Causal Dynamics Learning for Task-Independent State Abstraction

Zizhao Wang [1] **Xuesu Xiao** [2] **Zifan Xu** [2] **Yuke Zhu** [2] **Peter Stone** [2 3]

Abstract

Learning dynamics models accurately is an important goal for Model-Based Reinforcement Learning (MBRL), but most MBRL methods learn a *dense* dynamics model which is vulnerable to spurious correlations and therefore generalizes poorly to unseen states. In this paper, we introduce *Causal Dynamics Learning for Task-Independent State Abstraction* (CDL), which first learns a theoretically proved *causal* dynamics model that removes unnecessary dependencies between state variables and the action, thus generalizing well to unseen states. A state abstraction can then be derived from the learned dynamics, which not only improves sample efficiency but also applies to a wider range of tasks than existing state abstraction methods. Evaluated on two simulated environments and downstream tasks, both the dynamics model and policies learned by the proposed method generalize well to unseen states and the derived state abstraction improves sample efficiency compared to learning without it.

1. Introduction

Model-based Reinforcement Learning (MBRL) enables an agent to predict what would happen if it executed various actions, thus allowing it to learn from *imagined* experience (Hafner et al., 2019). However, such an approach relies on the model being learned *accurately*. Many model-based approaches rely on *dense* models that predict the next step value of each variable based on the action and all variables in the current state, as shown in Fig. 1 (a). Such dense models are sensitive to spurious correlations which lead to poor generalization. For example, when door B is at angles unseen during training or the clock is at unseen times, the prediction of door A can be inaccurate due to unnecessary dependence on those variables in the model.

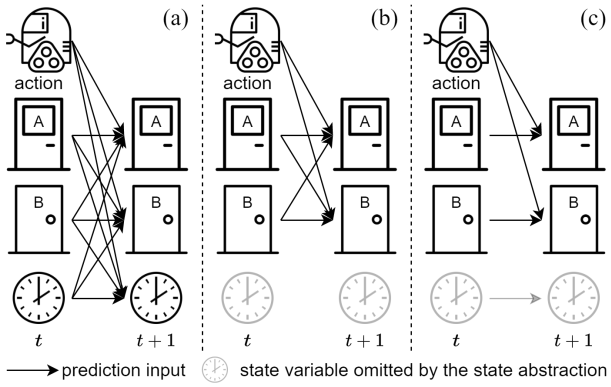

Figure 1. With two doors that the robot can open and go through and a clock on the wall, (a) *dense* dynamics models predict dynamics of each state variable unnecessarily using all variables; (b) based on a pre-defined reward (e.g., for navigation), existing state abstractions learn to omit the clock but still use a dense model for the remaining variables; (c) our *causal* models reason and only keep necessary dependencies (i.e., doors A and B depend on the action individually) and derives a state abstraction independent from any reward function.

Observing that unnecessary dependencies are the source of poor generalization, existing state abstractions mitigate this problem by removing some of those dependencies. Specifically, state abstractions group many states into an abstract state by omitting some state variables (Chapman & Kaelbling, 1991; McCallum, 1996; Jong & Stone, 2005). For example, bisimulation (Zhang et al., 2020b), which is particularly related to this paper, omits variables irrelevant to a pre-defined reward function. Though doing so removes unnecessary dependence on omitted variables, as shown in Fig. 1 (b), the same generalization issues persist as dense models are still used for the remaining variables, leaving unnecessary dependencies in the abstract state. Furthermore, despite bisimulation improving sample efficiency of task learning by reducing the learning space, such methods only find problem-dependent abstractions: they may omit variables that are irrelevant to the current task but that the agent can control and utilize for future tasks.

Given that good generalization requires only keeping necessary dependencies, this paper introduces Causal Dynamics Learning (CDL) for Task-Independent State Abstraction which learns a *causal* model that explicitly reasons about

[1]Department of Electrical and Computer Engineering, [2]Department of Computer Science, The University of Texas at Austin, Austin, USA [3]Sony AI. Correspondence to: Zizhao Wang <zizhao.wang@utexas.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

which actions and variables affect which variables from collected data, as shown in Fig. 1 (c). For example, by not depending on door B and the clock, the causal model's predictions about door A are unaffected by those variables and likely to be more accurate than dense models. Specifically, we prove that each variable's dependence on other variables (or actions) can be determined by a single conditional independence test. Such a test is then carried out by a novel architecture which estimates the conditional mutual information while learning the dynamics model.

Furthermore, by revealing all unnecessary dependencies, certain state variables which no other variables depend on (e.g., the clock) can be omitted for planning, forming a new form of state abstraction. Specifically, our model partitions state variables into those that it can change (*controllable variables*, e.g., doors A and B) with its actions, those that it cannot change but that influence actions' results on those that it can (*action-relevant variables*, e.g., an obstacle that may block door A's motion), and the remainder (*action-irrelevant variables*, e.g., the clock) which have no influence on others and thus can be omitted during planning. Also, in the abstract state, the dynamics model is still free of unnecessary dependencies and exhibits the same generalization benefits. Derived purely from dynamics, our state abstraction includes all controllable variables that the agent can use in the future, enabling it to solve a wider range of tasks than bisimulation which only retains variables specific to a single task.

CDL is compared against state-of-the-art dense models in two simulated environments: a chemical environment with causal relationships of different complexities, and a tabletop manipulation environment with challenging rigid body dynamics. We find that CDL learns causal relationships accurately and retains similar prediction accuracy on unseen states to the accuracy on seen states, while the prediction accuracies of dense models drop $60 \sim 90\%$ in some complex environments. When applied to downstream tasks, policies with the proposed causal state abstraction learn with higher sample efficiency and also generalize better than those with dense models.

2. Related Work

2.1. Model-based Reinforcement Learning

Model-based RL typically involves learning a dynamics model of the environment (including a reward predictor) by maximizing the likelihood of collected trajectories. Then the dynamics model and reward predictor are used for planning (Williams et al., 2017; Chua et al., 2018; Nagabandi et al., 2018), providing synthetic data (Kurutach et al., 2018; Janner et al., 2019), or improving Q-value estimates (Feinberg et al., 2018; Amos et al., 2021). However, most existing approaches model the dynamics in the dense form, where each state variable at the next time step depends on all current state variables and the action, e.g., as shown in Fig. 1 (a). This non-causal formulation suffers from spurious correlations and generalizes poorly for out-of-distribution states. Moreover, the reward predictor typically also has the same dense architecture, thus exacerbating the generalization issue.

In dynamics learning, there are models that explicitly consider modularity and sparsity (Goyal et al., 2019; 2020; 2021). However, without modelling the causal relationships, the sparse dependencies learned by those methods could still be spurious correlations and thus lead to poor generalization (see details in appendix Sec. B).

2.2. State Abstraction

Aiming at improving sample efficiency, state abstractions aggregate many states into a single abstract state (e.g., by omitting state variables that are irrelevant to the current task) while preserving policy optimality, thus enabling the agent to learn in a potentially much smaller state space. For example, bisimulation (Even-Dar & Mansour, 2003; Ravindran, 2004; Li, 2009) aggregates states that generate equivalent reward sequences given any action sequence originating from the states. Bisimulation is also closely tied to causality: work from Zhang et al. (2020a) shows that state variables kept by the bisimulation abstraction consist of only the causal ancestors of the reward variable. However, it still learns a dense model for those variables, as shown in Fig. 1 (b), and is subject to the same generalization issue mentioned earlier.

Furthermore, as the bisimulation abstraction is derived from the reward function, it is often limited to a small range of tasks. Specifically, the abstraction can only generalize to tasks where the new reward function is defined on the same or a subset of causal ancestors. For example, in the environment shown in Fig. 1, the abstraction learned from opening door A will only include door A, thus is unable to solve tasks like opening door B. In the real world, one generally wants a robot to be able to learn to perform a wide range of tasks (e.g., manipulating different objects); training this bisimulation abstraction for each task (object) is impractical. In contrast, CDL derives a state abstraction which includes all controllable state variables from the learned dynamics. As a result, it can be applied to a wider range of tasks than bisimulation.

2.3. Causality in Reinforcement Learning

Incorporating causality into reinforcement learning has received widespread interest recently, due to its generalization and transferability benefits, in directions that include but are not limited to representation learning (Zhang et al., 2019;

2020a; Sontakke et al., 2021; Volodin et al., 2020; Tomar et al., 2021; Fu et al., 2021), policy learning (Buesing et al., 2018; Nair et al., 2019; Mozifian et al., 2020; Lyle et al., 2021; Seitzer et al., 2021; Lu et al., 2020), and dynamics learning (Li et al., 2020). Among the literature, the work closest to ours is Wang et al. (2021) which also explicitly learns the causal dependencies but by regularizing the number of variables used when predicting each state variable. However, there is no theoretical guarantee that dynamics models learned by those methods are causal. Moreover, it is found to be difficult for the method from Wang et al. (2021) to balance between prediction performance and regularization. A large regularization penalty often sacrifices the prediction accuracy of rare events whereas a small regularization penalty fails to suppress spurious correlations related to frequent events. In contrast, our method is theoretically sound and does not have a regularization parameter to tune.

3. Causal Dynamics Learning for Task-Independent State Abstraction (CDL)

In this section, we introduce CDL, which learns a causal dynamics model and then derives a state abstraction that supports learning a wide range of tasks, from the learned causal relationships.

3.1. Background

3.1.1. Markov Processes

We consider the agent's interaction with the environment as a Markov Process defined as $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P})$, where $\mathcal{S} \subseteq \mathbb{R}^{d_\mathcal{S}}$ is the $d_\mathcal{S}$-dimensional state space, $\mathcal{A} \subseteq \mathbb{R}^{d_\mathcal{A}}$ is the $d_\mathcal{A}$-dimensional action space, $\mathcal{P}$ is the transition function. The state space consists of $d_\mathcal{S}$ state variables, one for each dimension of the state space, denoted as $\{\mathcal{S}^i\}_{i=1}^{d_\mathcal{S}}$.

3.1.2. Causal Graphical Models

A causal graphical model (Pearl, 2009) is defined by a distribution p_X over d random variables $X^1, \dots, X^d$ and a Directed Acyclic Graph $\mathcal{G} = (V, E)$. Each node $i \in V = \{1, \dots, d\}$ is associated with a random variable X^i and each edge $(i \to j) \in E$ represents that X^i is a **direct cause** of X^j, i.e., X^j depends on X^i during the data generation process. Then the distribution can be specified as

$$p_X(x^1, \dots, x^d) = \prod_{i=1}^{d} p_{X^i}(x^i | \mathbf{PA}_i), \tag{1}$$

where $\mathbf{PA}_i$ is the set of parents of the node i in the graph $\mathcal{G}$.

For simplicity, in the remainder of the paper, we denote $v^{1:n}$ (or $v_{1:n}$) for an abbreviation of a set of variables $\{v^i\}_{i=1}^n$ (or $\{v_i\}_{i=1}^n$) and will omit the subscript of distributions, i.e., $p_V(v) = p(v)$ and $p_V(v|w) = p(v|w)$.

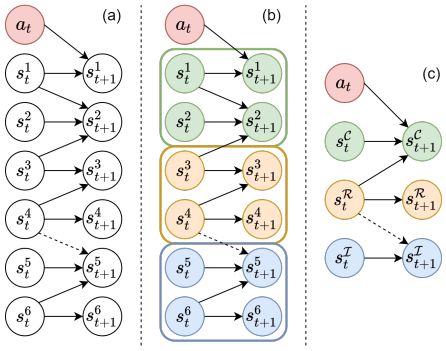

Figure 2. **(a)** an example causal dynamics model. **(b)** state variables can be split into three types: controllable (green), action-relevant (orange), and action-irrelevant (blue). The dashed arrow represents whether it exists does not affect s^5 to be action-irrelevant. **(c)** the causal graph can be split into three subgraphs, one for each type of state variable.

3.2. Problem Definition

We begin with a formal definition of controllable, action-relevant, and action-irrelevant state variables. If we are given the causal graphical model of a Markov Process with $V = \{s_t^{1:d_\mathcal{S}}, a_t, s_{t+1}^{1:d_\mathcal{S}}\}$ as nodes and E as edges describing causal relationships from $s_t^{1:d_\mathcal{S}}$ and a_t to $s_{t+1}^{1:d_\mathcal{S}}$, ancestors of a state variable s^i are defined as all nodes that have a directed path leading to node s^i (not necessarily from the immediate previous time step but can be from any previous step). For example, for the causal dynamics model shown in Fig. 2 (a), s^4 is an ancestor of s^2 as there is a path of $s_t^4 \to s_{t+1}^3 \to s_{t+2}^2$. Descendants of nodes are defined in the same way but in the opposite direction. Then we have:

Definition 1 (Controllable State Variables) $s^\mathcal{C}$ are the descendants of the action a_t.

Definition 2 (Action-Relevant State Variables) $s^\mathcal{R}$ are ancestors of controllable state variables, excluding those already belonging to $s^\mathcal{C}$.

Definition 3 (Action-Irrelevant State Variables) $s^\mathcal{I}$ are those that belong to neither $s^\mathcal{C}$ nor $s^\mathcal{R}$.

In the above definitions, $\mathcal{C}$, $\mathcal{R}$, and $\mathcal{I}$ are the set of state dimension indices for controllable, action-relevant, and action-irrelevant state variables respectively.

Given these definitions, the type of each state variable in the example causal dynamics model is shown in Fig. 2 (b). Further in (c), one may notice that the causal graph can be split into three parts, allowing us to rewrite the transition probabilities as $\mathcal{P}(s_{t+1}|s_t, a_t) = p(s_{t+1}^\mathcal{C}|s_t^\mathcal{C}, s_t^\mathcal{R}, a_t) \cdot p(s_{t+1}^\mathcal{R}|s_t^\mathcal{R}) \cdot p(s_{t+1}^\mathcal{I}|s_t^\mathcal{R}, s_t^\mathcal{I})$, where the edge from $s_t^\mathcal{R}$ to $s_{t+1}^\mathcal{I}$ is optional as it does not change the split of state variables ($s^\mathcal{I}$ are still neither a_t's descendants nor $s^\mathcal{C}$'s ancestors).

Then CDL forms the state abstraction ϕ by omitting action-irrelevant state variables, i.e., $\phi(s_t) = (s_t^\mathcal{C}, s_t^\mathcal{R})$, and the

dynamics in the abstract space can be expressed by removing the subgraph involving action-irrelevant state variables, remaining a causal dynamics model itself, as follows:

$$\mathcal{P}(\phi(s_{t+1})|\phi(s_t), a_t) = p(s^{\mathcal{C}}_{t+1}|s^{\mathcal{C}}_t, s^{\mathcal{R}}_t, a_t) \cdot p(s^{\mathcal{R}}_{t+1}|s^{\mathcal{R}}_t).$$

This state abstraction ϕ can be used to solve any *actively-accomplishable* downstream task. Here, downstream tasks are those defined in the same Markov Process so that the agent can use ϕ to solve the task by learning the provided rewards. Meanwhile, actively-accomplishable means that the reward function of the task only depends on controllable and action-related state variables $(s^{\mathcal{C}}, s^{\mathcal{R}})$. Additionally, if the task involves extra variables $\mathcal{V}$, e.g., a *varying* goal g_t for certain controllable state variables to reach, those variables should also be provided so that the agent can learn the reward accurately. Notice that actively-accomplishable tasks do not cover those involving action-irrelevant state variables (e.g., for the example in Fig. 1, a reward of $+1$ for opening door A when the clock shows 1 pm and 0 otherwise). However, as the reward function of a task can be arbitrarily designed using any state variable, being able to solve all tasks means that no state variable can be omitted (i.e., no state abstraction). We assume that in practice, it is relatively uncommon for a task's reward function to involve action-irrelevant state variables, making ϕ fairly generally applicable.

