] Jonathan Richens and Tom Everitt. Robust agents learn causal world models. *arXiv preprint arXiv:2402.10877*, 2024.

[7] Chaochao Lu, Yuhuai Wu, José Miguel Hernández-Lobato, and Bernhard Schölkopf. Invariant causal representation learning for out-of-distribution generalization. In *International Conference on Learning Representations*, 2021.

[8] Wenhao Ding, Haohong Lin, Bo Li, and Ding Zhao. Generalizing goal-conditioned reinforcement learning with variational causal reasoning. *Advances in Neural Information Processing Systems*, 35:26532–26548, 2022.

[9] Wenhao Ding, Laixi Shi, Yuejie Chi, and Ding Zhao. Seeing is not believing: Robust reinforcement learning against spurious correlation. *Advances in Neural Information Processing Systems*, 36, 2024.

[10] Yuren Liu, Biwei Huang, Zhengmao Zhu, Honglong Tian, Mingming Gong, Yang Yu, and Kun Zhang. Learning world models with identifiable factorization. *Advances in Neural Information Processing Systems*, 36, 2024.

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

[12] Mirco Mutti, Riccardo De Santi, Marcello Restelli, Alexander Marx, and Giorgia Ramponi. Exploiting causal graph priors with posterior sampling for reinforcement learning. *arXiv preprint arXiv:2310.07518*, 2023.

[13] Silviu Pitis, Elliot Creager, and Animesh Garg. Counterfactual data augmentation using locally factored dynamics. *Advances in Neural Information Processing Systems*, 33:3976–3990, 2020.

[14] Silviu Pitis, Elliot Creager, Ajay Mandlekar, and Animesh Garg. Mocoda: Model-based counterfactual data augmentation. *Advances in Neural Information Processing Systems*, 35: 18143–18156, 2022.

[15] Fan Feng and Sara Magliacane. Learning dynamic attribute-factored world models for efficient multi-object reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2023.

[16] Felix Leibfried, Sergio Pascual-Diaz, and Jordi Grau-Moya. A unified bellman optimality principle combining reward maximization and empowerment. *Advances in Neural Information Processing Systems*, 32, 2019.

[17] Alexander S Klyubin, Daniel Polani, and Chrystopher L Nehaniv. Empowerment: A universal agent-centric measure of control. In *2005 ieee congress on evolutionary computation*, volume 1, pages 128–135. IEEE, 2005.

[18] Alexander S Klyubin, Daniel Polani, and Chrystopher L Nehaniv. Keep your options open: An information-based driving principle for sensorimotor systems. *PloS one*, 3(12):e4018, 2008.

[19] Homanga Bharadhwaj, Mohammad Babaeizadeh, Dumitru Erhan, and Sergey Levine. Information prioritization through empowerment in visual model-based rl. *arXiv preprint arXiv:2204.08585*, 2022.

[20] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.

[21] Judea Pearl. *Causality*. Cambridge university press, 2009.

[22] Kevin P Murphy et al. Dynamic bayesian networks. *Probabilistic Graphical Models, M. Jordan*, 7:431, 2002.

[23] Carlos Guestrin, Daphne Koller, Ronald Parr, and Shobha Venkataraman. Efficient solution algorithms for factored mdps. *Journal of Artificial Intelligence Research*, 19:399–468, 2003.

[24] Carlos Guestrin, Daphne Koller, and Ronald Parr. Multiagent planning with factored mdps. *Advances in neural information processing systems*, 14, 2001.

[25] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2001.

[26] Christoph Salge, Cornelius Glackin, and Daniel Polani. Empowerment–an introduction. *Guided Self-Organization: Inception*, pages 67–114, 2014.

[27] Tobias Jung, Daniel Polani, and Peter Stone. Empowerment for continuous agent—environment systems. *Adaptive Behavior*, 19(1):16–39, 2011.

[28] Reuven Y Rubinstein. Optimization of computer simulation models with rare events. *European Journal of Operational Research*, 99(1):89–112, 1997.

[29] Nan Rosemary Ke, Aniket Didolkar, Sarthak Mittal, Anirudh Goyal, Guillaume Lajoie, Stefan Bauer, Danilo Rezende, Yoshua Bengio, Michael Mozer, and Christopher Pal. Systematic evaluation of causal discovery in visual model based reinforcement learning. *arXiv preprint arXiv:2107.00848*, 2021.

[30] Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. *Advances in neural information processing systems*, 28, 2015.

[31] Jongwook Choi, Archit Sharma, Honglak Lee, Sergey Levine, and Shixiang Shane Gu. Variational empowerment as representation learning for goal-based reinforcement learning. *arXiv preprint arXiv:2106.01404*, 2021.

[32] Maximilian Seitzer, Bernhard Schölkopf, and Georg Martius. Causal influence detection for improving efficiency in reinforcement learning. *Advances in Neural Information Processing Systems*, 34:22905–22918, 2021.

[33] Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. *Advances in neural information processing systems*, 32, 2019.

[34] Tung D Nguyen, Rui Shu, Tuan Pham, Hung Bui, and Stefano Ermon. Temporal predictive coding for model-based planning in latent space. In *International Conference on Machine Learning*, pages 8130–8139. PMLR, 2021.

[35] Mingde Zhao, Zhen Liu, Sitao Luan, Shuyuan Zhang, Doina Precup, and Yoshua Bengio. A consciousness-inspired planning agent for model-based reinforcement learning. *Advances in neural information processing systems*, 34:1569–1581, 2021.

[36] Masashi Okada and Tadahiro Taniguchi. Dreaming: Model-based reinforcement learning by latent imagination without reconstruction. In *2021 ieee international conference on robotics and automation (icra)*, pages 4209–4215. IEEE, 2021.

[37] Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33:14129–14142, 2020.

[38] Zhihai Wang, Jie Wang, Qi Zhou, Bin Li, and Houqiang Li. Sample-efficient reinforcement learning via conservative model-based actor-critic. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8612–8620, 2022.

[39] Brandon Amos, Samuel Stanton, Denis Yarats, and Andrew Gordon Wilson. On the model-based stochastic value gradient for continuous reinforcement learning. In *Learning for Dynamics and Control*, pages 6–20. PMLR, 2021.

[40] Inwoo Hwang, Yunhyeok Kwak, Suhyung Choi, Byoung-Tak Zhang, and Sanghack Lee. Quantized local independence discovery for fine-grained causal dynamics learning in reinforcement learning. 2023.

[41] Zizhao Wang, Caroline Wang, Xuesu Xiao, Yuke Zhu, and Peter Stone. Building minimal and reusable causal state abstractions for reinforcement learning. *arXiv preprint arXiv:2401.12497*, 2024.

[42] Zizhao Wang, Xuesu Xiao, Yuke Zhu, and Peter Stone. Task-independent causal state abstraction. In *Proceedings of the 35th International Conference on Neural Information Processing Systems, Robot Learning workshop*, 2021.

A Broader Impact

Our work explores leveraging causal structure to enhance empowerment for efficient policy learning, enabling better control of the environment in model-based reinforcement learning (MBRL). We propose a framework that can effectively combine diverse causal discovery methods. This holistic approach not only refines policy learning but also ensures that the causal model remains adaptable and accurate, even when faced with novel or shifting environmental conditions. ECL demonstrates improved learning efficiency and generalization compared to other causal MBRL methods across three different RL environments. Simultaneously, ECL achieves more accurate causal relationship discovery, overcoming spurious associations present in the environment.

While ECL demonstrated strengths in accurate causal discovery and overcoming spurious associations, disentangling controllable behavioral dimensions remains a limitation. Our implicit empowerment approach enhances the policy's control over the environment, but does not explicitly tease apart different behavioral axes. Explicitly disentangling controllable behavioral dimensions could be an important future work to further improve behavioral control and empowerment. Additionally, our current approach involves substantial data collection and model optimization efforts, which can hinder training efficiency. Moving forward, we aim to further streamline our framework to enable more efficient policy training and causal structure learning. Enhancing computational performance while maintaining accuracy will be a key focus area for future iterations of this work.

B Additional Related Works

B.1 Model-based Reinforcement Learning

MBRL involves training a dynamics model by maximizing the likelihood of collected transitions, known as the world model, as well as learning a reward model [1, 33]. Based on learned models, MBRL can execute downstream task planning [34, 35], data augmentation [14, 36, 37], and Q-value estimation [38, 39]. MBRL can easily leverage prior knowledge of dynamics, making it more effective at enhancing policy stability and generalization. However, when faced with high-dimensional state spaces and confounders in complex environments, the dense models learned by MBRL suffer from spurious correlations and poor generalization [2, 19]. To tackle these issues, causal inference approaches are applied to MBRL for state abstraction, removing unrelated components [40, 8, 41].