So far, we have defined three types of state variables and the derived state abstraction for a known causal dynamics model. However, for real-world problems, such a model is usually not accessible. Instead, agents can only collect transition data via its interactions with the environment. Hence, this paper introduces a novel method that: (1) learns a **causal** dynamics model $F_\theta : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$, (2) learns a transition collection policy $\pi : \mathcal{S} \to \mathcal{A}$ to learn F_θ efficiently, (3) derives the state abstraction $\phi : \mathcal{S} \to \mathcal{S}^{\mathcal{C}} \times \mathcal{S}^{\mathcal{R}}$ and dynamics F^{ϕ}_θ in the abstract space, (4) learns a reward predictor for any actively-accomplishable task in the abstract space $R_\varphi : \phi(\mathcal{S}) \times \mathcal{A}\ (\times\mathcal{V}) \to \mathbb{R}$, and (5) uses planning methods to solve the task with the learned F^{ϕ}_θ and R_φ.

3.3. Causal Dynamics Learning

The key challenge when learning a causal dynamics model is to determine whether a causal edge exists between two state variables, i.e., $s^i_t \to s^j_{t+1}$. First, adapting the work from Mastakouri et al. (2021), we present a method for inferring whether such a causal relationship exists, based on the following assumptions about the ground truth dynamics model:

Assumptions

A1. Causal Markov condition and Faithfulness in the underlying dynamics (definitions in appendix Sec. A.1).
A2. The state is fully observable and the dynamics is Markovian.
A3. The edge $s^i_t \to s^i_{t+1}$ exists for all state variables s^i.
A4. No simultaneous or backward edges in time, i.e., for all i, j, $s^i_t \nrightarrow s^j_t$ and $s^i_t \nrightarrow s^j_{t-1}$.
A5. The transitions for each state variable are independent, i.e., $\mathcal{P}(s_{t+1}|s_t, a_t) = \prod_{j=1}^{d_{\mathcal{S}}} p(s^j_{t+1}|s_t, a_t)$

Except for A4, which requires no redundant information in state variables (e.g., state variables that include both joint angles and the end-effector of a robot arm), and A5, which does not necessarily hold for rich observation spaces (e.g., images), these assumptions are commonly made for causal inference and dynamic systems. Moreover, for partially observable or high-dimensional state spaces, low-dimensional disentangled representations that encode the space can be learned to adhere to A2 and A5.

Theorem 3.1. *Assuming A1 - A5, we define the conditioning set* $\{a_t, s_t \setminus s^i_t\} = \{a_t, s^1_t, \ldots, s^{i-1}_t, s^{i+1}_t, \ldots, s^{d_{\mathcal{S}}}_t\}$. *Then, for any two state variables at indices* i *and* j, *if* $s^i_t \not\perp\!\!\!\perp s^j_{t+1}|\{a_t, s_t \setminus s^i_t\}$, *then* $s^i_t \to s^j_{t+1}$. *Similarly, if* $a_t \not\perp\!\!\!\perp s^j_{t+1}|s_t$, *then* $a_t \to s^j_{t+1}$.

Proof in appendix Sec. A.2. This result shows that the causal relationship between state variables (or between the action and any state variable) can be inferred with one Conditional Independence Test (CIT). For simplicity, in the remainder of the paper, we will only describe the CIT between two state variables, $s^i_t \not\perp\!\!\!\perp s^j_{t+1}|\{a_t, s_t \setminus s^i_t\}$. For the test between the action and the state variable, $a_t \not\perp\!\!\!\perp s^j_{t+1}|s_t$, the same method applies by changing s^i_t to a_t and the conditioning set according to Theorem 3.1.

In theory, the CIT between s^i_t and s^j_{t+1} can be made by measuring the Conditional Mutual Information, CMI^{ij}:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{s_t, a_t, s^j_{t+1}} \left[\log \frac{p(s^i_t, s^j_{t+1}|\{a_t, s_t \setminus s^i_t\})}{p(s^i_t|\{a_t, s_t \setminus s^i_t\}) p(s^j_{t+1}|\{a_t, s_t \setminus s^i_t\})} \right] \\
&= \mathop{\mathbb{E}}_{s_t, a_t, s^j_{t+1}} \left[\log \frac{p(s^j_{t+1}|a_t, s_t)}{p(s^j_{t+1}|\{a_t, s_t \setminus s^i_t\})} \right], \qquad (2)
\end{aligned}$$

where the expectation is over the joint distribution of $\{s_t, a_t, s^j_{t+1}\}$, all p are the ground truth conditional densities, and the derivation of Eq. (2) is in the appendix Sec. A.3. If $\text{CMI}^{ij} \geq \epsilon$ where ϵ is a pre-defined threshold, it suggests that s^i_t is necessary to predict s^j_{t+1} accurately, $s^i_t \not\perp\!\!\!\perp s^j_{t+1}|\{a_t, s_t \setminus s^i_t\}$ is true, and the causal edge $s^i_t \to s^j_{t+1}$ exists.

In practice, as the ground truth joint distribution and conditional densities are unknown, the expectation is computed over the collected transition data $\mathcal{D}$ and CDL learns predictive models $\hat{p}(s^j_{t+1}|a_t, s_t), \hat{p}(s^j_{t+1}|\{a_t, s_t \setminus s^i_t\})$ to approximate conditional densities.

However, computing CMI^{ij} for all state variable pairs

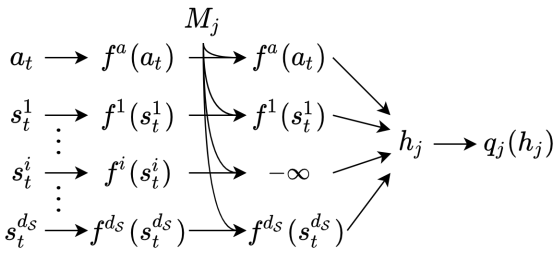

Figure 3. The predictive model for the state variable s_{t+1}^j. Different conditional densities can be represented by applying different masks M_j. $\hat{p}(s_{t+1}^j|\{a_t, s_t \setminus s_t^i\})$ is shown as an example in the figure.

requires training $(d_\mathcal{S})^2$ predictive models, which is intractable. Instead, we develop a novel architecture and training paradigm which reduces the requirement to $d_\mathcal{S}$ models. As shown in Fig. 3, to predict each state variable s_{t+1}^j: (1) the action and all current state variables are individually mapped to features $f_j^a(a_t), f_j^1(s_t^1), \ldots, f_j^{d_\mathcal{S}}(s_t^{d_\mathcal{S}})$ where each feature is a d_f-dimensional vector; (2) certain features are masked to $-\infty$ according to a binary map M_j; (3) the overall feature h_j is computed by taking the element-wise max of all features; and (4) a predictive network q_j takes h_j as input and predicts the distribution of s_{t+1}^j. This architecture can represent conditional densities with different masks M_j: (1) for $\hat{p}(s_{t+1}^j|a_t, s_t)$, none of the features is masked; (2) for $\hat{p}(s_{t+1}^j|\{a_t, s_t \setminus s_t^i\})$, the feature $f_j^i(s_t^i)$ is masked to $-\infty$ so that it will not be used to predict s_{t+1}^j, as shown in Fig. 3; and (3) for $\hat{p}(s_{t+1}^j|\mathbf{PA}_{s^j})$, after deriving parents of s_{t+1}^j from $\{\text{CMI}^{ij}\}_{i=1}^{d_\mathcal{S}}$), all non-parent features are masked to $-\infty$.

The architecture described above predicts one state variable s_{t+1}^j, and the whole causal dynamics model F_θ consists of $d_\mathcal{S}$ such models, where θ parameterizes all feature extractors $\{f_j^a, f_j^{1:d_\mathcal{S}}\}_{j=1}^{d_\mathcal{S}}$ and predictive networks $q_{1:d_\mathcal{S}}$. To train θ, we maximize the following log-likelihood:

$$\mathcal{L}_\theta = \sum_{j=1}^{d_\mathcal{S}} \big[\log \hat{p}(s_{t+1}^j|a_t, s_t) + \log \hat{p}(s_{t+1}^j|\{a_t, s_t \setminus s_t^i\}) + \log \hat{p}(s_{t+1}^j|\mathbf{PA}_{s^j})\big], \tag{3}$$

where i is uniformly sampled from $\{1, \ldots, d_\mathcal{S}\}$ for each j, and $\mathbf{PA}_{s^j}$ are inferred from $\{\text{CMI}^{ij}\}_{i=1}^{d_\mathcal{S}}$ learned so far. In Equation 3, the first two predictive likelihoods train models necessary to evaluate CMI^{ij}, and the last prediction likelihood finetunes the performance of the inferred causal dynamics model $\hat{p}(s_{t+1}^j|\mathbf{PA}_{s^j})$. We split the collected transition data $\mathcal{D}$ into the training part used to maximize $\mathcal{L}_\theta$ and the validation part for evaluating CMI.

After training, CDL evaluates the causal graph by checking if each $\text{CMI}^{ij} \geq \epsilon$ and derives the learned state abstraction $\phi(s) = (\hat{s}^\mathcal{C}, \hat{s}^\mathcal{R})$ from the learned causal graph, according to Definitions 1-3 (for a ground truth variable or distribution x, we denote $\hat{x}$ as its learned prediction or distribution). The dynamics in the abstract space F_θ^ϕ can also be derived by omitting the prediction networks for action-irrelevant state variables $\hat{s}^\mathcal{I}$.

3.4. Policy Learning for Transition Collection

The collected transition data $\mathcal{D}$ are important to accurate causal dynamics learning, as they are used in predictive model training (Eq. 3) and CMI evaluations (Eq. 2). An ideal transition collection policy π would cover all state-action pairs to expose causal relationships thoroughly and actively explore states where the causal dynamics model is not accurate.

To this end, an RL agent is used for transition collection with a reward function that is the prediction difference between the dense predictor and the causal predictor learned so far:

$$r_t = \tanh\left(\tau \cdot \sum_{j=1}^{d_\mathcal{S}} \log \frac{\hat{p}(s_{t+1}^j|s_t, a_t)}{\hat{p}(s_{t+1}^j|\mathbf{PA}_{s^j})}\right), \tag{4}$$

where τ is the scaling factor and $\tanh$ is used to keep the reward bounded. This reward motivates taking transitions where the dense predictor is better than the causal predictor, which usually suggests the learned causal graph is inaccurate.

3.5. Policy Learning for Downstream Tasks

When solving actively-accomplishable downstream tasks, like many MBRL algorithms, CDL simultaneously (1) learns a reward predictor with the abstract state, action, and any provided extra variables as input, $R_\varphi : \phi(\mathcal{S}) \times \mathcal{A}\ (\times \mathcal{V}) \to \mathbb{R}$, and (2) uses a planning algorithm with the learned dynamics and reward predictor for action selection. As the reward predictor is learned in an abstract space rather than the full state space, it is more sample-efficient and less vulnerable to spurious correlations brought about by excessive information (i.e., action-irrelevant state variables). Meanwhile, planning in the abstract space also reduces the computation cost by relieving the need to roll out action-irrelevant state variables.

The reward predictor is modeled as a neural network and trained by minimizing the prediction error,

$$\varphi^* = \arg\min_\varphi \mathbb{E}_{(s_t, a_t, g_t, r_t) \sim \mathcal{B}}\, \mathcal{L}(R_\varphi(\phi(s_t), a_t, g_t), r_t), \tag{5}$$

where $\mathcal{B}$ is the task data collected so far, and $\mathcal{L}$ can take any loss function (we experiment with the absolute value of the prediction error).

For planning, we use the cross entropy method (CEM, Rubinstein (1997)), a population-based optimization algorithm,

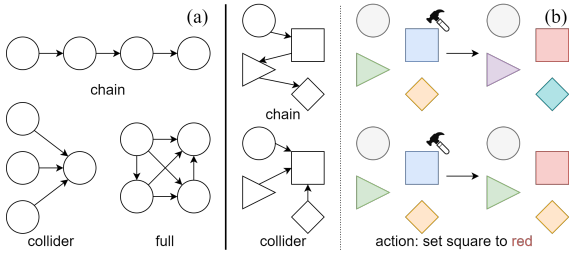

Figure 4. (a) different types of causal graphs. (b) illustration of the chemical environment (left: ground truth causal graphs, right: transitions after applying the action).

to search for the best action with the learned dynamics and reward predictor. For each time step t, depending on whether the action is continuous or discrete, CEM initializes a time-dependent belief over the optimal action sequence $a_{t:t+L} \sim \mathcal{N}(\mu_{t:t+L}, \sigma^2_{t:t+L})$ or Categorical$(\alpha_{t:t+L})$ where L is the planning length. Starting from a unit normal distribution for $\mathcal{N}$ or a uniform distribution for Categorical, it repeatedly samples J candidate action sequences, evaluates them based on cumulative rewards, and updates the action belief to the distribution of the top K candidates. After N iterations, the planner returns μ_t or $\arg\max \alpha_t$ as the current optimal action.

4. Experiments

Our experiments examine the following hypotheses:
1. CDL can infer the causal graph, with less tuning and better accuracy than a regularization-based method (Wang et al., 2021) (Sec. 4.3.1).
2. Causal models learned by CDL generalize better than dense models in both dynamics and downstream task learning on unseen states (Sec. 4.3.2).
3. The transition collection policy learned by CDL improves the accuracy of causal dynamics learning compared to a uniform policy or the policy learned from the curiosity reward (Pathak et al., 2017). (Sec. 4.4).
4. State abstractions derived by CDL improve sample-efficiency when learning downstream tasks compared to dense models which use the full state space (Sec. 4.5).
5. Policies learned by CDL for downstream tasks generalize better than those learned by dense models (Sec. 4.5).

4.1. Environments

We use (1) the Chemical environment modified from Ke et al. (2021) (Fig. 4 (b)) to examine CDL's performance on causal graphs with different complexities, and (2) a manipulation environment (Fig. 5) to validate CDL with complex rigid-body dynamics.

In the chemical environment, there are 10 objects whose x,y positions are continuous and whose colors are one of 5 colors; the action can change a chosen object to a specified

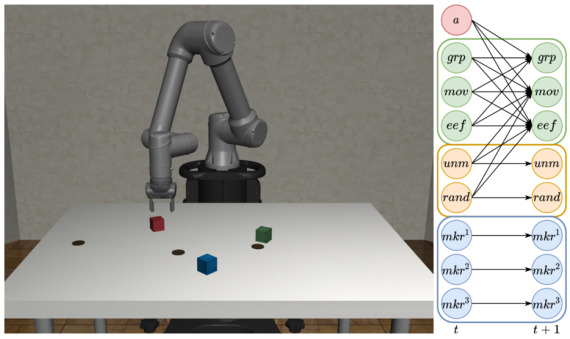

Figure 5. **(Left)** the manipulation environment. **(Right)** the learned causal graph in the object level which recovers the ground truth relationship between the state variables and the action.

new color. The interactions between objects take place according to the underlying causal graph - when one object is set to a new color, all its descendants change their colors in the order of the graph following a predefined conditional probability table. The dynamics of colors and positions is independent for all objects: positions are sampled from a normal distribution every step. We consider three graph structures shown in Fig. 4 (a): full, chain, and collider.

In the manipulation environment implemented with the `robosuite` simulation framework (Zhu et al., 2020), there are three objects that the robot can interact with: a *movable* red cube that the robot can manipulate (denoted as mov), an *unmovable* green cube that is fixed on the table (unm), and a *randomly moving* blue cube that also cannot be moved ($rand$). The purpose of the randomly moving object is to test whether CDL can learn that its motion is not caused by the robot's actions. There are also three randomly moving non-interactable markers (mkr^1, mkr^2, and mkr^3) on the table, as distractors to introduce potential spurious correlations. 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.

4.2. Implementation

4.2.1. BASELINES

We compare CDL with the following baselines:
1. **Reg**: regularization-based causal dynamics learning (Wang et al., 2021), where the causal mask M is learned as trainable parameters rather than inferred from CMI.
2. **Modular**: using a separate network to predict each state variable (Ke et al., 2021), which can be represented by our method when none of the features are masked.
3. **GNN**: a graph neural network, following (Ke et al., 2021), we consider the C-SWM model (Kipf et al., 2019) which assumes dense dependence for each state variable.

4. **Monolithic**: a multi-layer perceptron (MLP) network that takes all state variables and actions as inputs and predicts the next step values of all state variables, which is commonly used in MBRL.

For a fair comparison, we use the same architecture for CDL, Reg, and Modular. For GNN and Monolithic whose architectures are different from the others, we try our best to give them the same number of parameters as others. See architectures of all methods in appendix Sec. C.2.1.

4.2.2. Causal Dynamics Learning

We use MLP networks for feature extractors $f^a_{1:d_\mathcal{S}}, f^{1:d_\mathcal{S}}_{1:d_\mathcal{S}}$ and predictive networks $q_{1:d_\mathcal{S}}$. For continuous state variable s^j_{t+1}, q_j outputs the mean and standard deviation of a normal distribution. For discrete state variable, q_j outputs the logits of a categorical distribution. The threshold to infer causal relationships from CMI varies across environments. We specify it along with other hyperparameters in appendix Sec. C.2.2. Empirically, we find that reducing the frequency of optimizing causal prediction likelihood (i.e., $\log \hat{p}(s^j_{t+1}|\mathbf{PA}_{s^j})$ in Eq. 3) can stabilize training, so we only include this term once every 10 updates of the dynamics parameter θ according to Eq. 3 .

4.2.3. Transition Collection Policy

For the simple chemical environment, we use a random policy to collect transitions as it is enough to cover most of the state-action pairs. For the manipulation environment, we use Proximal Policy Optimization (PPO, Schulman et al. (2017)) as our RL agent. Meanwhile, as exploring the whole state space for manipulation domains is still a challenging research topic and not the focus of this paper, to reduce the complexity of exploration, we follow Nasiriany et al. (2021) and modify PPO's action space to long-horizon skills, such as position reaching and object grasping/pushing (details in appendix Sec. C.2.3). Notice that dynamics learning still uses low-level raw motor actions described in Sec. 4.1.

4.3. Causal Dynamics Learning Evaluation

After training all methods on the same set of collected transition data (details in appendix Sec. C.2.3), we evaluate the quality of their learned dynamics in the following aspects:

Causal Relationship Inference. We compare the causal graph inferred by CDL and Reg with the ground truth graph, in terms of edge accuracy.

Predicting Future States. Given the current state and a sequence of actions, we evaluate the accuracy of each method's k-step prediction, for states both in and out of the training distribution (denoted as ID and OOD states). For the chemical environment, we evaluate color prediction in terms of mean accuracy (MA). For manipulation, log-likelihoods is used.

We also provide results using other metrics in the appendix Sec. C.3.

Table 1. Causal Graph Accuracy (in %) for CDL and Reg

Environment	CDL (Ours)	Reg
Chemical (Collider)	**100.0** $\pm$ 0.0	99.4 $\pm$ 0.4
Chemical (Chain)	**100.0** $\pm$ 0.1	99.7 $\pm$ 0.1
Chemical (Full)	**99.1** $\pm$ 0.1	97.7 $\pm$ 0.4
Manipulation	**90.2** $\pm$ 0.3	84.4 $\pm$ 0.5

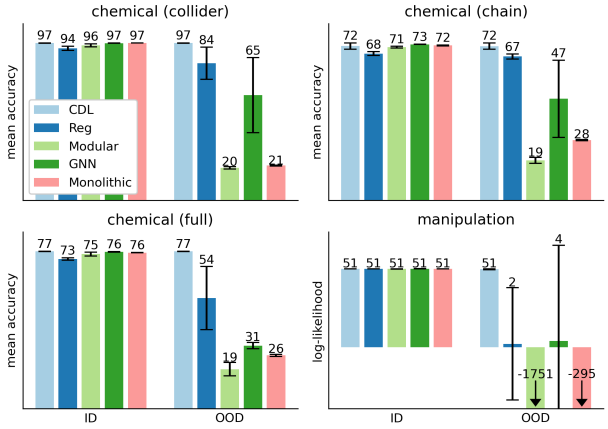

Figure 6. 1-step prediction performance on ID and OOD states. The mean is marked on top of each bar.

In the remainder of the paper, the performances shown for each method are mean and standard deviation computed from 3 models with different seeds.

4.3.1. Causal Relationship Inference

As shown in Table. 1, in all environments, CDL learns the causal graph more accurately than Reg. Other baselines are not compared as they do not learn causal relationships. Fig. 5 right shows the causal graph learned by CDL for the manipulation environment in the object level, and other quantitative and qualitative results can be found in appendix Sec. C.3.1 which collectively indicate that CDL learns both fewer false positives (i.e., spurious correlations) and false negatives (i.e., missing necessary dependencies).

4.3.2. Predicting Future States

We evaluate each method for $1 \sim 5$-step prediction on 5K transitions, for both ID and OOD states. To create OOD states, we change object positions in the chemical environment and marker positions in the manipulation environment to unseen values, and the prediction is measured on other unchanged state variables. The 1-step prediction performance for each method is shown in Fig. 6 (the Modular and Monolithic in the bottom right plot are cropped due to their bad performances, and their means are labeled instead), and results for $2 \sim 5$-steps are in appendix Sec.

Table 2. Causal Graph Accuracy for CDL with Different Transition Collection Policies

Method	PredDiff (Ours)	Uniform	Curiosity
Accuracy (%)	**90.2** $\pm$ 0.3	89.1 $\pm$ 0.6	88.9 $\pm$ 0.2

C.3.2. For ID states, all methods show similar prediction performance. However, for OOD states, the performance of dense models drops significantly as they suffer from spurious correlations while CDL and Reg mask those out with the learned dependencies. However, Reg also performs badly in the manipulation domain, because the trade-off between prediction accuracy and regularization leads to an inaccurate causal graph.

4.4. Transition Collection Policy Learning Evaluation

To evaluate the benefits of the proposed transition collection policy (denoted as **PredDiff**), we compare its performance with the following methods, in terms of causal graph accuracy in the manipulation environment,
Uniform: a fixed policy that uniformly selects between skills and skill parameters instead of active exploration;
Curiosity: a policy that uses the prediction error of the causal predictor $\hat{p}(s^j_{t+1}|\mathbf{PA}_{s^j})$ as rewards which encourages transitions where prediction errors are high (Pathak et al., 2017);

Table 4.4 shows the accuracies of the causal graphs learned by CDL from the same amount of data (1.8M transitions) but collected by those three different policies, measured in the same way as Sec. 4.3.1. Our PredDiff policy learns the graph with the highest accuracy. Though the accuracy differences between methods seem small, in a causal graph representing complex rigid body dynamics with 23×24 potential dependencies, 1% accuracy difference means prediction errors for $5 \sim 6$ edges, which may severely harm the generalizability of the learned causal dynamics model. Furthermore, as shown in appendix Sec. C.4, causal graphs learned by Uniform have more spurious correlations as it doesn't actively explore transitions where the causal graph is not learned well, and the Curiosity policy suffers from procrastination (i.e., repeats the transition with large stochasticity to maximize rewards) and fails to expose other causal dependencies.

4.5. Downstream Tasks Learning Evaluation

We evaluate each method with the following downstream tasks in the chemical environment (C) and in the manipulation environment (M):
Match (C): match the object colors with goal colors individually,
Reach (M): move the end-effector to the goal position,
Lift (M): lift the movable object to the goal position,
Stack (M): stack the movable object on the top of the unmovable object.

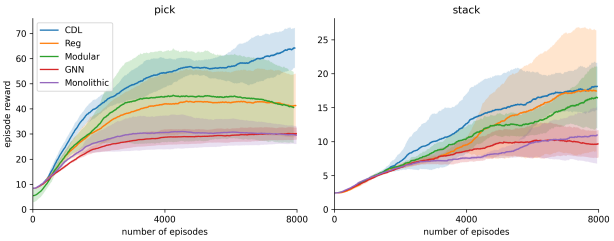

Figure 7. Training curves for two downstream tasks.

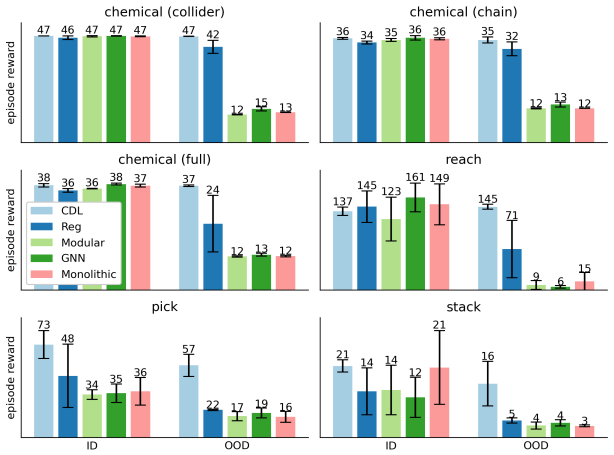

Figure 8. Episode reward of learned policies on ID and OOD states.

The tasks' reward functions are unknown to the agent and are specified in appendix Sec. C.5. During training, the dynamics model is frozen and only the reward predictor is learned as described in Sec. 3.5.

Compared to dense baselines, CDL and Reg also learn a causal graph so that they can learn tasks with the state abstraction described in Sec. 3.2. Fig. 7 shows the training curves of the Lift and Stack tasks (those of other tasks are in appendix Sec. C.5); CDL learns tasks the most sample-efficiently with the state abstraction. Though Reg also uses abstraction, it learns more slowly as some action-irrelevant state variables are also included in the abstraction because of the inaccurately learned causal graphs.

Furthermore, to test the generalization of the learned policies, we evaluate their performance for 100 episodes on both ID and OOD states, as shown in Fig. 8. Again, CDL generalizes the best across all methods on OOD states.

5. Conclusions

In this work, we introduce CDL, Causal Dynamics Learning for Task-Independent State Abstraction. In contrast to most existing MBRL methods that learn a dense dynamics model, CDL learns and, for each state variable, only keeps necessary dependencies on other variables and actions, thus generalizing better to unseen states than dense models. Furthermore, CDL derives a state abstraction from the learned

causal relationships, which can be applied to a wider range of tasks than existing state abstraction methods that only retain state variables specific to one task. Our experiments demonstrate CDL's generalization improvement both on the learned dynamics models and the policies for downstream tasks, as well as its sample efficiency resulting from its state abstraction. While this paper focuses on low-dimensional state spaces, an important future direction is to extend CDL to high-dimensional space, like images.

Acknowledgement

This work has taken place in the Learning Agents Research Group (LARG) and Robot Perception and Learning Lab at UT Austin, supported in part by NSF (CPS-1739964, IIS-1724157, NRI-1925082), ONR (N00014-18-2243, N00014-20-1-2115), FLI (RFP2-000), ARO (W911NF-19-2-0333), ARL (W911NF2020132), NASA (80NSSC19K0209), DARPA, Lockheed Martin, GM, and Bosch. Peter Stone serves as the Executive Director of Sony AI America and receives financial compensation for this work. The terms of this arrangement have been reviewed and approved by the University of Texas at Austin in accordance with its policy on objectivity in research.

References

Amos, B., Stanton, S., Yarats, D., and Wilson, A. G. On the model-based stochastic value gradient for continuous reinforcement learning. In *Learning for Dynamics and Control*, pp. 6–20. PMLR, 2021.

Buesing, L., Weber, T., Zwols, Y., Racaniere, S., Guez, A., Lespiau, J.-B., and Heess, N. Woulda, coulda, shoulda: Counterfactually-guided policy search. *arXiv preprint arXiv:1811.06272*, 2018.

Chapman, D. and Kaelbling, L. P. Input generalization in delayed reinforcement learning: An algorithm and performance comparisons. In *IJCAI*, volume 91, pp. 726–731. Citeseer, 1991.

Chua, K., Calandra, R., McAllister, R., and Levine, S. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *arXiv preprint arXiv:1805.12114*, 2018.

Even-Dar, E. and Mansour, Y. Approximate equivalence of markov decision processes. In *Learning Theory and Kernel Machines*, pp. 581–594. Springer, 2003.

Feinberg, V., Wan, A., Stoica, I., Jordan, M. I., Gonzalez, J. E., and Levine, S. Model-based value estimation for efficient model-free reinforcement learning. *arXiv preprint arXiv:1803.00101*, 2018.

Fu, X., Yang, G., Agrawal, P., and Jaakkola, T. Learning task informed abstractions. In *International Conference on Machine Learning*, pp. 3480–3491. PMLR, 2021.

Goyal, A., Lamb, A., Hoffmann, J., Sodhani, S., Levine, S., Bengio, Y., and Schölkopf, B. Recurrent independent mechanisms. *arXiv preprint arXiv:1909.10893*, 2019.

Goyal, A., Lamb, A., Gampa, P., Beaudoin, P., Levine, S., Blundell, C., Bengio, Y., and Mozer, M. Object files and schemata: Factorizing declarative and procedural knowledge in dynamical systems. *arXiv preprint arXiv:2006.16225*, 2020.

Goyal, A., Didolkar, A., Ke, N. R., Blundell, C., Beaudoin, P., Heess, N., Mozer, M. C., and Bengio, Y. Neural production systems. *Advances in Neural Information Processing Systems*, 34, 2021.

Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H., and Davidson, J. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, pp. 2555–2565. PMLR, 2019.

Janner, M., Fu, J., Zhang, M., and Levine, S. When to trust your model: Model-based policy optimization. *arXiv preprint arXiv:1906.08253*, 2019.

Jong, N. K. and Stone, P. State abstraction discovery from irrelevant state variables. In *IJCAI*, volume 8, pp. 752–757. Citeseer, 2005.

Ke, N. R., Didolkar, A. R., Mittal, S., Goyal, A., Lajoie, G., Bauer, S., Rezende, D. J., Mozer, M. C., Bengio, Y., and Pal, C. Systematic evaluation of causal discovery in visual model based reinforcement learning. 2021.

Kipf, T., van der Pol, E., and Welling, M. Contrastive learning of structured world models. *arXiv preprint arXiv:1911.12247*, 2019.

Kurutach, T., Clavera, I., Duan, Y., Tamar, A., and Abbeel, P. Model-ensemble trust-region policy optimization. *arXiv preprint arXiv:1802.10592*, 2018.

Li, A. C., Florensa, C., Clavera, I., and Abbeel, P. Sub-policy adaptation for hierarchical reinforcement learning. *arXiv preprint arXiv:1906.05862*, 2019.

Li, L. *A unifying framework for computational reinforcement learning theory*. Rutgers The State University of New Jersey-New Brunswick, 2009.

Li, Y., Torralba, A., Anandkumar, A., Fox, D., and Garg, A. Causal discovery in physical systems from videos. *arXiv preprint arXiv:2007.00631*, 2020.

Lu, C., Huang, B., Wang, K., Hernández-Lobato, J. M., Zhang, K., and Schölkopf, B. Sample-efficient reinforcement learning via counterfactual-based data augmentation. *arXiv preprint arXiv:2012.09092*, 2020.

Lyle, C., Zhang, A., Jiang, M., Pineau, J., and Gal, Y. Resolving causal confusion in reinforcement learning via robust exploration. In *Self-Supervision for Reinforcement Learning Workshop-ICLR 2021*, 2021.

Mastakouri, A. A., Schölkopf, B., and Janzing, D. Necessary and sufficient conditions for causal feature selection in time series with latent common causes. In *International Conference on Machine Learning*, pp. 7502–7511. PMLR, 2021.

McCallum, A. K. *Reinforcement learning with selective perception and hidden state*. University of Rochester, 1996.

Mozifian, M., Zhang, A., Pineau, J., and Meger, D. Intervention design for effective sim2real transfer. *arXiv preprint arXiv:2012.02055*, 2020.

Nagabandi, A., Kahn, G., Fearing, R. S., and Levine, S. Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7559–7566. IEEE, 2018.

Nair, S., Zhu, Y., Savarese, S., and Fei-Fei, L. Causal induction from visual observations for goal directed tasks. *arXiv preprint arXiv:1910.01751*, 2019.

Nasiriany, S., Liu, H., and Zhu, Y. Augmenting reinforcement learning with behavior primitives for diverse manipulation tasks. In *arXiv preprint arXiv:2110.03655*, 2021.

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pp. 2778–2787. PMLR, 2017.

Pearl, J. *Causality*. Cambridge university press, 2009.

Ravindran, B. *An algebraic approach to abstraction in reinforcement learning*. University of Massachusetts Amherst, 2004.

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

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Seitzer, M., Schölkopf, B., and Martius, G. Causal influence detection for improving efficiency in reinforcement learning. *arXiv preprint arXiv:2106.03443*, 2021.

Sontakke, S. A., Mehrjou, A., Itti, L., and Schölkopf, B. Causal curiosity: Rl agents discovering self-supervised experiments for causal representation learning. In *International Conference on Machine Learning*, pp. 9848–9858. PMLR, 2021.