B.2 Causality in MBRL

Due to the exclusion of irrelevant factors from the environment through causality, the application of causal inference in MBRL can effectively improve sample efficiency and generalization [29, 11]. Wang [42] proposes a regularization-based causal dynamics learning method that explicitly learns causal dependencies by regularizing the number of variables used when predicting each state variable. GRADER [8] execute variational inference by regarding the causal graph as a latent variable. CDL [2] is a causal dynamics learning method based on conditional independence testing. CDL employs conditional mutual information to compute the causal relationships between different dimensions of states and actions, thereby explicitly removing unrelated components. However, it is challenging to strike a balance between explicit causal discovery and prediction performance, and the learned policy has lower controllability over the system.

C Theoretical Analyses

Assumption A.1. (d-separation [21]) d-separation is a graphical criterion used to determine, from a given causal graph, if a set of variables X is conditionally independent of another set Y, given a third set of variables Z. In a directed acyclic graph (DAG) $\mathcal{G}$, a path between nodes n_1 and n_m is said to be blocked by a set S if there exists a node n_k, for $k = 2, \cdots, m-1$, that satisfies one of the following two conditions:

(i) $n_k \in S$, and the path between n_{k-1} and n_{k+1} forms $(n_{k-1} \rightarrow n_k \rightarrow n_{k+1})$, $(n_{k-1} \leftarrow n_k \leftarrow n_{k+1})$, or $(n_{k-1} \leftarrow n_k \rightarrow n_{k+1})$.

(ii) Neither n_k nor any of its descendants is in S, and the path between n_{k-1} and n_{k+1} forms $(n_{k-1} \rightarrow n_k \leftarrow n_{k+1})$.

In a DAG, we say that two nodes n_a and n_b are d-separated by a third node n_c if every path between nodes n_a and n_b is blocked by n_c, denoted as $n_a \perp\!\!\!\perp n_b | n_c$.

Proposition A.2. *Under the assumptions that the causal graph is Markov and faithful to the observations, there exists an edge from* $a_t^i \rightarrow s_{t+1}^j$ *if and only if* $a_t^i \not\perp\!\!\!\perp s_{t+1}^j | \{a_t \setminus a_t^i, s_t\}$, *then* $a_t^i \rightarrow s_{t+1}^j$.

Proof. We first prove that if there exists an edge from a_t^i to s_{t+1}^j, then $a_t^i \not\perp\!\!\!\perp s_{t+1}^j | \{a_t \setminus a_t^i, s_t\}$. We prove it by contradiction. Suppose that a_t^i is independent of s_{t+1}^j given $\{a_t \setminus a_t^i, s_t\}$. According to the faithfulness assumption, we can infer this independence from the graph structure. If a_t^i is independent of s_{t+1}^j given $\{a_t \setminus a_t^i, s_t\}$, then there cannot be a directed path from a_t^i to s_{t+1}^j in the graph. Hence, there is no edge between a_t^i and s_{t+1}^j. This contradicts our initial statement about the existence of this edge.

Now, we prove the converse: if $a_t^i \not\perp\!\!\!\perp s_{t+1}^j | \{a_t \setminus a_t^i, s_t\}$, then there exists an edge from a_t^i to s_{t+1}^j. Again, we use proof by contradiction. Suppose there is no edge between a_t^i and s_{t+1}^j in the graph. Due to the Markov assumption, the lack of an edge between these variables implies their conditional independence given $\{a_t \setminus a_t^i, s_t\}$. This contradicts our initial statement that $a_t^i \not\perp\!\!\!\perp s_{t+1}^j | \{a_t \setminus a_t^i, s_t\}$. Therefore, there must exist an edge from a_t^i to s_{t+1}^j.

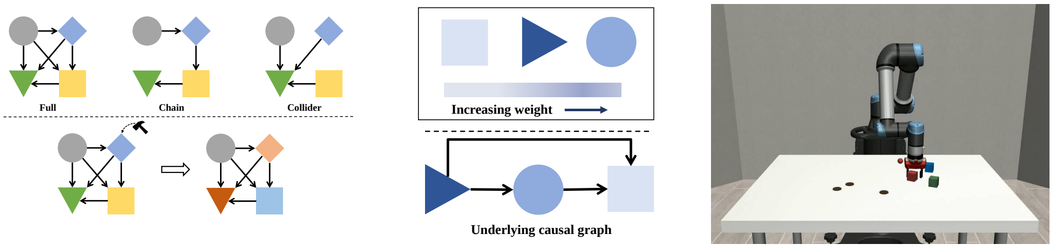
Proposition A.3. *Under the assumptions that the causal graph is Markov and faithful to the observations, there exists an edge from* $s_t^i \rightarrow s_{t+1}^j$ *if and only if* $s_t^i \not\perp\!\!\!\perp s_{t+1}^j | \{a_t, s_t \setminus s_t^i\}$.

The proof of Proposition A.3 follows a similar line of reasoning to that of Proposition A.2. Consequently, the two propositions collectively serve as the foundation for deriving Theorem 1.

D Details on Experimental Design and Results

D.1 Experimental environments

We select three different types environments for experimental evaluation, as shown in Figure 7.



(a) Chemical (b) Physical (c) Manipulation

Figure 7: Three experimental environments.

Chemical In chemical environment, we aim to discovery the causal relationship (Chain, Collider & Full) of chemical items which will prove the learned dynamics and explain the behavior without spurious correlations. Meanwhile, in the downstream tasks, we evaluate the proposed methods by episodic reward and success rate. The reward function is defined as follows:

Match: match the object colors with goal colors individually:

$$r^{\text{match}} = \sum_{i=1}^{10} \mathbb{1}[m_t^i = g^i] \tag{13}$$

where $\mathbb{1}$ is the indicator function, m_t^i is the current color of the i-object, and g^i is the goal color of the i-object.

Manipulation In manipulation environment, we aim to prove the learned dynamics and policy for difficult settings with spurious correlations and multi-dimension action causal influence. The state space consists of the robot end-effector (EEF) location ($\mathbb{R}^3$), gripper (grp) joint angles ($\mathbb{R}^2$), and locations of objects and markers ($6 \times \mathbb{R}^3$). The action space includes EEF location displacement ($\mathbb{R}^3$) and the degree to which the gripper is opened ([0, 1]). In each episode, the objects and markers are reset to randomly sampled poses on the table. The task reward functions of **Reach**, **Pick** and **Stack** are followed [2].

Physical In addition to the chemical and manipulation environment, we also evaluate our method in the physical environment. In a 5 × 5 grid-world, there are 5 objects and each of them has a unique weight. The state space is 10-dimensional, consisting of x, y positions (a categorical variable over 5 possible values) of all objects. At each step, the action selects one object, moves it in one of 4 directions or lets it stay at the same position (a categorical variable over 25 possible actions). During the movement, only the heavier object can push the lighter object (the object won't move if it tries to push an object heavier than itself). Meanwhile, the object cannot move out of the grid-world nor can it push other lighter objects out of the grid-world. Moreover, the object cannot push two objects together, even when both of them are lighter than itself (Dense model mode). The task reward function is defined as follows:

Push: calculate the average distance between the current node and the target location:

$$r^{\text{match}} = \frac{1}{5}\sum_{i=1}^{5} \text{dis}(o_i, t_i) \tag{14}$$

where $\text{dis}(\cdot)$ is the distance between two objects position. o_i is the position of current node and t_i is the position of target node.

D.2 Experimental setup

D.2.1 Dynamics learning implementation details

We present the architectures of the proposed method across all environments in Table 2. For all activation functions, the Rectified Linear Unit (ReLU) is employed. Additionally, we summarize the hyperparameters for causal mask learning used in all environments for ECL-CIT and ECL-ASR in Table 3. Regarding the other parameter settings, we adhered to the parameter configurations established in CDL [2] and REG [42].

The $\mathcal{L}_{\text{causal}}$ of CIT-based causal discovery method used in ECL is :

$$\mathcal{L}_{\text{causal}}^{\text{CIT}} = \sum_{j=1}^{d_S} \left[\log \hat{p}(s_{t+1}^j | \{a_t, s_t \setminus s_t^i\}) + \log\hat{p}(s_{t+1}^j | \text{PA}_{s^j})\right] \tag{15}$$

The $\mathcal{L}_{\text{causal}}$ of regularization-based causal discovery method used in ECL is :

$$\mathcal{L}_{\text{causal}}^{\text{ASR}} = \mathop{\mathbb{E}}_{\mathcal{D}} \log P(s_{t+1;t+H} | s_t, a_{t:t+H-1} - \lambda_M ||M||_1) \tag{16}$$

where $\mathcal{D}$ is the transition data and λ_M is regularization coefficient.

D.2.2 Task learning implementation details

We list the downstream task learning architectures of the proposed method across all environments in Table 4. We outline the parameter configurations for the reward predictor, as well as the settings employed for the Cross-Entropy Method (CEM) that is applied.

D.3 Results of causal dynamics learning

We compare the performance of causal dynamics learning with CIT-based method GRADER [8], CDL [2] and regularization-based method REG [42] across different environments. The experimental results, presented in Table 5, reveal that although GRADER exhibits superior performance in the chemical full environment, ECL-based methods overall achieve better results than GRADER across

Table 2: Architecture settings in all environments.

Architecture	Environments		
	Chemical	Physical	Manipulation
feature dimension	64	128	128
predictive networks	[64,32]	[128,128]	[128,64]
number of transitions	500K	500K	32M
max step of environment	50	100	250
batch size	64		
learing rate	1e-4		
max sample time	128		
prediction step during training	2		

Table 3: Hyperparameters for causal mask learning in all environments.

Method	hyperparameters	Environments		
		Chemical	Physical	Manipulation
ECL-CIT	CMI threshold	0.02	0.01	0.002
	optimization frequency	10		
	evaluation frequency	10		
	evaluation batch size	32		
	evaluation step	1		
	prediction reward weight	1.0		
ECL-ASR	coefficient	0.002	0.02	0.001
	regularization starts after N steps	100K	100K	750K

three chemical environments. In the accuracy assessment metrics, ECL-CIT attains 100% precision, and across the chain and collider environments, all evaluation metrics achieve perfect 100% scores. Furthermore, in the physical environment, our proposed methods attain 100% performance. The result of rigorous evaluation metrics substantiate that incorporating ECL has boosted the dynamics model performance. These experimental results further validate the effectiveness of the proposed ECL approach in both sparse and dense modal environments.

Furthermore, we analyze the prediction accuracy performance of the causal dynamics constructed by our proposed method. The multi-step (1-5 steps) prediction experimental results across four environments are illustrated in Figure 8. ECL-CIT and CDL exhibit smaller declines in accuracy as the prediction steps increase, benefiting from the causal discovery realized based on conditional mutual information. Compared to REG, ECL-ASR achieves a significant improvement in accuracy under different settings. Concurrently, we find that the outstanding out-of-distribution experimental results further corroborate the strong generalization capability of our proposed method. Overall, we can demonstrate that the proposed ECL framework realizes efficient and robust causal dynamics learning.

D.4 Visualization on the learned causal graphs

We conduct a detailed comparative analysis by visualizing the learned causal graphs. In each causal graph, these are $d_{\mathcal{S}}$ rows and $d_{\mathcal{S}} + 1$ columns, and the element at the j-th row and i-th column represents whether the variable s_{t+1}^{j} depends on the variable s_{t+1}^{i} if $j < d_{\mathcal{S}} + 1$ or a_t if $j = d_{\mathcal{S}} + 1$, measured by CMI for CIT-based methods and Bernoulli success probability for Reg. First, the causal graph learning scenario in the chemical chain environment is shown in Figure 9. Compared to CDL and REG, ECL-CIT accurately uncovers the causal relationships among crucial elements, such as all different dimensions between states and actions, outperforming the other two methods. Moreover, we

Table 4: Hyperparameters for downstream task learning in all environments.

Method	hyperparameters	Environments		
		Chemical	Physical	Manipulation
Reward Predictor	training step	300K	1.5M	2M
	optimizer	Adam		
	learing rate	3e-4		
	batch size	32		
CEM	number of candidates	64	128	
	number of iterations	5	10	
	number of top candidates	32		
	action_noise	0.03		

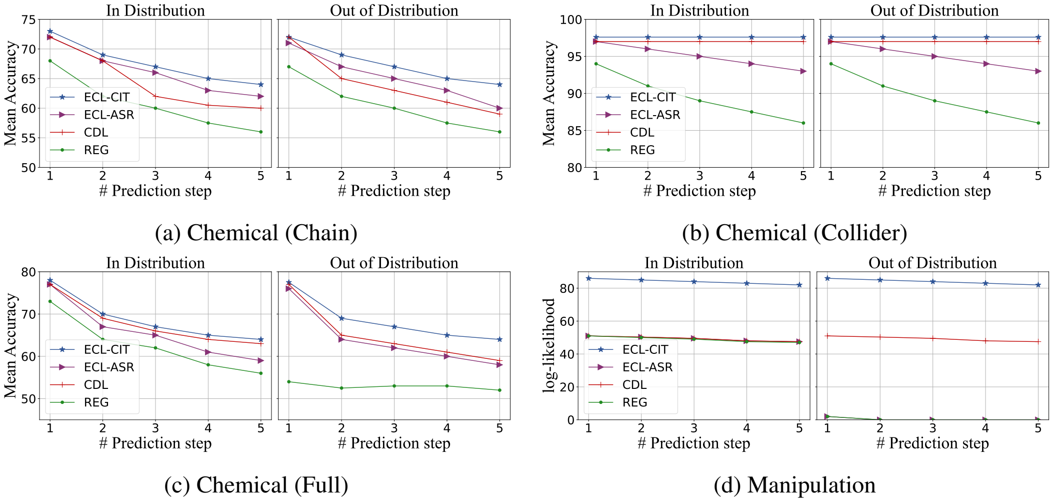
Table 5: Compared results of causal graph learning on three chemical and physical environments.

Metrics	Methods	Chain	Collider	Full	Physical
Accuracy	ECL-CIT	**1.00±0.00**	**1.00±0.00**	**1.00±0.00**	**1.00±0.00**
	ECL-ASR	0.99±0.00	0.99±0.00	0.99±0.01	**1.00±0.00**
	GRADER	0.99±0.00	0.99±0.00	0.99±0.00	-
Recall	ECL-CIT	**1.00±0.00**	**1.00±0.00**	**0.97±0.00**	**1.00±0.00**
	ECL-ASR	**1.00±0.00**	0.99±0.01	0.90±0.02	**1.00±0.00**
	GRADER	0.96±0.03	0.99±0.02	0.96±0.02	-
Precision	ECL-CIT	**1.00±0.00**	**1.00±0.00**	0.96±0.02	**1.00±0.00**
	ECL-ASR	0.99±0.01	0.99±0.01	0.97±0.03	**1.00±0.00**
	GRADER	0.94±0.04	0.90±0.05	**1.00±0.00**	-
F1 Score	ECL-CIT	**1.00±0.00**	**1.00±0.00**	0.97±0.01	**1.00±0.00**
	ECL-ASR	0.99±0.00	0.99±0.00	0.93±0.02	**1.00±0.00**
	GRADER	0.95±0.03	0.94±0.03	**0.98±0.01**	-
ROC AUC	ECL-CIT	**1.00±0.00**	**1.00±0.00**	**0.98±0.01**	**1.00±0.00**
	ECL-ASR	0.99±0.01	0.99±0.01	0.95±0.01	**1.00±0.00**
	GRADER	0.94±0.02	0.99±0.01	0.96±0.01	-

achieve extensive elimination of causality between irrelevant factors. These results demonstrate the accuracy of the proposed method in causal inference within the chemical chain environment.

Furthermore, for the chemical collider environment, the compared causal graphs are depicted in Figure 10. We can observe that both CDL and ECL-CIT achieved optimal discovery of causal relationships. Moreover, in contrast to the REG method, ECL-CIT is not impeded by interference from irrelevant causal factors. For the chemical full environment, the causal graph is illustrated in Figure 11. Compared to CDL, ECL-CIT better excludes interference from irrelevant causal factors. In comparison with the REG method, ECL-CIT attains superior overall performance in discovering causal relationships. Additionally, ECL-CIT reaches optimal learning performance when provided the true causal graph.

Moreover, for the manipulation environment, the experimental results are presented in Figures 12 and 13. From the results in Figure 6, we can discern that ECL-CIT achieves around 90% overall fitting degree with the true causal graph and accurately learns the causal association between state and action. Compared to CDL shown in Figure 13, ECL-CIT learns more causal associations from relevant causal components related to the gripper, movable states, and actions. Conversely, in contrast to REG, ECL-CIT better excludes interference from irrelevant causal factors, such as unmovable and marker states. In summary, the proposed method achieves more accurate and efficient learning performance in causal dynamics learning.