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. *Causation, prediction, and search*. MIT press, 2000.

Tomar, M., Zhang, A., Calandra, R., Taylor, M. E., and Pineau, J. Model-invariant state abstractions for model-based reinforcement learning. *arXiv preprint arXiv:2102.09850*, 2021.

Volodin, S., Wichers, N., and Nixon, J. Resolving spurious correlations in causal models of environments via interventions. *arXiv preprint arXiv:2002.05217*, 2020.

Wang, Z., Xiao, X., Zhu, Y., and Stone, P. Task-independent causal state abstraction. In *Proceedings of the 35th International Conference on Neural Information Processing Systems, Robot Learning workshop*, 2021.

Williams, G., Wagener, N., Goldfain, B., Drews, P., Rehg, J. M., Boots, B., and Theodorou, E. A. Information theoretic mpc for model-based reinforcement learning. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1714–1721. IEEE, 2017.

Zhang, A., Lipton, Z. C., Pineda, L., Azizzadenesheli, K., Anandkumar, A., Itti, L., Pineau, J., and Furlanello, T. Learning causal state representations of partially observable environments. *arXiv preprint arXiv:1906.10437*, 2019.

Zhang, A., Lyle, C., Sodhani, S., Filos, A., Kwiatkowska, M., Pineau, J., Gal, Y., and Precup, D. Invariant causal prediction for block mdps. In *International Conference on Machine Learning*, pp. 11214–11224. PMLR, 2020a.

Zhang, A., McAllister, R., Calandra, R., Gal, Y., and Levine, S. Learning invariant representations for reinforcement learning without reconstruction. *arXiv preprint arXiv:2006.10742*, 2020b.

Zhu, Y., Wong, J., Mandlekar, A., and Martín-Martín, R. robosuite: A modular simulation framework and benchmark for robot learning. In *arXiv preprint arXiv:2009.12293*, 2020.

Appendix for the paper: Causal Dynamics Learning for Task-Independent State Abstraction

A. Causal Dynamics Learning

A.1. Definitions

Here we provide fundamental definitions in the causal inference that we use for proving Theorem. 3.1. For a thorough study see Pearl (2009).

Definition A.1. (d-separation (Pearl, 2009)) In a directed acyclic graph (DAG) $\mathcal{G}$, a path between nodes I_1 and I_m is blocked by a set S (with neither I_1 nor I_m in S) whenever there is a node $I_k, k = 2, \dots, m-1$, such that one of the following two possibilities holds:
(i) $I_k \in S$ and $I_{k-1} \to I_k \to I_{k+1}$ or $I_{k-1} \leftarrow I_k \leftarrow I_{k+1}$ or $I_{k-1} \leftarrow I_k \to I_{k+1}$.
(ii) Neither I_k nor any of its descendants is in S and $I_{k-1} \to I_k \leftarrow I_{k+1}$.

In a DAG $\mathcal{G}$, we say that two nodes A and B are d-separated by a third node C if every path between nodes A and B is blocked by C, denoted as $A \perp\!\!\!\perp_{\mathcal{G}} B|C$.

Definition A.2. (Causal Markov Condition (Spirtes et al., 2000)) Let $\mathcal{G}$ be a causal graph with vertex set V and p be a probability distribution over the vertices in V generated by the causal structure represented by $\mathcal{G}$. $\mathcal{G}$ and p satisfy the Causal Markov Condition if and only if for every node v in V, v is independent of $V \setminus (\text{Descendants}(v) \cup \text{Parents}(v))$ given Parents(v).

Here we use the global version of the Markov condition, which reads: if $A \perp\!\!\!\perp_{\mathcal{G}} B|C \Rightarrow A \perp\!\!\!\perp B|C$ for all disjoint vertex sets A, B, C (where $\perp\!\!\!\perp_{\mathcal{G}}$ denotes d-separation, as defined above, and $\perp\!\!\!\perp$ denote independence in the probability).

Definition A.3. (Causal Faithfulness). A distribution p is faithful to a DAG $\mathcal{G}$ if no conditional independence relations other than the ones entailed by the Markov property are present.

A.2. Proof of Theorem. 3.1

We need to show that if $s_t^i \not\to s_{t+1}^j$, then the condition $s_t^i \not\perp\!\!\!\perp s_{t+1}^j | \{a_t, s_t \setminus s_t^i\}$ or the assumption is violated.

Assume there is no direct edge from s_t^i to s_{t+1}^j but they are dependent, i.e., $s_t^i \not\to s_{t+1}^j$ and $s_t^i \not\perp\!\!\!\perp s_{t+1}^j$, then there is a confounder $Q_{t'}$ with the confounding path $s_t^i \dashleftarrow Q_{t'} \dashrightarrow s_{t+1}^j$ where $t' < t$ as we assume there is no simultaneous edge. There are two cases for $Q_{t'}$:
(1) If $Q_{t'}$ is observed, i.e., $Q_{t'}$ is one of the state variable, then it violates the condition because we condition on $\{a_t, s_t \setminus s_t^i\}$ which block all possible paths for $Q_{t'} \dashrightarrow s_{t+1}^j$ and thus d-separate s_t^i and s_{t+1}^j.
(2) If $Q_{t'}$ is unobserved, then it must be memoryless (i.e., $Q_t \not\to Q_{t+1}$) to adhere to A2 (the state is fully observed and the dynamics is Markovian). In that case, there exists such a path $s_{t-1}^i \to s_t^i \dashleftarrow Q_{t'} \dashrightarrow s_{t+1}^j$ given A3 ($s_t^i \to s_{t+1}^i$ exists for all state variables s^i). This path suggests $s_{t-1}^i \not\perp\!\!\!\perp s_{t+1}^j | s_t^i$ where s_t^i serves as the collider connecting s_{t-1}^i and s_{t+1}^j. If we further condition the dependence on all other state variables at t to block all potential paths $s_{t-1}^i \to s_t^k \to s_{t+1}^j$, we have $s_{t-1}^i \not\perp\!\!\!\perp s_{t+1}^j | s_t$. However this conditional dependence violates A2, i.e., Markovian property of the dynamics $s_{t-1} \perp\!\!\!\perp s_{t+1} | s_t$.

In both cases either the condition or A2 is violated. Therefore we show that if $s_t^i \not\perp\!\!\!\perp s_{t+1}^j | \{a_t, s_t \setminus s_t^i\}$, then $s_t^i \to s_{t+1}^j$. □

The proof of that if $a_t \not\perp\!\!\!\perp s_{t+1}^j | s_t$, then $a_t \to s_{t+1}^j$ is even simpler: as a_t has no parent in the causal graph, there cannot exist the confounding path $a_t \dashleftarrow Q_{t'} \dashrightarrow s_{t+1}^j$. Then following A1 (causal faithfulness), conditioning on all other potential parents of s_{t+1}^j (i.e., s_t), the conditional dependence $a_t \not\perp\!\!\!\perp s_{t+1}^j | s_t$ must result from the existence of the edge $a_t \to s_{t+1}^j$. □

Table 3. prediction accuracy (%) in the chemical collider environment

Method	**CDL** (Ours)	**NPS**
in-distribution states	**97.3**	88.0
OOD states	**97.3**	35.7

A.3. Conditional Mutual Information

We provide the full derivation of Eq. 2 as follows:

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{s_t,a_t,s^j_{t+1}}\left[\log \frac{p(s^i_t, s^j_{t+1}|\{a_t, s_t \setminus s^i_t\})}{p(s^i_t|\{a_t, s_t \setminus s^i_t\})\, p(s^j_{t+1}|\{a_t, s_t \setminus s^i_t\})}\right] \\
=&\mathop{\mathbb{E}}_{s_t,a_t,s^j_{t+1}}\left[\log \frac{p(s^i_t|\{a_t, s_t \setminus s^i_t\})\, p(s^j_{t+1}|\{a_t, s_t\})}{p(s^i_t|\{a_t, s_t \setminus s^i_t\})\, p(s^j_{t+1}|\{a_t, s_t \setminus s^i_t\})}\right] && \text{// by expanding } p(s^i_t, s^j_{t+1}|\{a_t, s_t \setminus s^i_t\}) \\
=&\mathop{\mathbb{E}}_{s_t,a_t,s^j_{t+1}}\left[\log \frac{p(s^j_{t+1}|a_t, s_t)}{p(s^j_{t+1}|\{a_t, s_t \setminus s^i_t\})}\right]. && \text{// by cancelling out } p(s^i_t|\{a_t, s_t \setminus s^i_t\})
\end{aligned}
$$

B. Additional Related Work

In this section, we given a more detailed comparison between our work and methods that consider explicit sparsity and modularity, e.g., Neural Production Systems (NPS, Goyal et al. (2021)) and Object Files and Schemata (SCOFF, Goyal et al. (2020)). Specifically:

1. **Sparsity**: Compared to our method in which each variable depends on all of its causal parents, NPS constrains each variable (i.e., "slot" in NPS) to only depend on one parent. However, in many problems, the dynamics of one variable does depend on multiple variables, e.g., a hammer handed from one robot to the other robot, or the chemical environments in the paper with the collider graph. In those cases, NPS does not perform well due to its one-parent constraint, as shown in Table. 3. SCOFF assumes dense dependencies, so we leave it out of the sparsity discussion.

2. **Modularity**: NPS and SCOFF consider modular dynamics, i.e., multiple variables can share the same dynamics function, while our dynamics model does not permit such modularization. Extending the model accordingly is a good direction for future work.

3. **Global vs local dependencies**: Our method considers global dependencies that are static across all state values. In contrast, NPS's and SCOFF's local dependencies depend on the value of the state.
Cons of local dependencies: Since the selection of causal parent depends on the value of state variables, out-of-distribution (OOD) variables may be assigned as the wrong parent, resulting in inaccurate prediction, as shown by OOD state evaluation in Table. 3. In contrast, in our method, dependencies are independent of variable values, and thus the prediction of variables will not be affected by OOD variables that are not their parents.
Pros of local dependencies: They may induce an even sparser dynamics. For example, globally, an object depends on the robot hand which can move it, but, locally, it only depends on the hand when being grasped by the hand and is independent otherwise. As such, another promising future direction is to incorporate locality into global dependencies for further sparsity.

C. Experiment Details

C.1. Environment Details

C.1.1. Chemical Environment

In the chemical environment modified from (Ke et al., 2021), there are 10 objects and each of them has its different x, y position (two continuous variables) and color (a categorical variable out of 5 possible colors), forming 30 state variables in total. At each step, the x, y positions of objects are sampled from a normal distribution $\mathcal{N}(0, 1e^{-4})$. Meanwhile, the action changes a chosen object to a specified new color (a categorical variable out of 50 possible actions). After the chosen object changes its color, all its decadents in the underlying graph will change their colors in the breadth-first order. When changing to its new color, each descendant follows a pre-defined conditional probability table modeled as a randomly initialized multi-layer perceptron (MLP). The MLP takes the newest color of each parent as inputs and outputs a categorical

distribution over 5 colors out of which one color is sampled for the object. For example, in the chain graph shown in Fig. 4 (a) top, if the color of the leftmost object is changed by the action, then the second left object will change its color given the new color of the leftmost object changes to, then the third object changes, etc.

In the ground truth causal graph, each object's x position or y position only depends on itself and thus is action-irrelevant, while the color depends on the action and the colors of its parents and thus is controllable. However, if an object only has one parent, then its color doesn't depend on the color of that parent, as whenever its color needs to change, it's either (1) because the color is directly set by the action where the parent color has no effect or; (2) because that its only parent changes the color, and this new parent color it is going to depend on is not included in s_t (s_t only includes the parent color before the change). As a result, there is no dependence from the parent color at t to its new color at $t+1$.

C.1.2. MANIPULATION ENVIRONMENT

In the manipulation environment, the x-axis is along the width of the table, the y-axis is along the length, and z-axis is along the height. If we set the center of the table top as $(x, y, z) = (0, 0, 0)$, at each episode, the positions of all objects are uniformly sampled from $(x, y) \sim \text{Uniform}([-0.3, 0.3] \times [-0.4, 0.4])$ without overlapping and markers are sampled from $(x, y) \sim \text{Uniform}([-0.35, 0.35] \times [-0.5, 0.5])$. The end-effector of the robot is constrained in a box with the range of $[-0.3, 0.3] \times [-0.4, 0.4] \times [0, 0.2]$. At each step, the position change of the randomly moving object is sampled from $\mathcal{N}(0, 4e-6)$ and the changes of markers are sampled from $\mathcal{N}(0, 1e-4)$, and they are constrained in their initialization ranges.

In the ground truth causal graph shown in Fig. 13 top left, the controllable state variables include EEF position, gripper joint angles, and movable object positions; the action-relevant state variables are positions of the unmovable object and the randomly moving object as they can block the motion of the EEF, gripper, and the movable object (when pushed or held by the gripper); and action-irrelevant state variables include positions of all non-interactable markers.

C.1.3. PHYSICAL ENVIRONMENT

In addition to the chemical and manipulation environment, we also evaluate our method in the physical environment proposed by (Ke et al., 2021). 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.

For the ground truth causal graph, though the original authors state that there are only edges from heavier objects to lighter objects, we want to point out that the graph is actually a dense graph that also includes edges from lighter objects to heavier objects, because lighter objects can also block the movement of heavier objects if they are at the boundary or two lighter ones block one heavier object together, as described above. However, in the physical environment, as all state variables are controllable and each state variable depends on all other variables, causal dynamics learning would learn a dense model and thus has no advantages over dense dynamic models.

C.2. Implementation Details

C.2.1. ARCHITECTURES

We listed the dynamics model architectures of our method (CDL) and all baselines in Table. 4 (note that CDL, Reg, and Modular share the same architecture) and hyperparameters shared across methods in Table. 5. Meanwhile, the same architecture and hyperparameters are used for the chemical environment with different underlying graphs. For the MLP network, a list is used to represent its architecture, e.g., [64, 128] represents an MLP of 2 hidden layers with 64 neurons in the first layer and 128 neurons in the second. Notice that the input layer and the output layer of the MLP are not listed, so an empty list [] represents a linear mapping from inputs to outputs. For all activation functions, ReLU is used.

Table 4. Architectures for CDL and Baselines in All Environments

Architectures		**Environments**		
Methods	Modules	Chemical	Manipulation	Physical
CDL, **Reg**, **Modular**	feature extractors, $\{f_j^a, f_j^{1:d_\mathcal{S}}\}_{j=1}^{d_\mathcal{S}}$	[]	[]	[]
	feature dimension, $h_{1:d_\mathcal{S}}$	64	128	128
	predictive networks, $q_{1:d_\mathcal{S}}$	[64, 32]	[128, 64]	[128, 128]
GNN	embedders (map state variables to node attributes)	[]	[]	[]
	node attribute dimension	256	256	256
	edge attribute dimension	256	256	256
	node networks	[256, 256]	[256, 256]	[256, 256]
	edge networks	[256, 256]	[256, 256]	[256, 256]
	projects (map node attributes to state variables)	[]	[]	[]
Monolithic	MLP network	[256, 256, 256]	[512, 512, 512]	[256, 256, 256]

Table 5. Hyperparameters Shared across All Methods and Environments if not Specified

Name	**Environments**		
	Chemical	Manipulation	Physical
number of transitions	100K	28.8M	2M
training step	500K	1.8M	2M
optimizer		Adam	
learning rate		3e-4	
batch size		32	
prediction step during training, H		3	

C.2.2. Dynamics Learning Implementation Details

For all methods, during training, instead of optimizing 1-step prediction loss, we roll out the model for H steps and optimize the sum of H-step prediction loss. Also, we keep 10% of the data as the validation data to select the best model for each method. Moreover, CDL also evaluate the conditional mutual information (CMI) using the validation data rather than training data for more accurate measurements. For Reg whose performance is sensitive to the regularization coefficient, we conduct a grid search for the best coefficients in terms of the causal graph accuracy, and its value for each environment is listed in Table. 6.

During CDL training, for the continuous state variable where CDL predicts its next value as the mean and standard deviation of a normal distribution, we find that clipping the range of standard deviation between $[0.001, 0.01]$ reduces overfitting and apply it to other baselines as well. For CMI evaluation, as it's expensive to evaluate it using all validation data whenever the dynamics model is updated, we instead sample a batch of transitions from the validation data and evaluate CMI using it once every 10 updates of dynamics parameters θ. To get a smooth estimate of CMI while assigning greater weights to estimates from the newest model, we apply an exponential moving average to the CMI calculated from each batch. All hyperparameters of causal dynamics learning are listed in Table. 6.

C.2.3. Transition Collection Policy Implementation Details

For the chemical environment where the state-action space is simple, we just use a random policy that uniformly samples actions over 50 possible values. For the physical environment where the interactions between objects are infrequent, we use a scripted policy that iteratively selects the object to move and the object to push with the former object so that it covers all iterations between objects. For the chemical and physical environment, all methods are trained with the same transitions collected by policies mentioned earlier.

For the manipulation environment where efficient exploring is still a challenging research topic and not the focus of this paper, to reduce the complexity of exploration, we follow Nasiriany et al. (2021) and modify the action space of the RL agent to following long-horizon skills:
(1) Reach: move the end-effector to the position of a chosen object or a target position(x, y, z) . Execution takes at most 50 atomic actions.

Table 6. Hyperparameters for Dynamics Learning of CDL and Reg (Shared Across Environments if not Specified)

Method	Name	Environments				
		Chemical (collider)	Chemical (chain)	Chemical (full)	Manipulation	Physical
CDL	CMI threshold, ϵ	0.02	0.05	0.01	0.002	0.01
	$\hat{p}(s^j_{t+1}\|\mathbf{PA}_{s^j})$ optimization frequency	once every 10 updates of θ				
	CMI evaluation frequency	once every 10 updates of θ				
	CMI evaluation batch size	32				
	exponential moving average discount	0.999				
Reg	regularization coefficient	0.002	0.002	0.002	0.001	0.0
	regularization starts after N training steps	100K	100K	100K	750K	0

Table 7. Architecture and Hyperparameters for Transition Collection Policy in the Manipulation Environment

Name	Value
critic network	[128, 128]
skill selection network	[64, 64]
object selection network	[64, 64]
skill parameter network	[64, 64]
optimizer	Adam
learning rate	1e-4
batch size	32
discount	0.95
generalized advantage estimator parameter, λ	0.98
clip ratio	0.1
entropy weight	0.2

(2) Grasp: move the end-effector to the position of a chosen object or a target position (x, y, z). Execution takes at most 50 atomic actions.
(3) Lift: lift one of the objects to a target position (x, y, z). Execution takes at most 50 atomic actions.
(4) Push: select one of the objects whose position is (x, y, z) and then move the end-effector from $(x + \delta_x, y + \delta_y, z + \delta_z)$ to $(x + \Delta_x, y + \Delta_y, z + \Delta_z)$. Execution takes at most 50 atomic actions.
(5) Open: open the gripper. Execution takes at most 4 atomic actions.
(6) Atomic: apply a single atomic action, which allows the agent to fill in any gaps that cannot be fulfilled by other skills.

Notice that dynamics learning still uses low-level raw motor actions (i.e., atomic actions). To train the agent whose action space becomes low-frequency high-level skills but whose rewards are still computed from high-frequency low-level atomic actions, we adopt HiPPO (Li et al., 2019). Similar to PPO, in HiPPO, the critic uses the current state and atomic action as inputs and is trained to fit state values computed from rewards, and the actor is trained to maximize the advantage but only at timestamps when it selects skills. The actor consists of (1) a skill selection network that selects one of six skills based on s_t, (2) an object selection network that selects the object to apply the skill to based on s_t and the selected skill, and (3) a skill parameter network that outputs skill parameter (e.g., goal position (x, y, z)) based on s_t, the selected skill and the selected object. If the selected skill does not apply to an object or does not need parameters, the selected object or skill parameters will be ignored. The architecture and hyperparameters of HiPPO are listed in Table. 7. For the manipulation environment, for a fair comparison, all methods are trained with the same transitions that the HiPPO agent collects for CDL training.

C.3. Causal Dynamics Learning Evaluation Details

Apart from the results shown in Sec. 4.3, we also evaluate the dynamics model learned by CDL and baselines with other metrics and experiments which we show their results in this section.

C.3.1. Causal Relationship Inference Details

For CDL, after training the dynamics model and evaluating CMI, we derive the causal graph by examining whether CMI $\geq \epsilon$ for each edge. The threshold ϵ is an important hyperparameter that determines the accuracy of the derived causal graph and

Table 8. Performance on Causal Graph Learning for CDL and Reg on All Environments

Metrics	Method	Environments				
		Chemical (collider)	Chemical (chain)	Chemical (full)	Manipulation	Physical
Accuracy	CDL	**1.000** $\pm$ 0.000	**1.000** $\pm$ 0.001	**0.991** $\pm$ 0.001	**0.902** $\pm$ 0.003	1.00 $\pm$ 0.000
	Reg	0.994 $\pm$ 0.004	0.997 $\pm$ 0.001	0.977 $\pm$ 0.004	0.844 $\pm$ 0.005	1.00 $\pm$ 0.000
Recall	CDL	**1.000** $\pm$ 0.000	**0.992** $\pm$ 0.012	**0.917** $\pm$ 0.016	**0.612** $\pm$ 0.009	1.00 $\pm$ 0.000
	Reg	0.891 $\pm$ 0.086	0.942 $\pm$ 0.012	0.794 $\pm$ 0.006	0.459 $\pm$ 0.048	1.00 $\pm$ 0.000
Precision	CDL	1.000 $\pm$ 0.000	1.000 $\pm$ 0.000	**0.968** $\pm$ 0.016	**0.980** $\pm$ 0.006	1.00 $\pm$ 0.000
	Reg	0.993 $\pm$ 0.010	0.991 $\pm$ 0.012	0.917 $\pm$ 0.051	0.839 $\pm$ 0.068	1.00 $\pm$ 0.000
F1 Score	CDL	**1.000** $\pm$ 0.000	**0.996** $\pm$ 0.006	**0.941** $\pm$ 0.009	**0.754** $\pm$ 0.008	1.00 $\pm$ 0.000
	Reg	0.937 $\pm$ 0.046	0.966 $\pm$ 0.006	0.850 $\pm$ 0.019	0.589 $\pm$ 0.025	1.00 $\pm$ 0.000
ROC AUC	CDL	**1.000** $\pm$ 0.000	0.996 $\pm$ 0.006	0.958 $\pm$ 0.008	0.805 $\pm$ 0.005	1.00 $\pm$ 0.000
	Reg	0.932 $\pm$ 0.042	0.990 $\pm$ 0.009	0.956 $\pm$ 0.008	**0.893** $\pm$ 0.002	1.00 $\pm$ 0.000

thus the accuracy of the dynamics model. Even though we can use the same value used for training, as the best threshold is unknown before training, that value is just roughly estimated from several runs and may not be the best threshold. As a result, to select the best threshold, we conduct a linear search and select the one that gives the best F1 score. For a fair comparison, we apply the same threshold selection method to the regularization-based causal dynamics learning (Reg): Reg learns each edge of the causal graph as an individual Bernoulli distribution where its success probability is the learnable parameter. During training, as a rule-of-thumb, we use 0.5 as the threshold when judging whether an edge exists for each sample. After training and when evaluating the causal graph based on the learned success probability, we conduct a linear search and again select the threshold that gives the best F1 score.

After deriving the causal graph for each run with its best threshold, we evaluate the performance of CDL and Reg on causal graph learning, in terms of accuracy, recall, precision, F1 score, and area under the receiver operating characteristic curve (ROC AUC). For all metrics, the higher the better. The results for different environments are shown in Table. 8. Again, CDL outperforms Reg for most of the metrics and environments, except for ROC AUC in the manipulation environments. In the physical environment, both methods learn the dense causal graph correctly and thus have values of 1.000 for all metrics.

In Fig. 11 $\sim$ 16, we also show the causal graph learned by both methods in all environments except for the physical environment as both methods learn it equally well. In each quadruplet figure, the top left one shows the ground truth causal graph of the environment which is binary, while the other three show the causal graph learned by one of the methods with three runs. In each causal graph, there 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^j_{t+1} depends on the variable s^i_t if $j < d_{\mathcal{S}} + 1$ or a_t if $j = d_{\mathcal{S}} + 1$, measured by CMI for CDL or Bernoulli success probability for Reg. As CMI and success probability are continuous values, we plot them as a heatmap where darker color represents smaller values and vice versa. Also, the color in each heatmap is capped at the threshold selected as described in Sec. C.3 so that all inferred causal edges are shown in the same white (brightest) color, and the selected threshold is marked on top of each heatmap. Finally, we mark all wrong edge predictions (i.e., false positives and false negatives) using red boxes for easier identification.

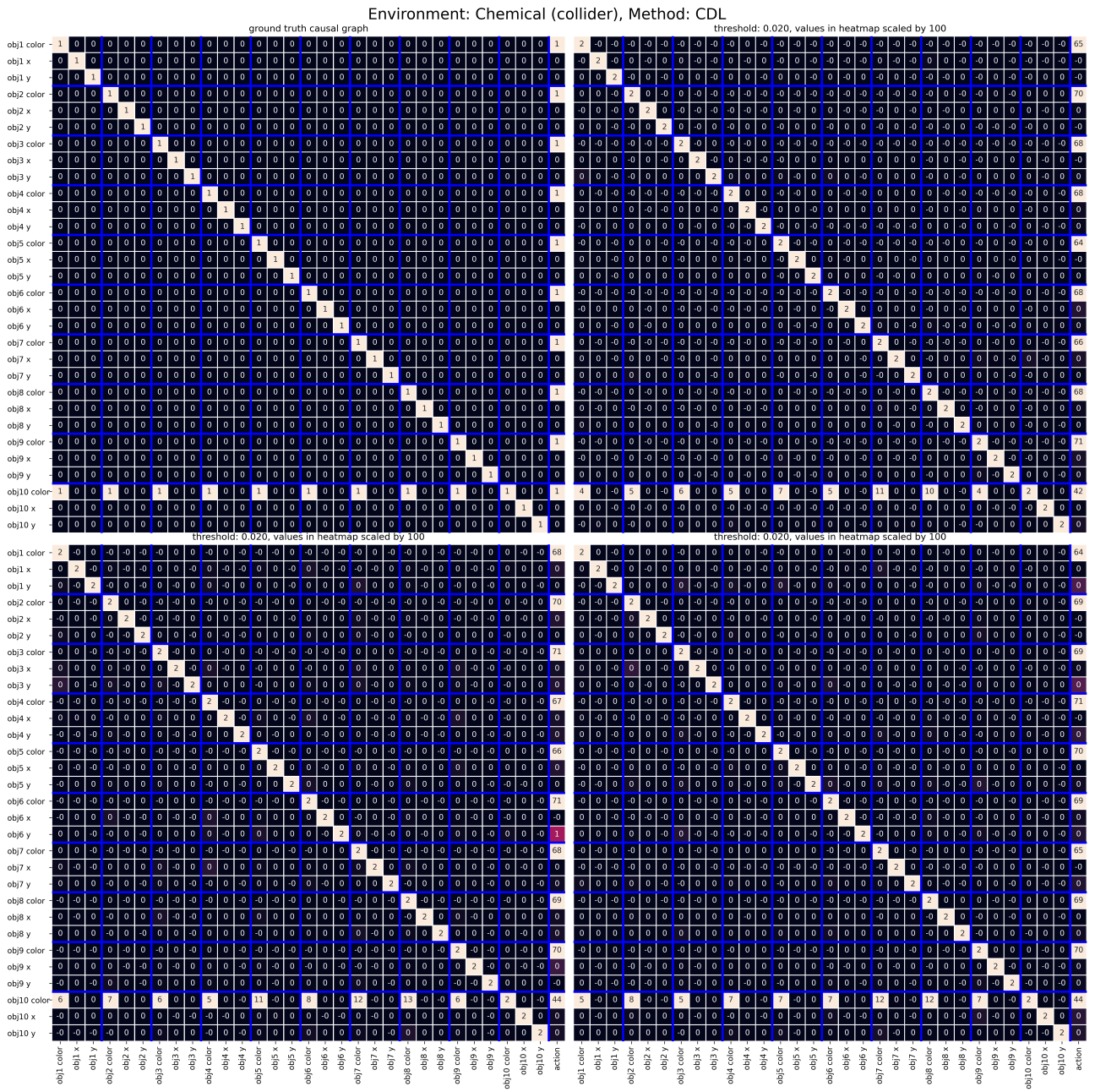

Figure 9. Causal graph for the chemical environment with the collider graph. **(top left)** the ground truth causal graph, **(others)** causal graphs learned by **CDL** with 3 different runs, where all of them learn the causal graph accurately.

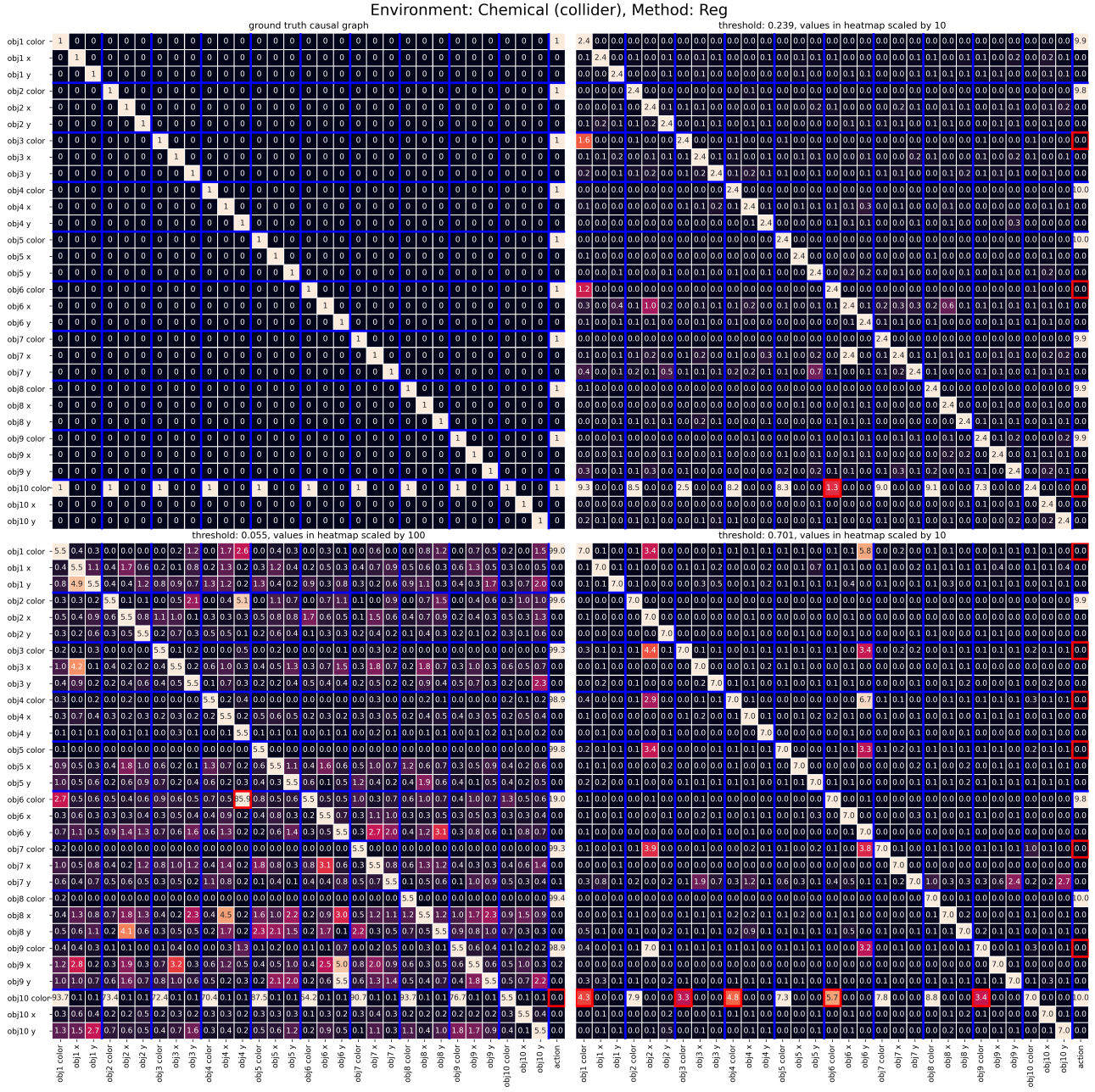

Figure 10. Causal graph for the chemical environment with the collider graph. **(top left)** the ground truth causal graph, **(others)** causal graphs learned by **Reg** with 3 different runs, where there are multiple missing dependencies (false negatives) and spurious correlations (false positives).