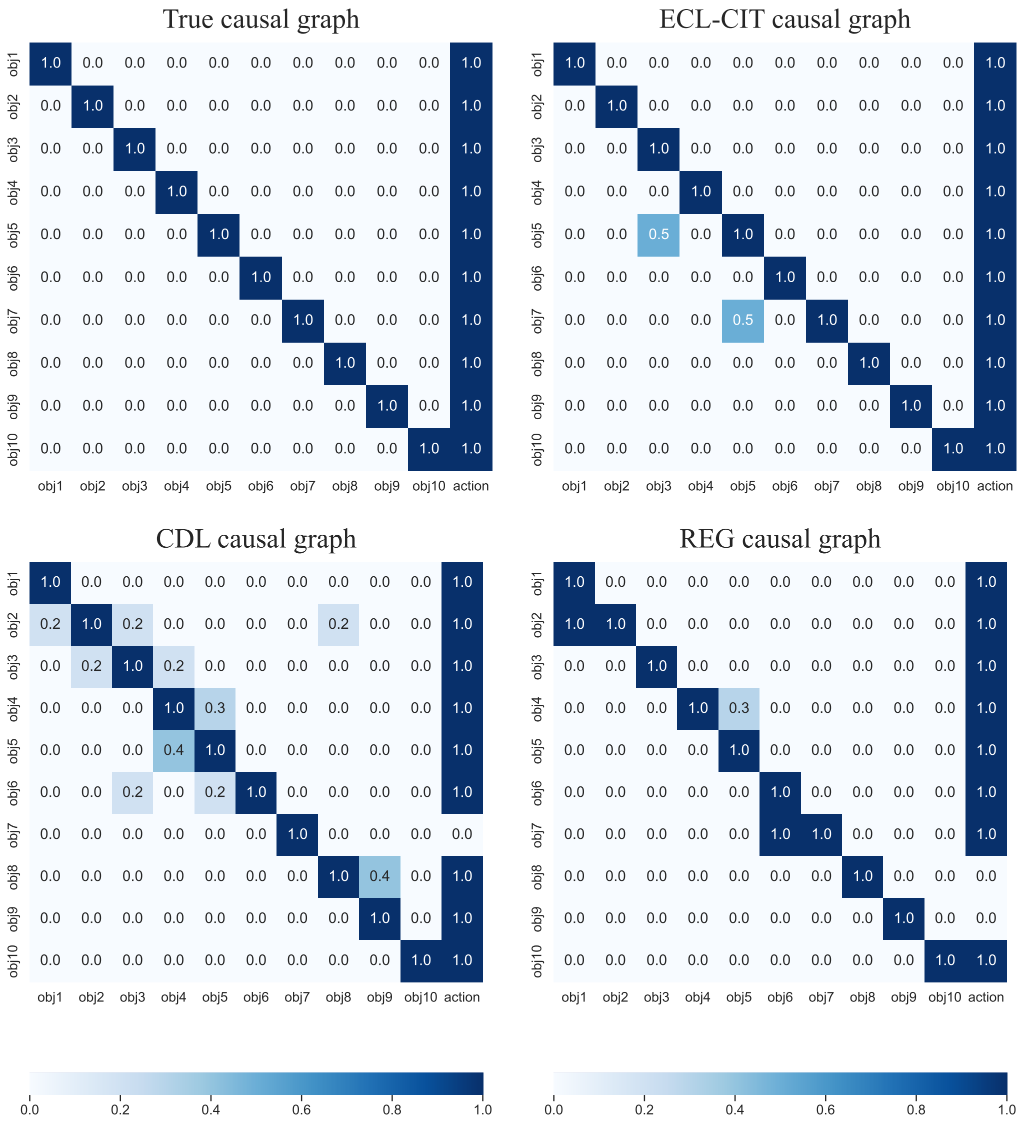
(a) Chemical (Chain)

(b) Chemical (Collider)

(c) Chemical (Full)

(d) Manipulation

Figure 8: Multi-step prediction performance for all environments. (**Left**) prediction on in distribution states. (**Right**) prediction on out of distribution states.



True causal graph

	obj1	obj2	obj3	obj4	obj5	obj6	obj7	obj8	obj9	obj10	action
obj1	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
obj2	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
obj3	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
obj4	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
obj5	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0
obj6	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0
obj7	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0
obj8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0
obj9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0
obj10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0

ECL-CIT causal graph

	obj1	obj2	obj3	obj4	obj5	obj6	obj7	obj8	obj9	obj10	action
obj1	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
obj2	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
obj3	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
obj4	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
obj5	0.0	0.0	0.5	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0
obj6	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0
obj7	0.0	0.0	0.0	0.0	0.5	0.0	1.0	0.0	0.0	0.0	1.0
obj8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0
obj9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0
obj10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0

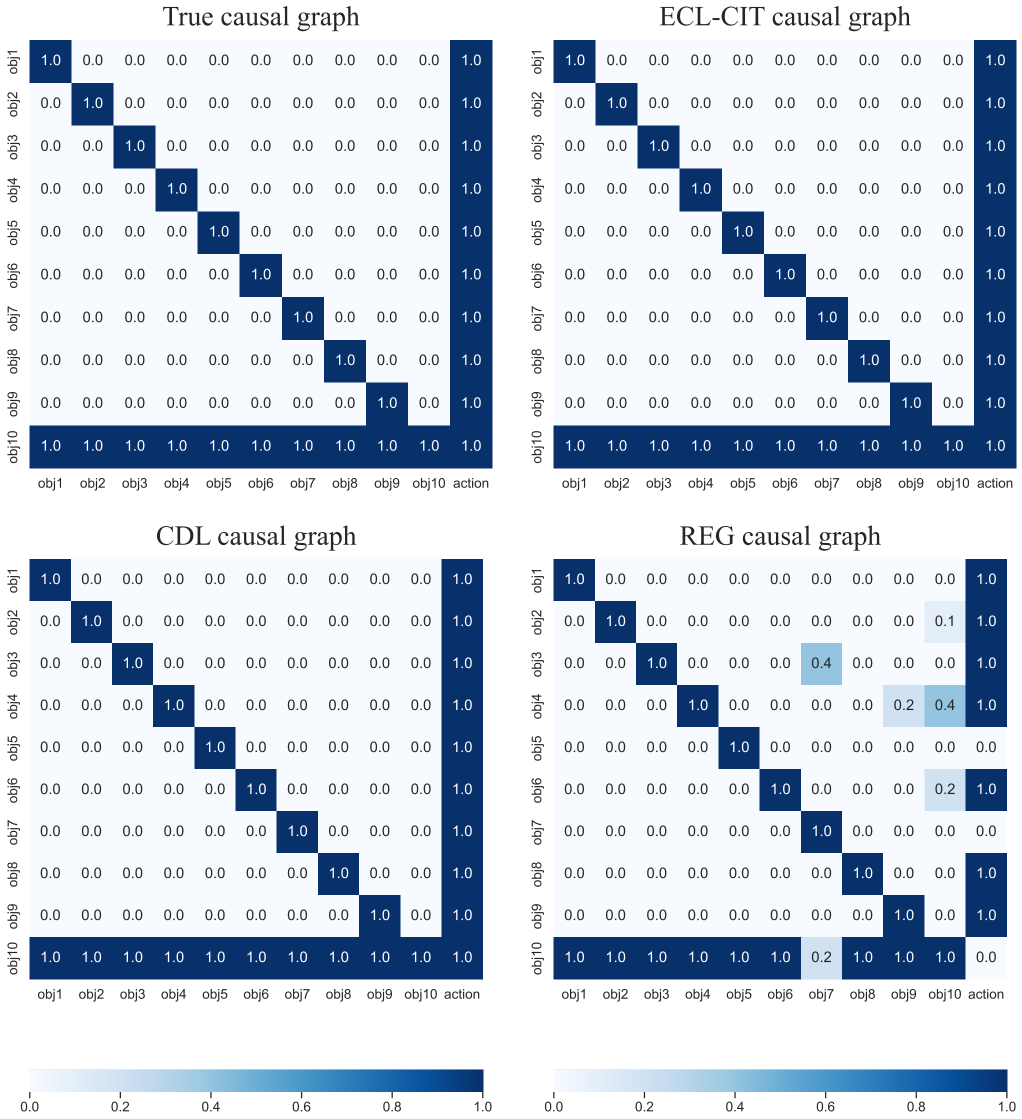
CDL causal graph

	obj1	obj2	obj3	obj4	obj5	obj6	obj7	obj8	obj9	obj10	action
obj1	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
obj2	0.2	1.0	0.2	0.0	0.0	0.0	0.0	0.2	0.0	0.0	1.0
obj3	0.0	0.2	1.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	1.0
obj4	0.0	0.0	0.0	1.0	0.3	0.0	0.0	0.0	0.0	0.0	1.0
obj5	0.0	0.0	0.0	0.4	1.0	0.0	0.0	0.0	0.0	0.0	1.0
obj6	0.0	0.0	0.2	0.0	0.2	1.0	0.0	0.0	0.0	0.0	1.0
obj7	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
obj8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.4	0.0	1.0
obj9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0
obj10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0

REG causal graph

	obj1	obj2	obj3	obj4	obj5	obj6	obj7	obj8	obj9	obj10	action
obj1	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
obj2	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
obj3	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
obj4	0.0	0.0	0.0	1.0	0.3	0.0	0.0	0.0	0.0	0.0	1.0
obj5	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0
obj6	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0
obj7	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	1.0
obj8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
obj9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
obj10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0

Figure 9: Causal graph for the chemical chain environment learned by the ECL, CDL and REG.