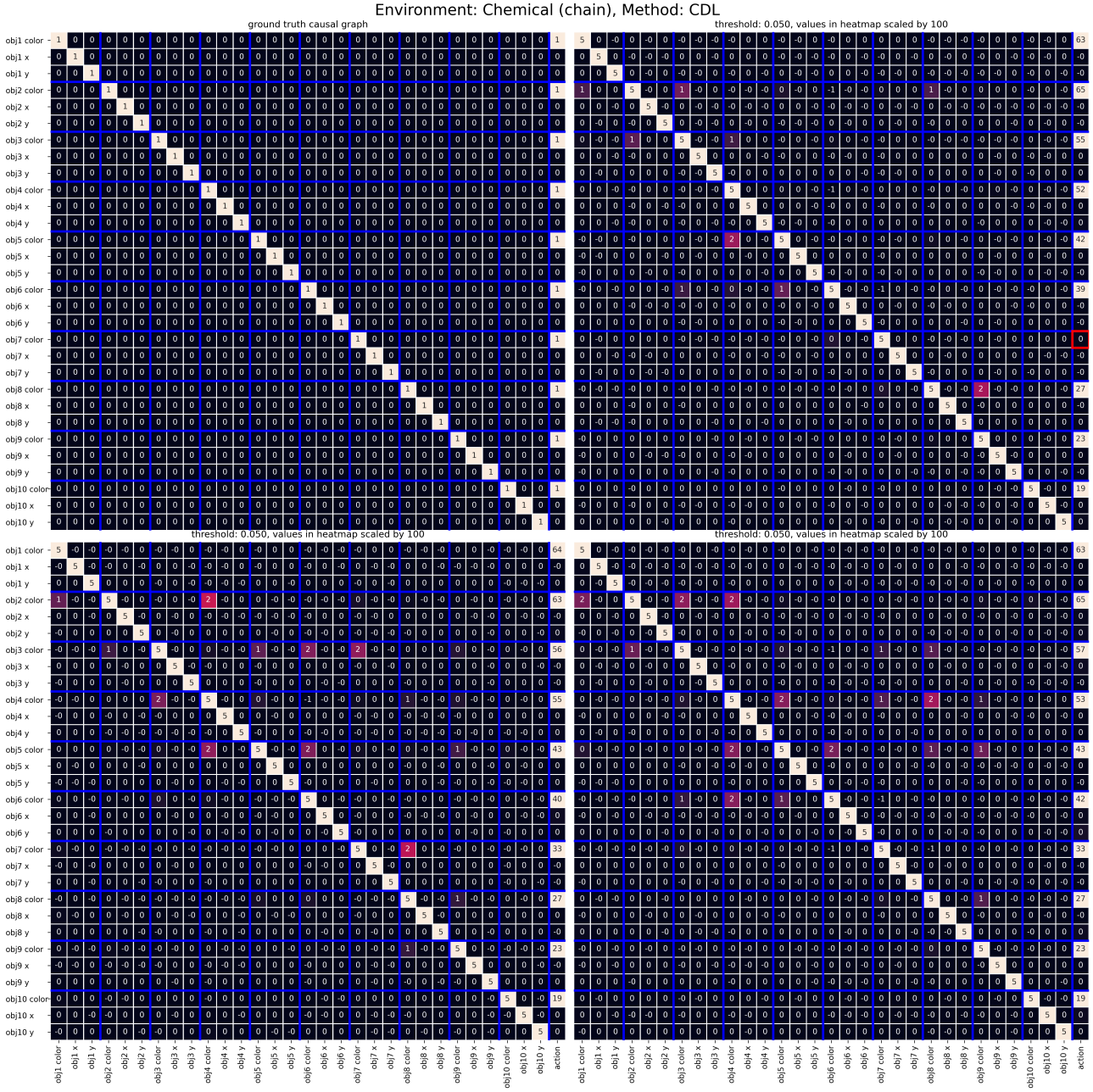

Figure 11. Causal graph for the chemical environment with the chain graph. **(top left)** the ground truth causal graph, **(others)** causal graphs learned by **CDL** with 3 different runs. Despite that the underlying graph is more challenging than the collider graph with long-term depdencies, all three runs still learn the causal graph accurately except for one missing dependency in top right.

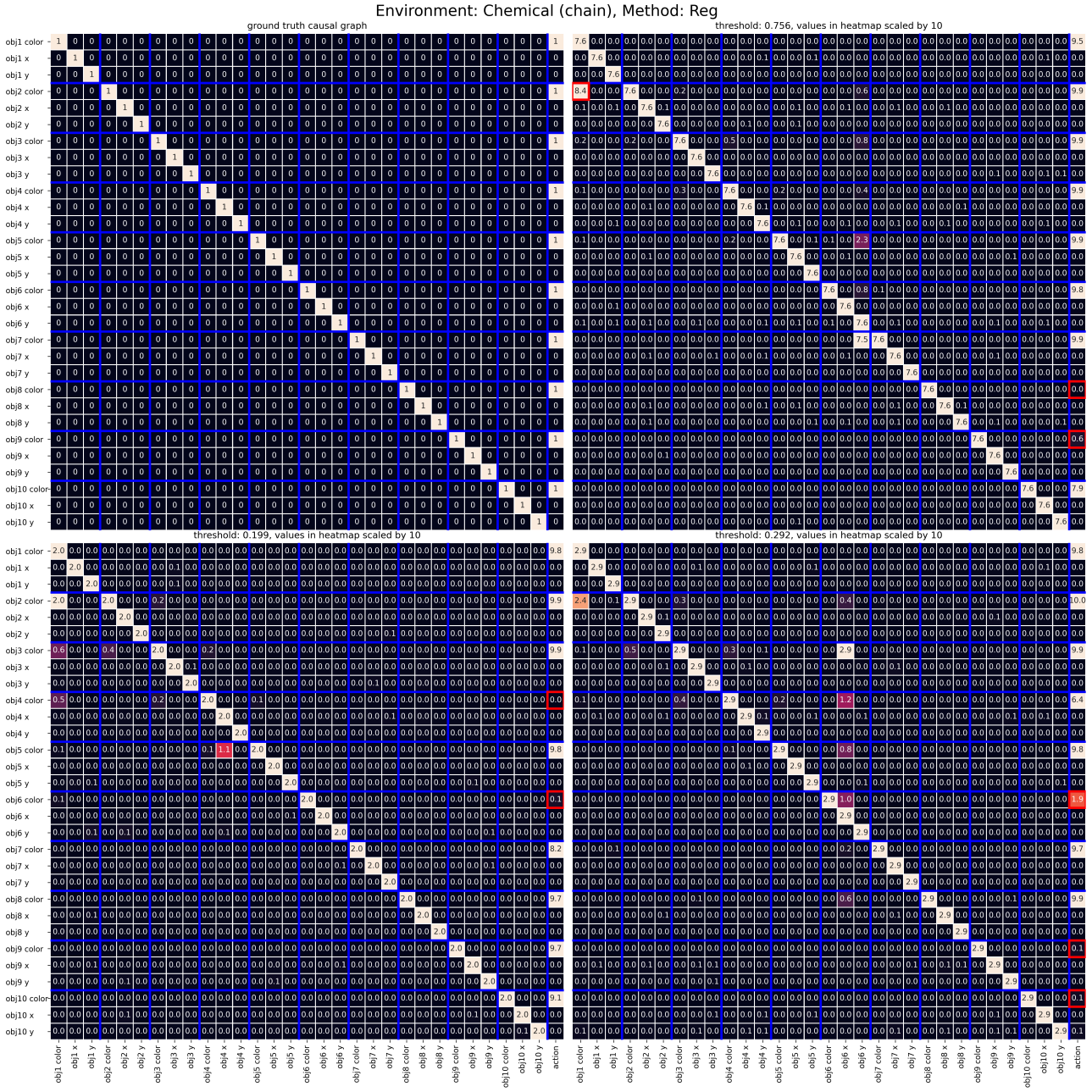

Figure 12. Causal graph for the chemical environment with the chain graph. **(top left)** the ground truth causal graph, **(others)** causal graphs learned by **Reg** with 3 different runs, where there again are multiple missing dependencies (false negatives) and spurious correlations (false positives) in all runs. Furthermore, Reg misses several dependencies on the action which are crucial to downstream task learning.

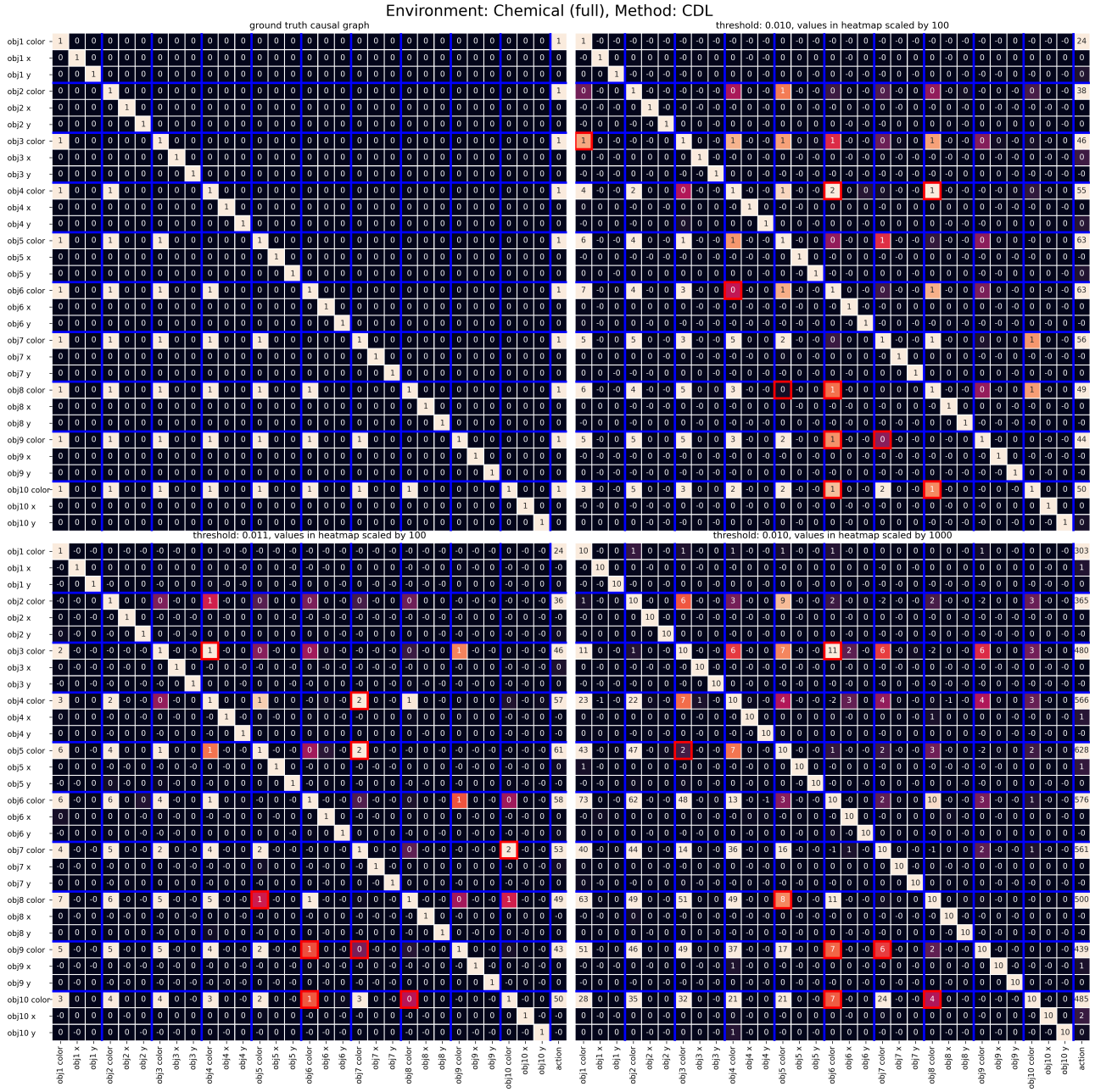

Figure 13. Causal graph for the chemical environment with the full graph. **(top left)** the ground truth causal graph, **(others)** causal graphs learned by **CDL** with 3 different runs. Among the three graphs for the chemical environment, the dense graph is the most complex one, with both long-term dependencies and much more dependencies to capture. As a result, CDL makes more prediction errors, but the learn the causal graphs are still much more accurate than those learned by Reg.

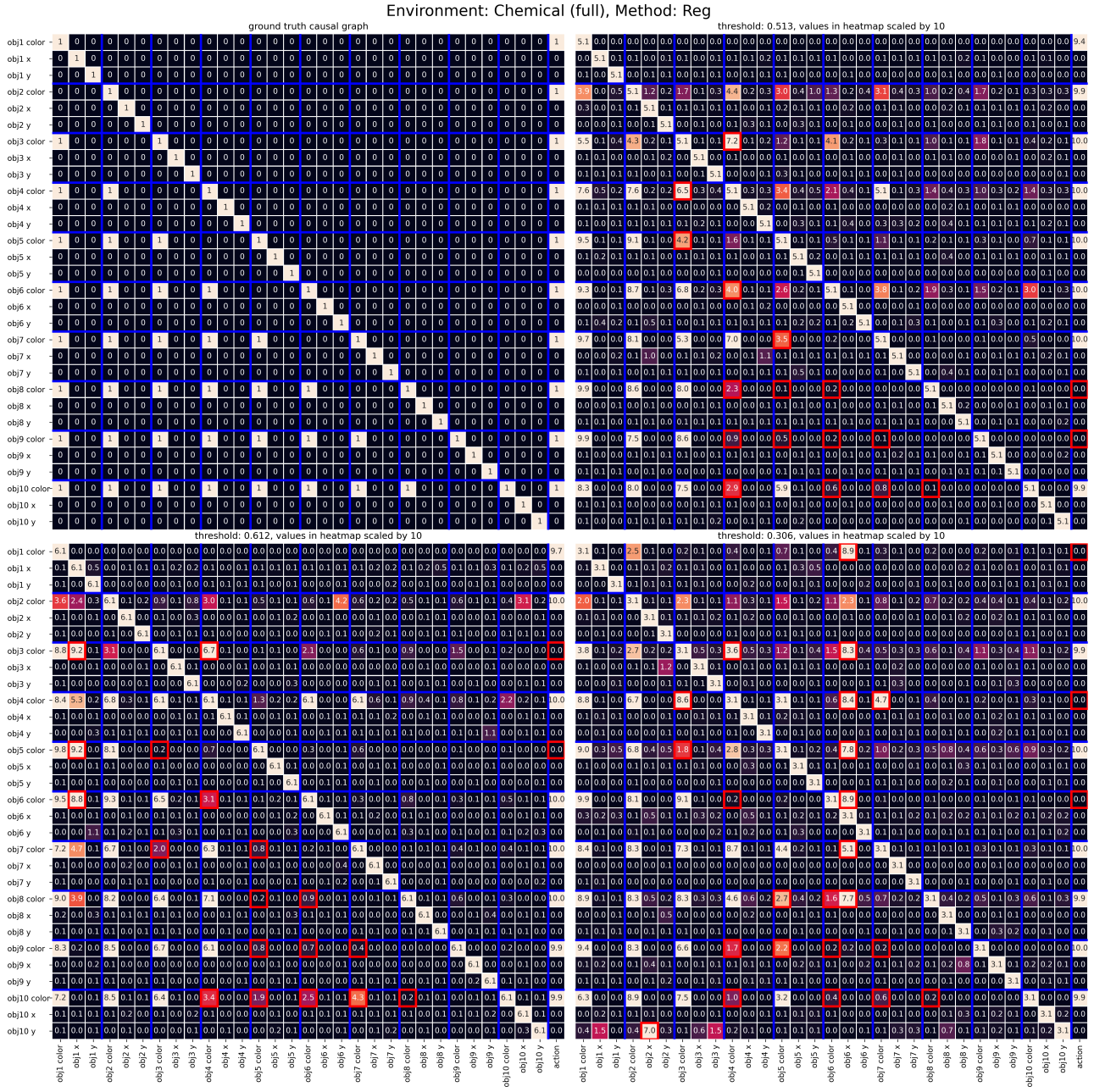

Figure 14. Causal graph for the chemical environment with the full graph. **(top left)** the ground truth causal graph, **(others)** causal graphs learned by **Reg** with 3 different runs, where there are much more missing dependencies (false negatives) and spurious correlations (false positives) than the graphs learned by CDL.