True causal graph

	obj1	obj2	obj3	obj4	obj5	obj6	obj7	obj8	obj9	obj10	action
obj1	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
obj2	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
obj3	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
obj4	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
obj5	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0
obj6	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0
obj7	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0
obj8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0
obj9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0
obj10	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

ECL-CIT causal graph

	obj1	obj2	obj3	obj4	obj5	obj6	obj7	obj8	obj9	obj10	action
obj1	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
obj2	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
obj3	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
obj4	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
obj5	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0
obj6	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0
obj7	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0
obj8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0
obj9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0
obj10	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

CDL causal graph

	obj1	obj2	obj3	obj4	obj5	obj6	obj7	obj8	obj9	obj10	action
obj1	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
obj2	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
obj3	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
obj4	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
obj5	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0
obj6	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0
obj7	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0
obj8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0
obj9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0
obj10	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

REG causal graph

	obj1	obj2	obj3	obj4	obj5	obj6	obj7	obj8	obj9	obj10	action
obj1	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
obj2	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	1.0
obj3	0.0	0.0	1.0	0.0	0.0	0.0	0.4	0.0	0.0	0.0	1.0
obj4	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.2	0.4	1.0
obj5	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
obj6	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.2	1.0
obj7	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
obj8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0
obj9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0
obj10	1.0	1.0	1.0	1.0	1.0	1.0	0.2	1.0	1.0	1.0	0.0

Figure 10: Causal graph for the chemical collider environment learned by the ECL, CDL and REG.

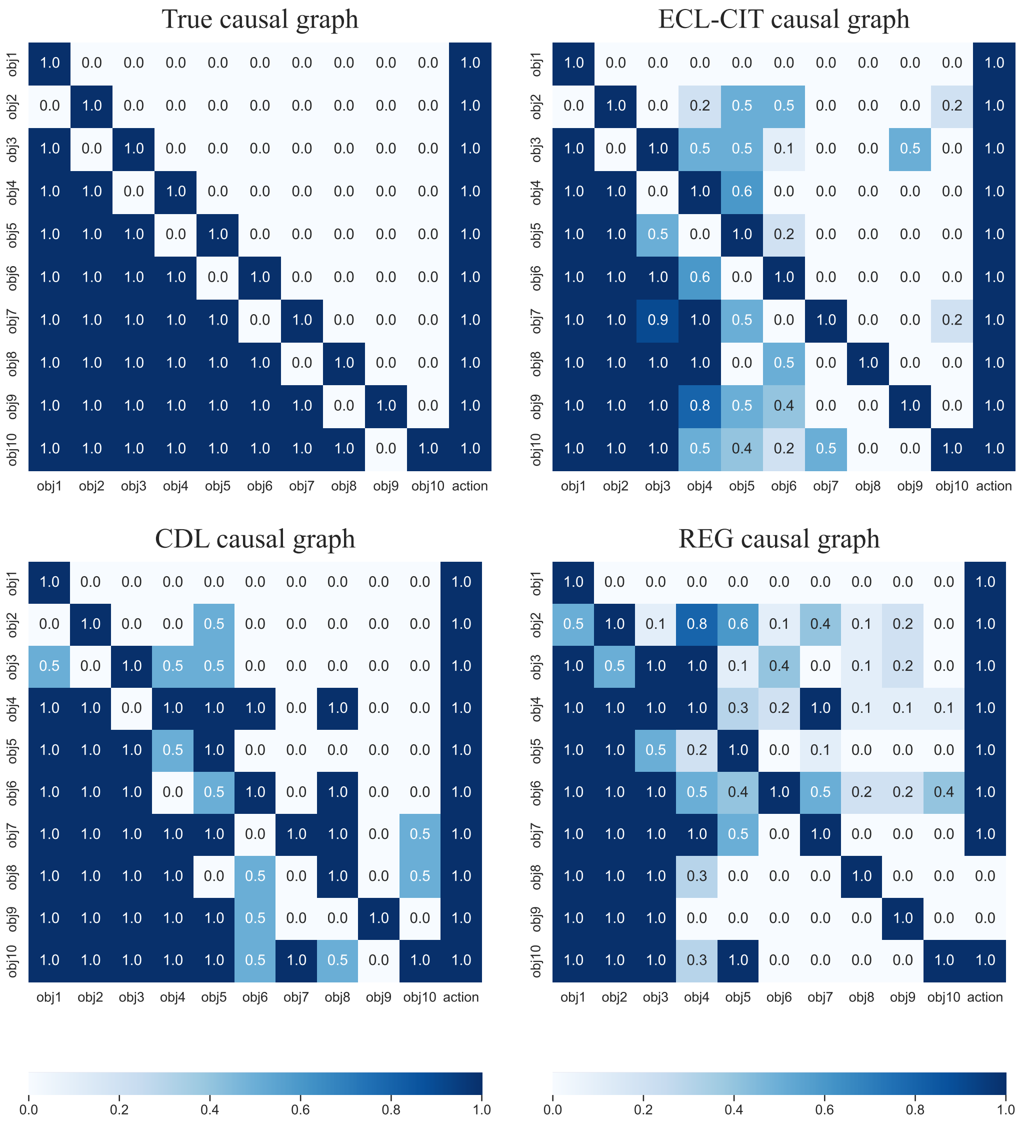
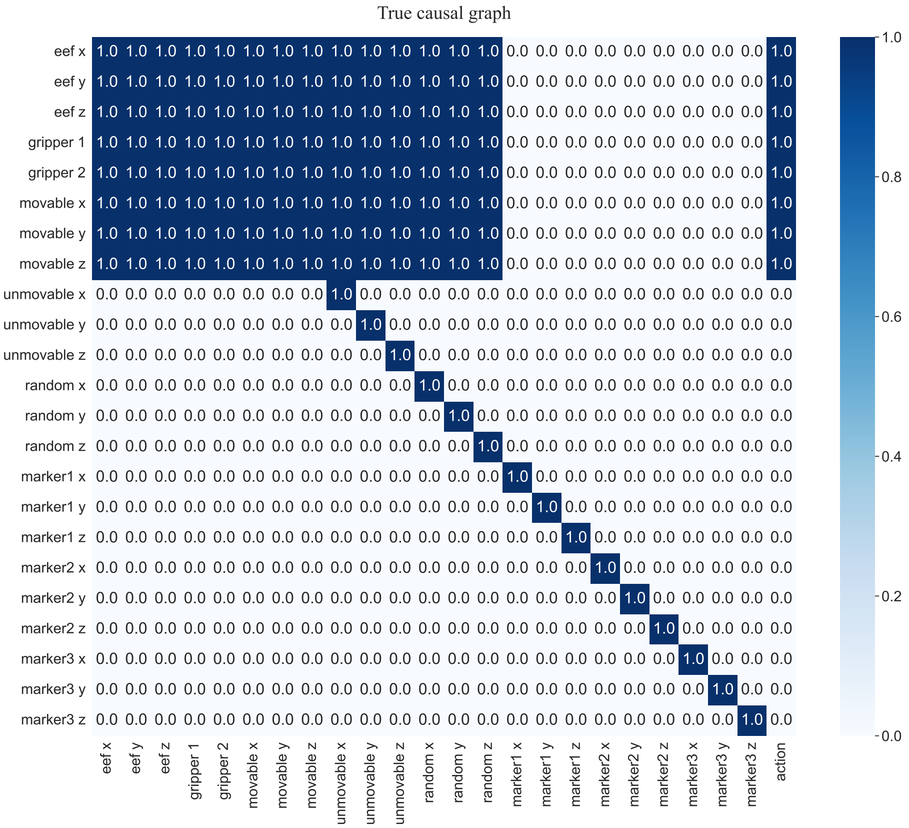
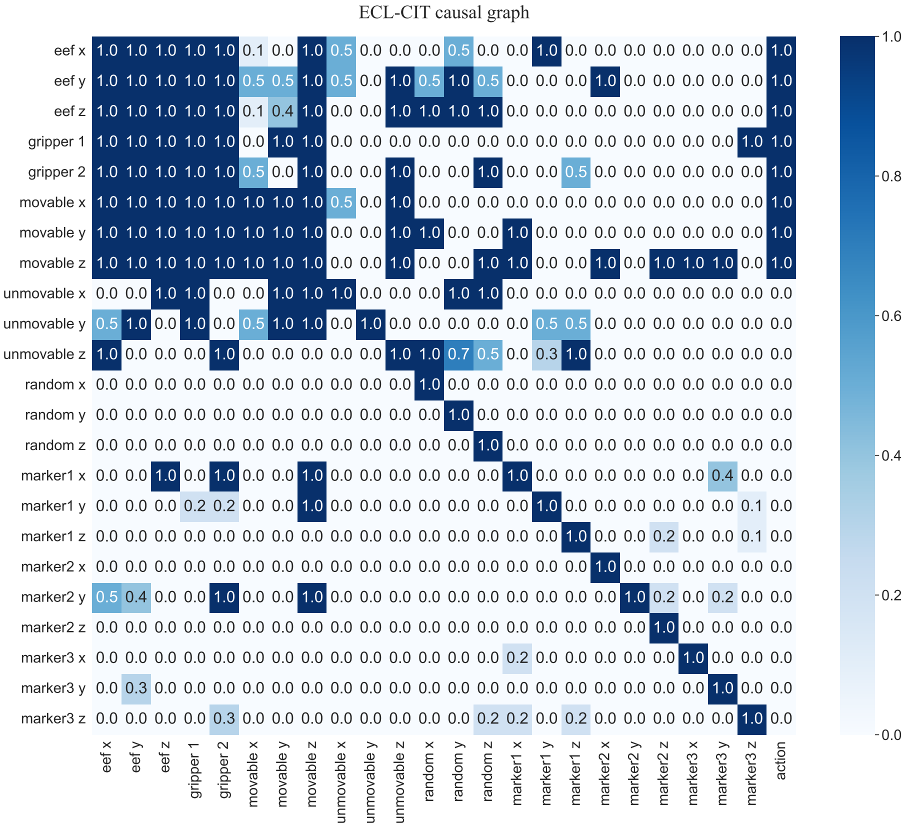