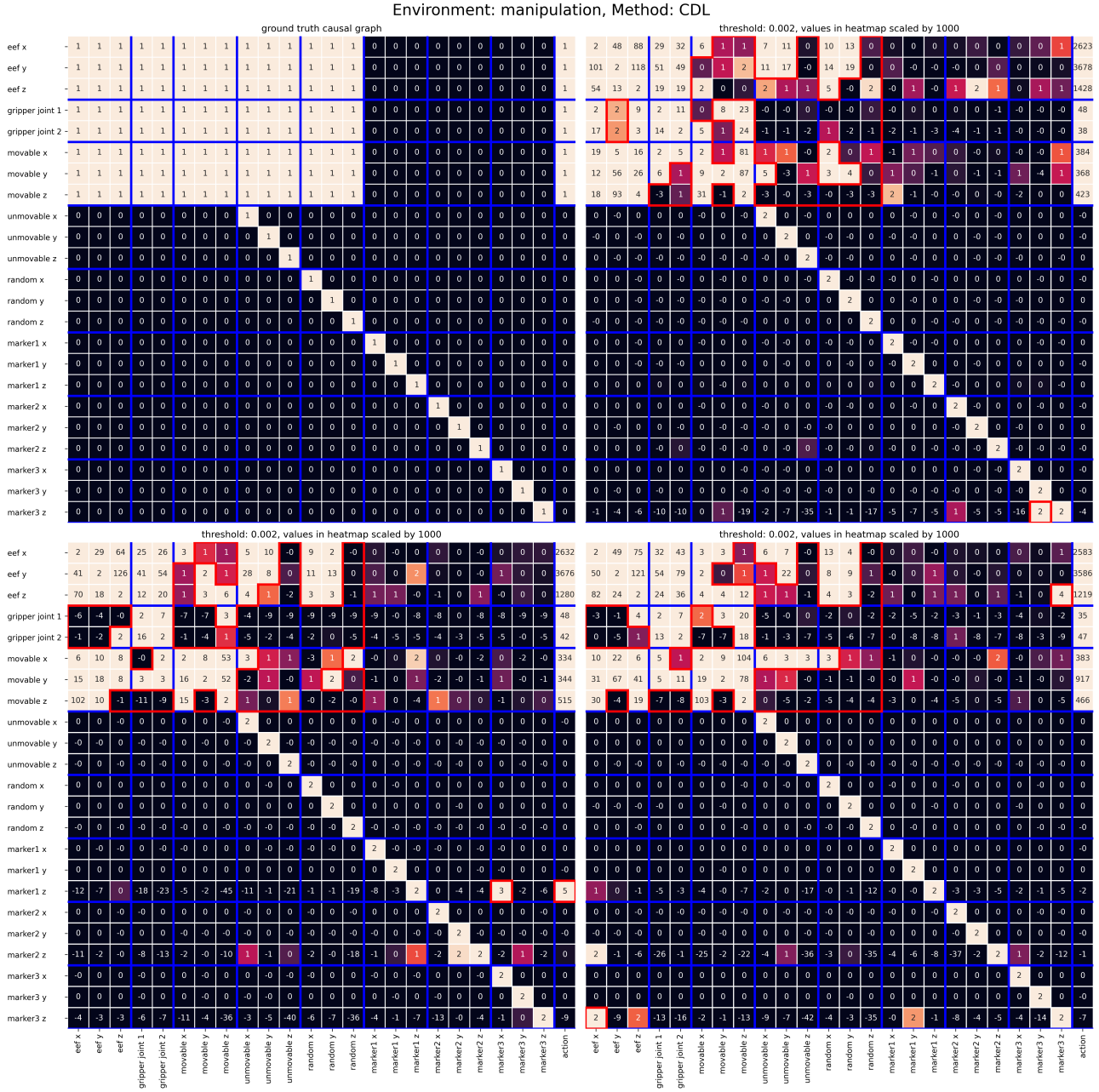

Figure 15. Causal graph for the manipulation environment. **(top left)** the ground truth causal graph, **(others)** causal graphs learned by **CDL** with 3 different runs. In all runs, though CDL successfully learns the controllable variables: eef, gripper, and movable object, it only captures part of their dependencies on the unmovable and the randomly moving object as the interactions between them are infrequent. Also, it gives one or two false positives in each run, but it still learns much cleaner graphs compared to those learned by Reg.

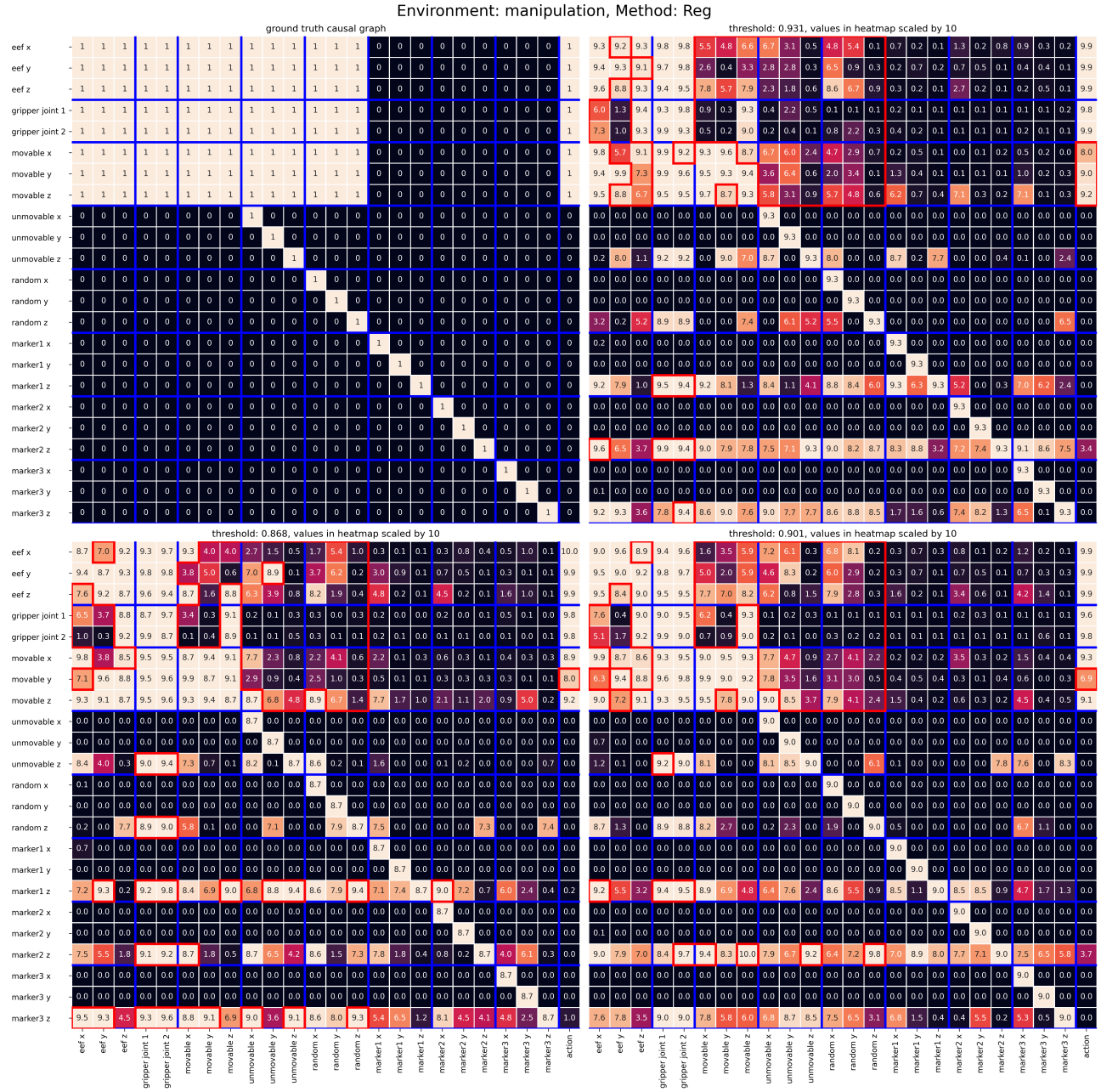

Figure 16. Causal graph for the manipulation environment. **(top left)** the ground truth causal graph, **(others)** causal graphs learned by **Reg** with 3 different runs, which clearly illustrate the challenging tradeoff between prediction and regularization for regularization-based method. Even though trained under the same regularization coefficient, frequent events, like the movement of markers, use lots of spurious correlations to improve prediction performance, while rare events, like eef's dependency on the unmovable objects which may block it, are sacrificed for lower regularization penalty. Consequently, the causal graphs learned by Reg have many false positives and false negatives at the same time.

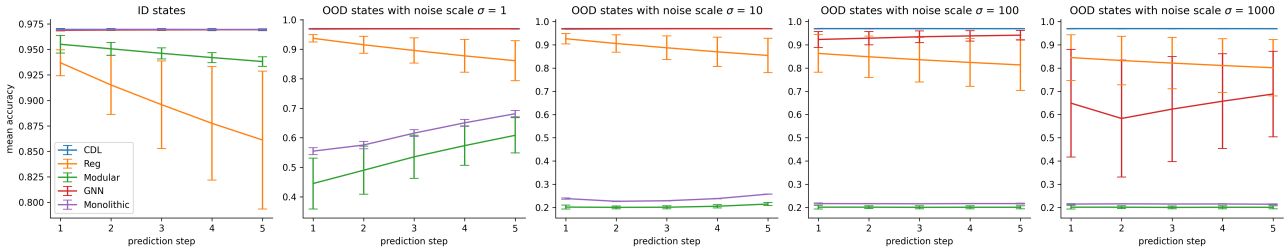

Figure 17. Multi-step prediction performance for the chemical environment with the **collider** graph. **(leftmost)** prediction on ID states. **(others)** prediction on OOD states with increasing noise scale σ.

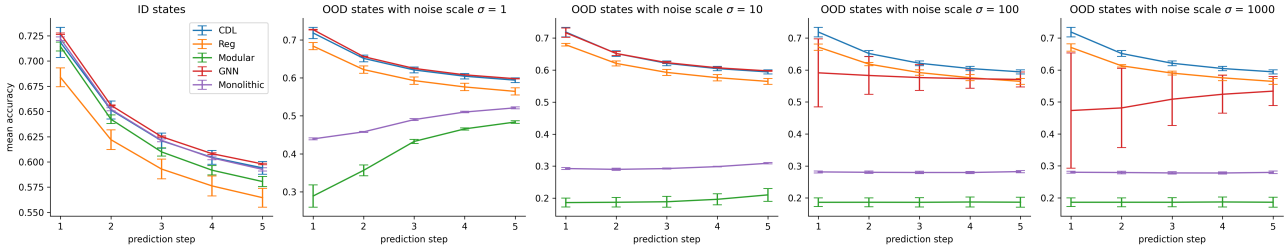

Figure 18. Multi-step prediction performance for the chemical environment with the **chain** graph. **(leftmost)** prediction on ID states. **(others)** prediction on OOD states with increasing noise scale σ.

C.3.2. Predicting Future States Details

In this section, we provide the results for $1 \sim 5$-step prediction results for each method, both evaluated on states in and out of training distribution (denoted as ID and OOD states) except for the physical environment for which it is not practical to create OOD states. To create OOD states, we change object positions in the chemical environment and marker positions in the manipulation environment to random values sampled from a normal distribution $\mathcal{N}(0, \sigma^2)$ where σ controls to what extent the OOD states are different from ID states. Both ID and OOD prediction are measured on other unchanged state variables which do not depend on those changed variables, so ideally prediction performance on OOD states should be the same as the performance on ID states. For 1-step prediction results shown in Sec. 4.3.2, we use $\sigma = 1000$ for the chemical environment and $\sigma = 1$ for the manipulation environment.

The $1 \sim 5$-step prediction results are shown in Fig. $17 \sim 21$. In all environments, CDL achieves similar performance to other methods on ID states. On OOD states, CDL retains similar performance while the performance of dense methods degrades severely. One exception is in the manipulation environment where one run of CDL wrongly uses the spurious correlation on the marker, as shown in Fig. 15, and thus its prediction performance also degrades a little. Meanwhile, although Reg also learns a causal graph and shares the generalization benefits, as it does not learn the causal graphs as accurately as CDL, its prediction performance is worse than CDL both on ID and OOD states.

We also evaluate how much the prediction performance of dense model decreases under different σ values, as shown in Fig. $17 \sim 19$. Surprisingly, though GNN assumes a dense underlying graph between state variables, it still performs

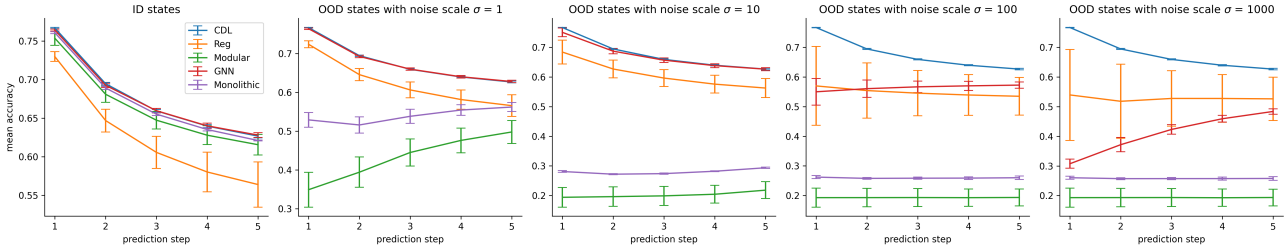

Figure 19. Multi-step prediction performance for the chemical environment with the **full** graph. **(leftmost)** prediction on ID states. **(others)** prediction on OOD states with increasing noise scale σ.

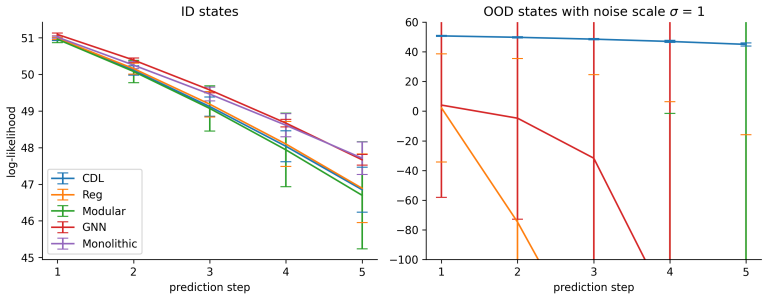

Figure 20. Multi-step prediction performance for the manipulation environment. **(leftmost)** prediction on ID states. **(others)** prediction on OOD states with noise scale $\sigma = 1$.

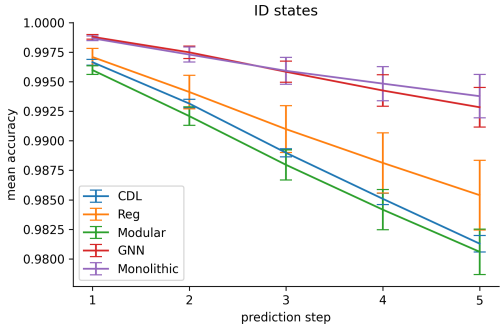

Figure 21. Multi-step prediction performance for the physical environment.

comparably with CDL when $\sigma = 100$ in the chemical environment with the collider graph and when $\sigma = 10$ in the chemical environment with the chain or dense graph. This suggests that the dense model has certain generalizability when being used in environments with simple underlying causal graphs (like the collider graph) and being trained with enough data. However, this is barely the case for more complex environments. For example, in the manipulation environment, GNN's prediction degrades severely even with small noise $\sigma = 1$. Furthermore, we find that, counter-intuitively, for certain dense models, environments, and noise scales, the prediction performance increases with larger prediction steps, such as Modular and Monolithic in Fig. 17 with $\sigma = 1$. We hypothesize that this increasing performance results from the generalizability of dense models just described: though the state given to the models is OOD, as the model rolls out to predict longer into the future, it prefers to predict ID values for states and gradually recovers from the interruption brought by the OOD value. As mentioned, the generalizability of dense models is limited, so for larger values of σ, Modular and Monolithic cannot recover any more and thus the prediction performance no longer increases with the prediction step.