Figure 11: Causal graph for the chemical full environment learned by the ECL, CDL and REG.

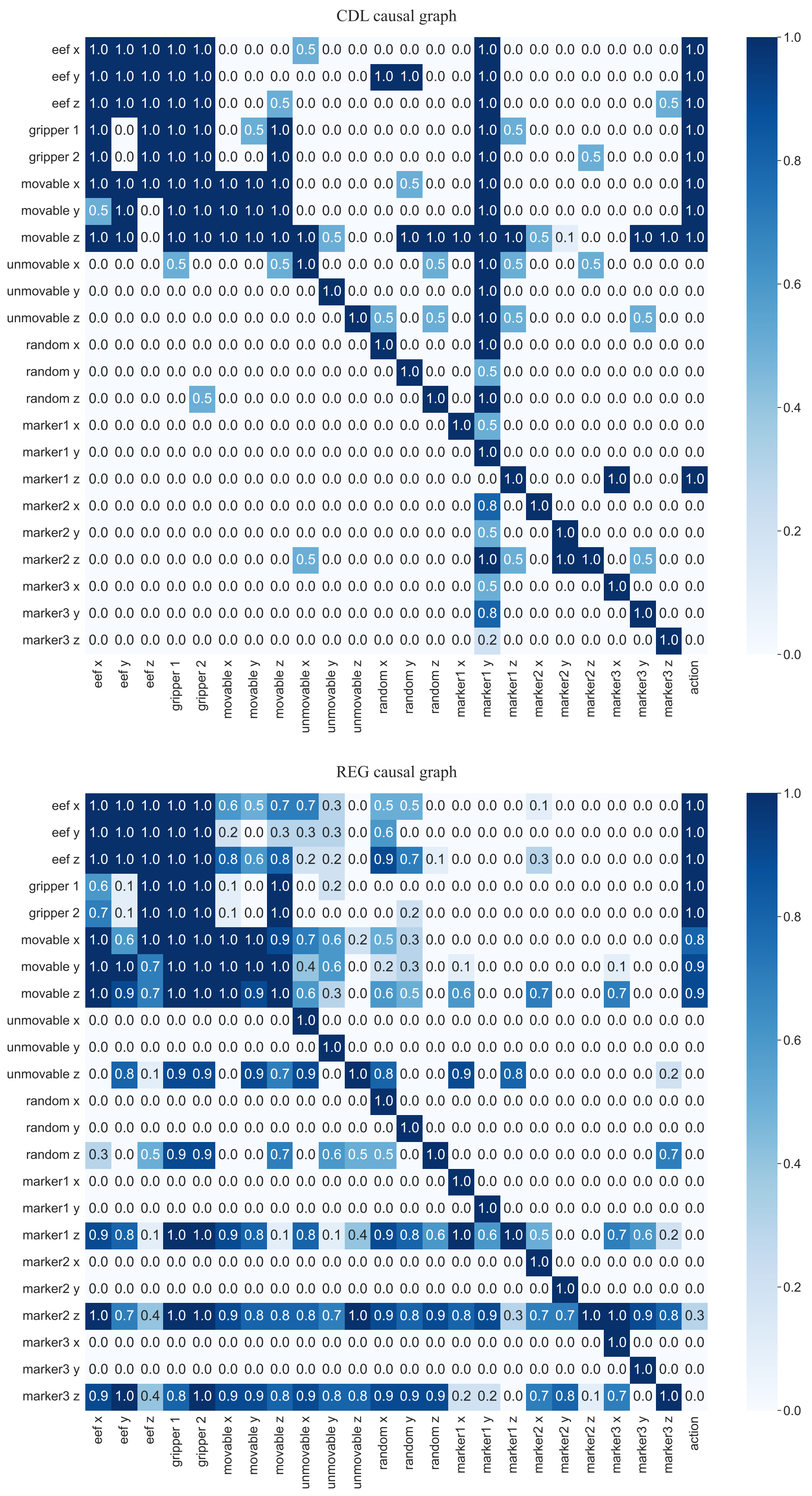


	eef x	eef y	eef z	gripper 1	gripper 2	movable x	movable y	movable z	unmovable x	unmovable y	unmovable z	random x	random y	random z	marker1 x	marker1 y	marker1 z	marker2 x	marker2 y	marker2 z	marker3 x	marker3 y	marker3 z	action
eef x	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
eef y	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
eef z	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
gripper 1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
gripper 2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
movable x	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
movable y	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
movable z	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
unmovable x	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
unmovable y	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
unmovable z	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
random x	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
random y	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
random z	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
marker1 x	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
marker1 y	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
marker1 z	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
marker2 x	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
marker2 y	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
marker2 z	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
marker3 x	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
marker3 y	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
marker3 z	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0



	eef x	eef y	eef z	gripper 1	gripper 2	movable x	movable y	movable z	unmovable x	unmovable y	unmovable z	random x	random y	random z	marker1 x	marker1 y	marker1 z	marker2 x	marker2 y	marker2 z	marker3 x	marker3 y	marker3 z	action
eef x	1.0	1.0	1.0	1.0	1.0	0.1	0.0	1.0	0.5	0.0	0.0	0.0	0.5	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
eef y	1.0	1.0	1.0	1.0	1.0	0.5	0.5	1.0	0.5	0.0	1.0	0.5	1.0	0.5	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0
eef z	1.0	1.0	1.0	1.0	1.0	0.1	0.4	1.0	0.0	0.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
gripper 1	1.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0
gripper 2	1.0	1.0	1.0	1.0	1.0	0.5	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	1.0
movable x	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.5	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
movable y	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
movable z	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0	0.0	1.0	0.0	1.0	1.0	1.0	0.0	1.0
unmovable x	0.0	0.0	1.0	1.0	0.0	0.0	1.0	1.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
unmovable y	0.5	1.0	0.0	1.0	0.0	0.5	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.5	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0
unmovable z	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.7	0.5	0.0	0.3	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
random x	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
random y	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
random z	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
marker1 x	0.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0
marker1 y	0.0	0.0	0.0	0.2	0.2	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
marker1 z	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.2	0.0	0.0	0.1	0.0
marker2 x	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
marker2 y	0.5	0.4	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.2	0.0	0.2	0.0	0.0
marker2 z	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
marker3 x	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
marker3 y	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
marker3 z	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.2	0.0	0.2	0.0	0.0	0.0	0.0	0.0	1.0	0.0

Figure 12: Causal graph for the manipulation environment learned by the true graph and ECL.



CDL causal graph

	eef x	eef y	eef z	gripper 1	gripper 2	movable x	movable y	movable z	unmovable x	unmovable y	unmovable z	random x	random y	random z	marker1 x	marker1 y	marker1 z	marker2 x	marker2 y	marker2 z	marker3 x	marker3 y	marker3 z	action
eef x	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
eef y	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
eef z	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	1.0
gripper 1	1.0	0.0	1.0	1.0	1.0	0.0	0.5	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	1.0
gripper 2	1.0	0.0	1.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	1.0
movable x	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
movable y	0.5	1.0	0.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
movable z	1.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	0.5	0.0	0.0	1.0	1.0	1.0	1.0	1.0	0.5	0.1	0.0	0.0	1.0	1.0	1.0
unmovable x	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.5	1.0	0.0	0.0	0.0	0.0	0.5	0.0	1.0	0.5	0.0	0.0	0.5	0.0	0.0	0.0	0.0
unmovable y	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
unmovable z	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.5	0.0	0.5	0.0	1.0	0.5	0.0	0.0	0.0	0.0	0.5	0.0	0.0
random x	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
random y	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
random z	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
marker1 x	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
marker1 y	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
marker1 z	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0
marker2 x	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
marker2 y	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
marker2 z	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.5	0.0	1.0	1.0	0.0	0.5	0.0	0.0
marker3 x	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
marker3 y	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
marker3 z	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0

REG causal graph

	eef x	eef y	eef z	gripper 1	gripper 2	movable x	movable y	movable z	unmovable x	unmovable y	unmovable z	random x	random y	random z	marker1 x	marker1 y	marker1 z	marker2 x	marker2 y	marker2 z	marker3 x	marker3 y	marker3 z	action
eef x	1.0	1.0	1.0	1.0	1.0	0.6	0.5	0.7	0.7	0.3	0.0	0.5	0.5	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	1.0
eef y	1.0	1.0	1.0	1.0	1.0	0.2	0.0	0.3	0.3	0.3	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
eef z	1.0	1.0	1.0	1.0	1.0	0.8	0.6	0.8	0.2	0.2	0.0	0.9	0.7	0.1	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	1.0
gripper 1	0.6	0.1	1.0	1.0	1.0	0.1	0.0	1.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
gripper 2	0.7	0.1	1.0	1.0	1.0	0.1	0.0	1.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
movable x	1.0	0.6	1.0	1.0	1.0	1.0	1.0	0.9	0.7	0.6	0.2	0.5	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8
movable y	1.0	1.0	0.7	1.0	1.0	1.0	1.0	1.0	0.4	0.6	0.0	0.2	0.3	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.9
movable z	1.0	0.9	0.7	1.0	1.0	1.0	0.9	1.0	0.6	0.3	0.0	0.6	0.5	0.0	0.6	0.0	0.0	0.7	0.0	0.0	0.7	0.0	0.0	0.9
unmovable x	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
unmovable y	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
unmovable z	0.0	0.8	0.1	0.9	0.9	0.0	0.9	0.7	0.9	0.0	1.0	0.8	0.0	0.0	0.9	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.2	0.0
random x	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
random y	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
random z	0.3	0.0	0.5	0.9	0.9	0.0	0.0	0.7	0.0	0.6	0.5	0.5	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.0
marker1 x	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
marker1 y	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
marker1 z	0.9	0.8	0.1	1.0	1.0	0.9	0.8	0.1	0.8	0.1	0.4	0.9	0.8	0.6	1.0	0.6	1.0	0.5	0.0	0.0	0.7	0.6	0.2	0.0
marker2 x	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
marker2 y	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
marker2 z	1.0	0.7	0.4	1.0	1.0	0.9	0.8	0.8	0.8	0.7	1.0	0.9	0.8	0.9	0.8	0.9	0.3	0.7	0.7	1.0	1.0	0.9	0.8	0.3
marker3 x	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
marker3 y	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
marker3 z	0.9	1.0	0.4	0.8	1.0	0.9	0.9	0.8	0.9	0.8	0.8	0.9	0.9	0.9	0.2	0.2	0.0	0.7	0.8	0.1	0.7	0.0	1.0	0.0

Figure 13: Causal graph for the manipulation environment learned by CDL and REG.

D.5 Downstream tasks learning

As illustrated in Figures 14 and 15, ECL-CIT attains the highest reward across three environments when compared to dense models like GNN and MLP, as well as causal approaches such as CDL and REG. Notably, ECL-CIT outperforms other methods in intricate manipulation tasks. Furthermore, ECL-ASR surpasses REG, enhancing model performance and achieving a reward comparable to CDL. The proposed curiosity reward encourages exploration and avoids local optimality during the policy learning process. Moreover, ECL excels not only in accurately uncovering causal relationships but also in enabling efficient learning for downstream tasks.

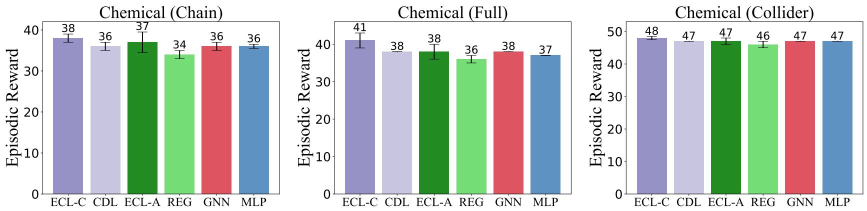


Figure 14: The task learning of episodic reward in three environments with ECL-CIT (ECL-C), ECL-ASR (ECL-A) and baselines.

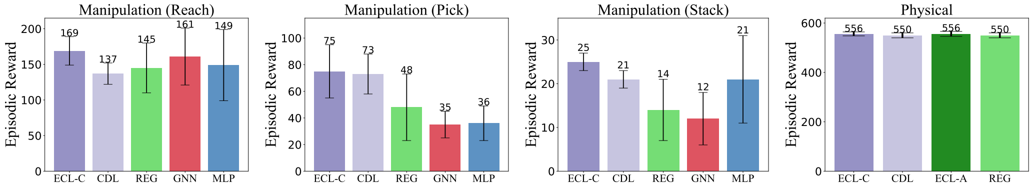


Figure 15: The task learning of episodic reward in three manipulation and pyhsical environments.

Sample efficiency analysis. We perform comparative analysis of downstream tasks learning across all environments. As depicted in Figure 16 for experiments in three chemical environments, we can find that ECL-CIT and ECL-ASR achieve outstanding performance in all three environments. Furthermore, the policy learning exhibits relative stability, reaching a steady state after approximately 400 episodes. Additionally, Figure 17 illustrates the reward learning scenarios in the other four environments. Within the intricate manipulation environment, ECL-CIT facilitates more expeditious policy learning. Moreover, in the dense physical environment, ECL-CIT and ECL-ASR also exhibit the most expeditious learning efficiency. The experimental results demonstrate that the proposed methods outperform CDL. Moreover, compared to CDL, ECL enhances sample efficiency, further corroborating the effectiveness of the proposed intrinsic-motivated empowerment method.

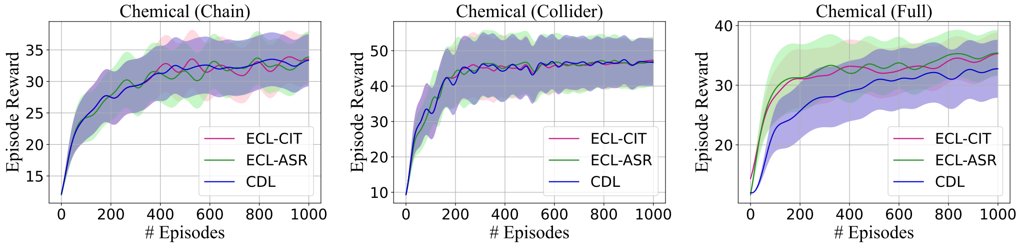


Figure 16: The task learning curves of episodic reward in three chemical environments and the shadow is the standard error.

D.6 Property analysis

Training steps analysis. For property analysis, we set different training steps for causal dynamics learning of ECL-CIT. As depicted in Figure 18, in the chemical chain environment, we observe that

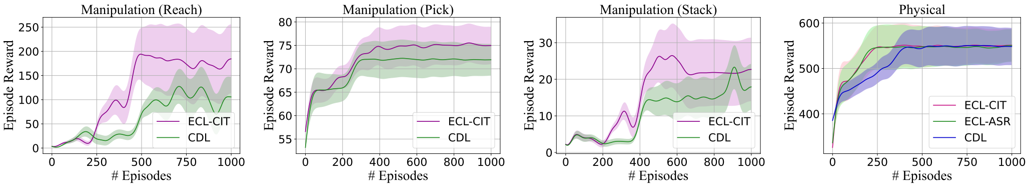


Figure 17: The task learning curves of episodic reward in four environments and the shadow is the standard error..

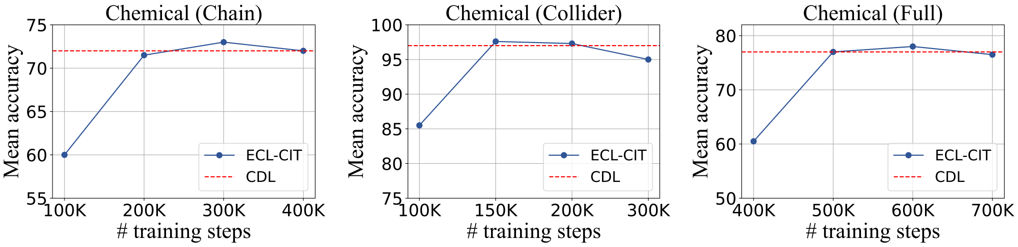


Figure 18: The mean accuracy of prediction with different training steps in chemical environments.

the mean prediction accuracy reaches its peak at 300k training steps. A similar trend is observed in the collider environment, where the maximum accuracy is achieved at 150k training steps. Although in the full environment, ECL attains its maximum accuracy at 600k steps, which is higher than the 500k steps used for training CDL, we notice that at 500k steps, ECL has already achieved performance comparable to CDL. These results substantiate that our proposed causal action empowerment method effectively enhances sample efficiency and dynamics performance.