C.4. Transition Collection Policy Learning Evaluation Details

Apart from the results shown in Sec. 4.4, we provide other quantitative and qualitative results of the transition collection policy evaluation in this section.

For the causal graph derivation, we use the same method described in Sec. C.3.1 for threshold selection and the same metrics to evaluate the causal graphs learned from data collected by different policies, which are shown in Table. 9. The

Table 9. Performance on Causal Graph Learning for CDL and Reg on All Environments

Metrics	PredDiff	Uniform	Curiosity
Accuracy	**0.902** $\pm$ 0.003	0.891 $\pm$ 0.006	0.889 $\pm$ 0.002
Recall	0.612 $\pm$ 0.009	0.593 $\pm$ 0.042	0.573 $\pm$ 0.009
Precision	0.980 $\pm$ 0.006	0.943 $\pm$ 0.022	0.956 $\pm$ 0.022
F1 Score	0.754 $\pm$ 0.008	0.726 $\pm$ 0.026	0.716 $\pm$ 0.005
ROC AUC	0.805 $\pm$ 0.005	0.792 $\pm$ 0.019	0.784 $\pm$ 0.003

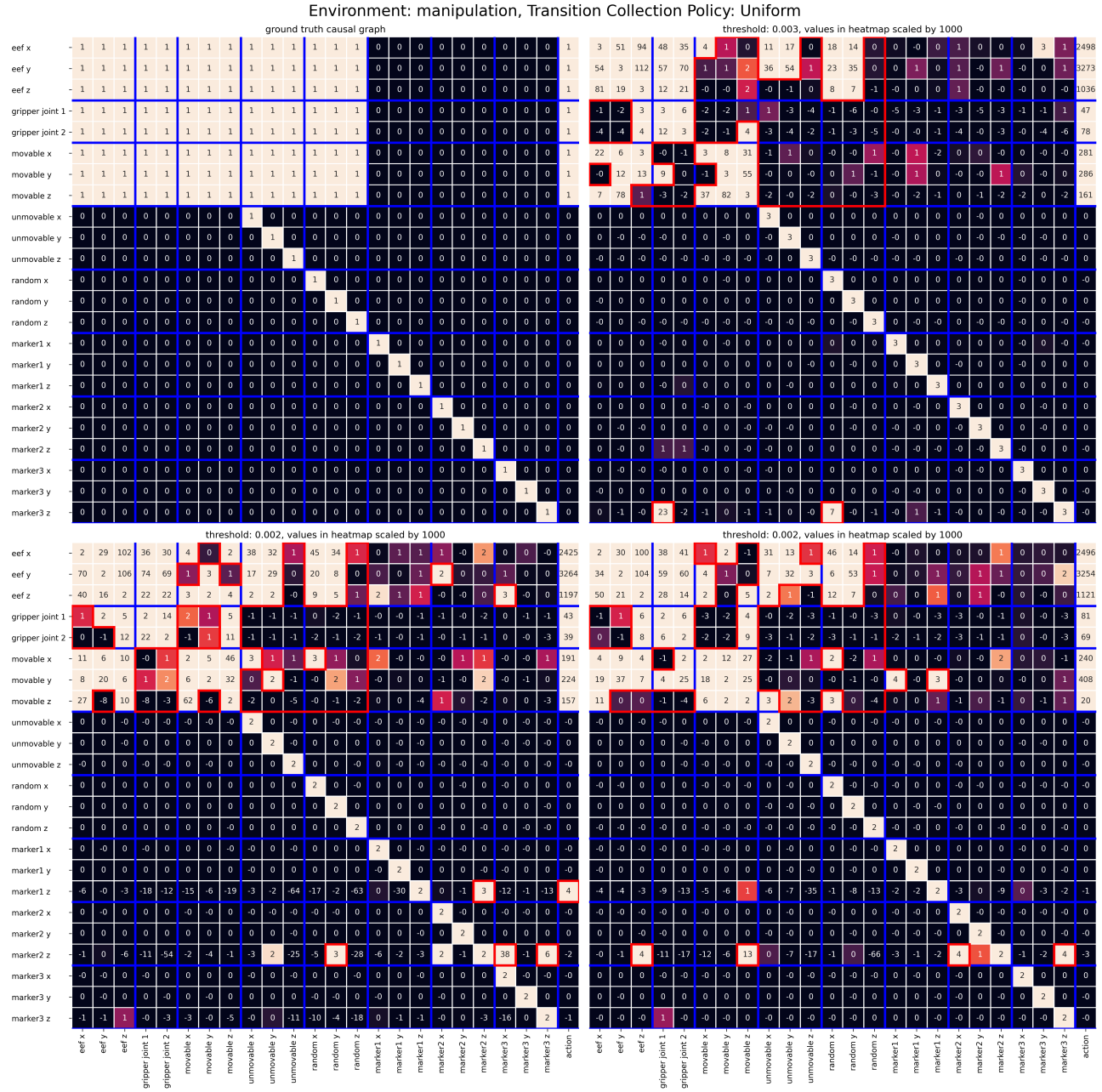

Figure 22. Causal graph for the manipulation environment learned with CDL and transitions collected by **Uniform**. **(top left)** the ground truth causal graph, **(others)** causal graphs learned with data from **Uniform** for 3 different runs. Compared to the graphs learned with data from **PredDiff** shown in Fig. 15, there are more spurious correlations (which is also reflected by the lower precision scores), especially in two bottom graphs, and those spurious correlations limit the generalizability of the learned causal dynamics models.

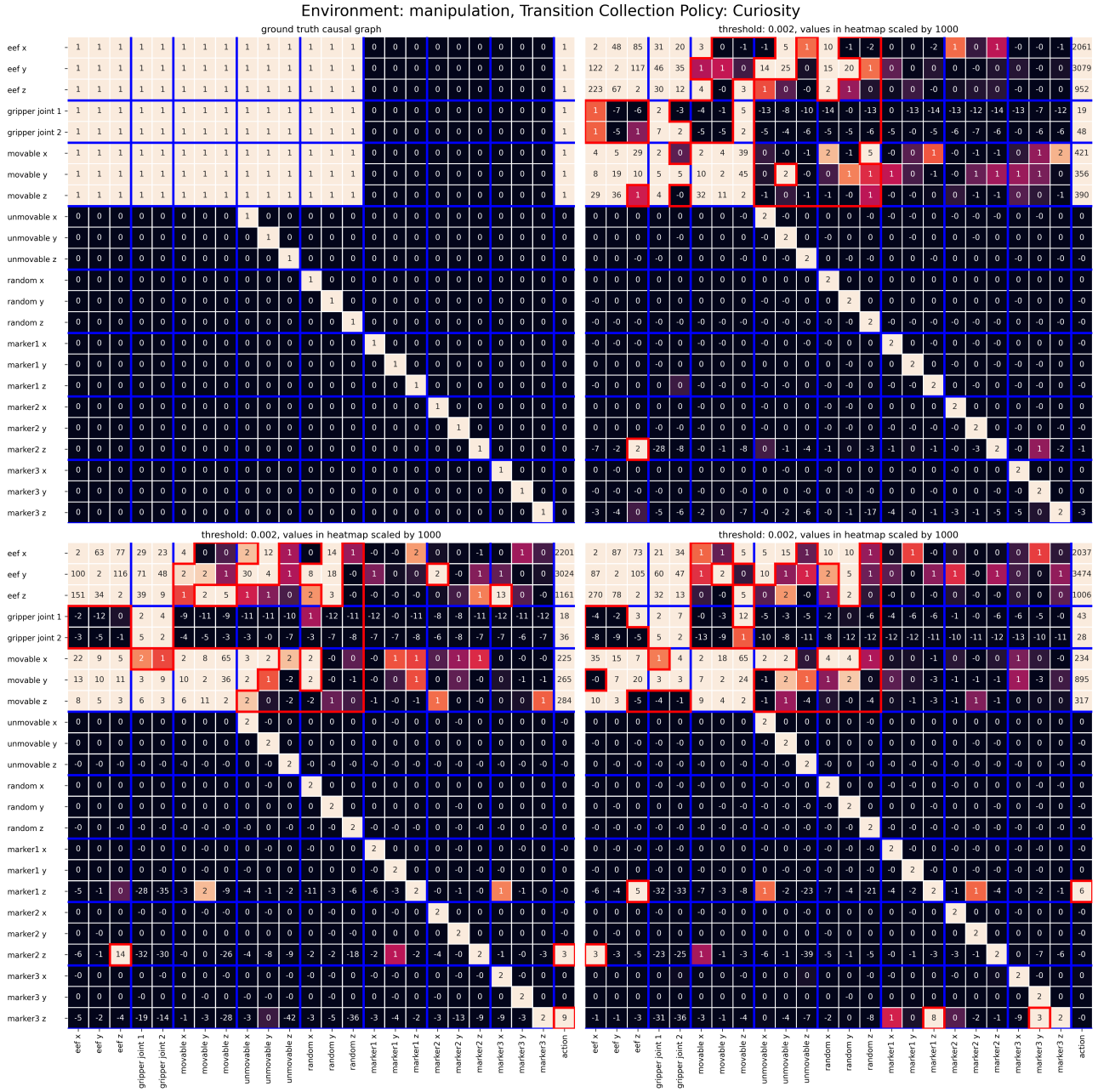

Figure 23. Causal graph for the manipulation environment learned with CDL and transitions collected by **Curiosity**. **(top left)** the ground truth causal graph, **(others)** causal graphs learned with data from **Curiosity** for 3 different runs. Compared to the graphs learned with data from **PredDiff** shown in Fig. 15, there are more missing dependencies (which is also reflected by the lower recall scores), especially eef's dependencies on the unmovable and randomly moving object. Those missing dependencies result from procrastination, a common problem for curiosity agents.

Table 10. Parameters of the Reward Predictor and CEM (Shared Across Tasks if not Specified)

Method	Name	Tasks				
		Chemical (collider)	Chemical (chain, full)	Reach	Manipulation	Physical
Reward Predictor	architecture	[64, 64]		[128, 64]	[64, 64, 64]	
	activation functions	[ReLU, ReLU]		[ReLU, Tanh]	[ReLU, ReLU, ReLU]	
	training step	300K	1M	2M	2M	2M
	optimizer	Adam				
	learning rate	3e-4				
	batch size	32				
CEM	planning length, L	3		1		
	number of candidates, J	64		128		
	number of top candidates, K	32		32		
	number of iterations, N	5		10		
	exploration noise	N/A		0.03		
	exploration probability	0.2		N/A		

causal graphs learned by PredDiff is the same as Fig. 15 and the ones learned by Uniform and Curiosity are shown in Fig, 22 and Fig. 23 respectively. For the Curiosity policy, we also observe procrastination which is a common problem for curiosity-based exploration. Specifically, as it is harder to predict the next time step value of the movable object when it is being moved around than when it stays still, the agent keeps moving the movable object to maximize rewards, rather than actively exploring and exposing unlearned causal relationships, e.g., interacting with the unmovable and randomly moving object to expose eef's dependencies on them.

C.5. Downstream Tasks Learning Evaluation Details

In this section, we give more details on task setup, reward predictor implementations and downstream task learning results.

For the tasks described in Sec. 4.5, their reward functions are defined as follows:

Matching (C): match the object colors with goal colors individually,

$$r_t = \sum_{i=1}^{10} \mathbb{1}\left[c_t^i = g^i\right],$$

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

Reach (M): move the end-effector to the goal position,

$$r_t = \tanh(10 \cdot \|eef_t - g\|_1),$$

where $\|\cdot\|_1$ is L1 norm, $eef_t \in \mathbb{R}^3$ is the current end-effector position, and $g \in \mathbb{R}^3$ is the goal position in this episode.

Lift (M): lift the movable object to the goal position,

$$r_t = \begin{cases} 0.3 + 0.5 \cdot (1 - \frac{\|mov_t - g\|_1}{1.2}) & \text{if } mov \text{ is grasped} \\ 0.1 \cdot (1 - \frac{\|eef_t - mov_t\|_1}{1.2}) \cdot \mathbb{1}\,[\text{gripper is open}] & \text{otherwise} \end{cases}.$$

Stack (M): stack the movable object on the top of the unmovable object.

$$r_t = \begin{cases} 0.35 + 1.0 \cdot (1 - \frac{\|mov_t - g\|_1}{1.2}) & \text{if } mov \text{ is grasped} \\ 2.0 & \text{if } mov \text{ is not grasped and is on top of } unm \\ 0.1 \cdot (1 - \frac{\|eef_t - mov_t\|_1}{1.2}) \cdot \mathbb{1}\,[\text{gripper is open}] & \text{otherwise} \end{cases}.$$

During training, the dynamics model is frozen and only the reward predictor is learned as described in Sec. 3.5. The architecture and hyperparameters of the reward predictor are listed in Table. 10, along with the parameters of Cross-Entropy Method (CEM) that is used for planning.

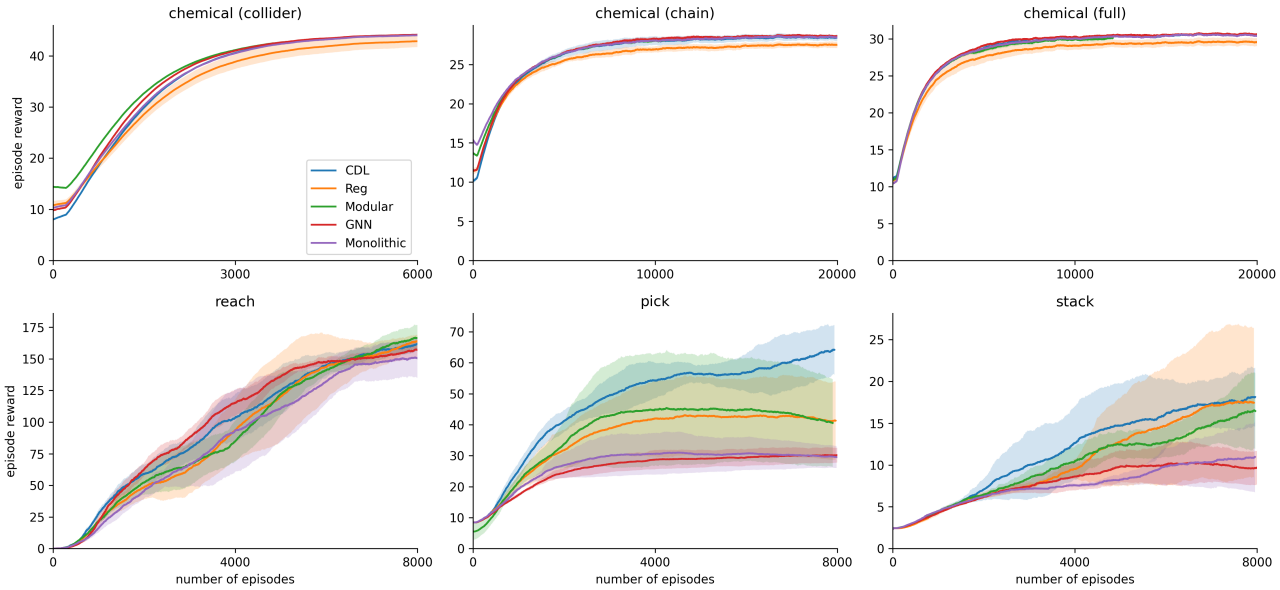

Figure 24. Traninig curve for all downstream tasks.

The training curves for all tasks are shown in Fig. 24. Compared to lift and stack, the sample efficiency advantages of the derived are less obvious for match and reach tasks, because those tasks are relatively simple and have dense reward signals that are easy to learn.