Hyperparameter analysis. We further analyze the impact of the hyperparameter λ introduced in the downstream task reward function with CUR. We compare four different threshold settings, and the experimental results are depicted in Figure 19. From the results, we observe that when the parameter is set to 1, the policy learning performance is optimal. When the parameter is set to 0, the introduced CUR cannot encourage exploratory behavior in the policy. Nonetheless, it still achieves reward performance comparable to CDL. This finding further corroborates the effectiveness of our method for dynamics learning. Conversely, when this parameter is set excessively high, it causes the policy to explore too broadly, subjecting it to increased risks, and thus more easily leading to policy divergence. Through comparative analysis, we ultimately set this parameter to 1. In our future work, we will further optimize the improvement scheme for the reward function.

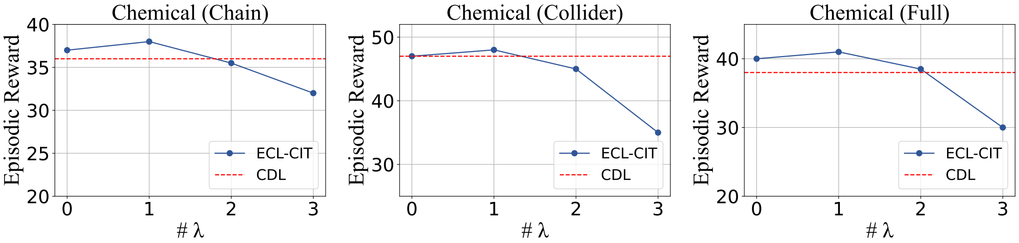


Figure 19: The episodic reward with different hyperparameter λ in three chemical environments.

D.7 Ablation Studies

To further validate the effectiveness of the various components comprising the proposed ECL method, we designed a series of ablation experiments for verification. First, we implement the method without offline model learning, simultaneously conducting causal model and task learning (w/i Sim) to verify the effectiveness of the proposed three-stage optimization framework. Second, we replace the curiosity reward introduced in the task learning with a causality motivation-driven reward (w/i Cau)

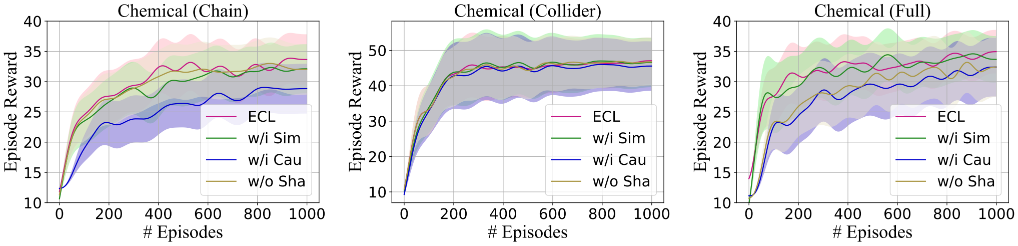


Figure 20: Learning curves of ablation Studies in three chemical environments and the shadow is the standard error. w/i represents with. w/o represents without.

$r_{\text{cau}} = \mathbb{E}_{(s_t,a_t,s_{t+1}\sim\mathcal{D})}\left[\mathbb{KL}\left(P_{\text{true}}||P_{\phi}\right) - \mathbb{KL}\left(P_{\text{true}}||P_{\phi_c}\right)\right]$, and a method without reward shaping (w/o Sha), respectively, to verify the effectiveness of incorporating the curiosity reward.

The results presented in Figure 20 clearly demonstrate the superior performance of the ECL over all other comparative approaches. ECL achieves the highest reward scores among the evaluated methods. Moreover, when compared to the method with Sim, ECL not only attains higher cumulative rewards but also exhibits greater stability in its performance during training. Additionally, ECL significantly outperforms the methods with Cau and method without Sha, further highlighting the efficacy of our proposed curiosity-driven exploration strategy in mitigating overfitting issues. By encouraging the agent to explore novel states and gather diverse experiences, the curiosity mechanism effectively prevents the policy from becoming overly constrained.

In summary, ECL facilitates effective and controllable policy learning for agents operating in complex environments. The curiosity-driven reward enables the agent to acquire a comprehensive understanding of the environment while simultaneously optimizing for the desired task objectives, resulting in superior performance and improved sample efficiency.

E Details on the Proposed Framework

Algorithm 1 lists the full pipeline of ECL below.

F Experimental Platforms and Licenses

F.1 Platforms

All experiments of this approach are implemented on 2 Intel(R) Xeon(R) Gold 6444Y and 4 NVIDIA RTX A6000 GPUs.

F.2 Licenses

In our code, we have utilized the following libraries, each covered by its respective license agreements:

- PyTorch (BSD 3-Clause "New" or "Revised" License)
- Numpy (BSD 3-Clause "New" or "Revised" License)
- Tensorflow (Apache License 2.0)
- Robosuite (MIT License)
- CausalMBRL (MIT License)

Algorithm 1 Towards Empowerment Gain through Causal Structure Learning in Model-Based RL

Input: policy network π_e, π_θ, transition collect policy π_{collect}, epoch length of dynamics model training, causal empowerment and downstream task policy learning H_{dyn}, H_{cau}, and H_{task}, evaluation frequency for causal mask learning f_{eval}

```
Step 1: Offline Model Learning

for each environment step $t$ do
    Collect transitions $\{(s_i, a_i, r_i, s'_i)\}_{i=1}^{|\mathcal{D}_{\text{env}}|}$ with $\pi_{\text{collect}}$ from environment
    Add transitions to replay buffer $\mathcal{D}_{\text{collect}}$
end for
for $epoch = 1, \cdots, H_{\text{dyn}}$ do
    Sample transitions $\{(s_i, a_i, s'_i)\}_{i=1}^{|\mathcal{D}_{\text{dyn}}|}$ from $\mathcal{D}_{\text{collect}}$
    Train dynamics model $\phi_{\text{c}}$ with $\{(s_i, a_i, s'_i)\}_{i=1}^{|\mathcal{D}_{\text{dyn}}|}$
    if $epoch \ \% \ f_{\text{eval}}$ == 0 then
        Sample transitions $\{(s_i, a_i, s'_i)\}_{i=1}^{|\mathcal{D}_{\text{cau}}|}$ from $\mathcal{D}_{\text{collect}}$
        Learn causal dynamics model with causal mask $\phi_{\text{c}-\text{dyn}}$ with different causal discovery methods
    end if
    Sample transitions $\{(s_i, a_i, r_i, s'_i)\}_{i=1}^{|\mathcal{D}_{\text{rew}}|}$ from $\mathcal{D}_{\text{collect}}$
    Train reward model $\varphi_r$ with $\{(s_i, a_i, r_i, s'_i)\}_{i=1}^{|\mathcal{D}_{\text{rew}}|}$ and $\psi(\cdot)$
end for

Step 2: Online Model Learning

Collection transitions $\{(s_i, a_i, r_i, s'_i)\}_{i=1}^{|\mathcal{D}_{\text{emp}}|}$ with policy $\pi_e$
for $epoch = 1, \cdots, H_{\text{cau}}$ do
    Maximize $(\mathcal{E}_{\phi_c}(s_{t+1}) - \mathcal{E}_{\phi}(s_{t+1}))$ with transitions sampled from $\mathcal{D}_{\text{emp}}$ for policy $\pi_e$ learning
    Add transitions sampled with $\pi_e$ to $\mathcal{D}_{\text{emp}}$
    if $epoch \ \% \ f_{\text{eval}}$ == 0 then
        Optimize causal mask $M$ with fixed $\phi_c$ and transitions sampled from $\mathcal{D}_{\text{emp}}$
    end if
end for

Step 3: Policy Learning

for $epoch = 1, \cdots, H_{\text{task}}$ do
    Collect transitions $\{(s_i, a_i, r_i, s'_i)\}_{i=1}^{|\mathcal{D}_{\text{task}}|}$ with $\pi_\theta$
    Compute predicted rewards $r_{\text{task}}$ by learned reward predictor
    Calculate curiosity reward $r_{\text{cur}}$ by Eq. 11
    Calculate $r \leftarrow r_{\text{task}} + \lambda r_{\text{cur}}$
    Optimize policy $\pi_\theta$ by the CEM planning
end for
return policy $\pi_\theta$
```