Appendix

Table of Contents

A Theoretical Proofs **18**

A.1 Formulation of Latent Variable Models 18

A.2 Transition Model Learning 19

A.3 Derivation of Planner Module 20

A.4 Causal Discovery 23

A.5 Overall Performance Guarantee of Iterative Optimization 26

B Additional Experiment **27**

B.1 Overall Performance 27

B.2 Causal Graph Analysis 28

B.3 Distance between Goal and State Distribution 28

B.4 Task Performance of Chemistry Experiment 28

C Additional Information **29**

C.1 Details about Conditional Independence Test 29

C.2 Experiment Details 30

C.3 Broader Social Impact and Additional Limitation 31

A Theoretical Proofs

In Appendix A, we first show the derivation of the latent variable models A.1, then provide some analytical results in the iterative optimization of model learning A.2, planning A.3, and causal discovery A.4. Finally, we give the proof of the theorem of overall performance guarantee A.5 given some common assumptions.

To compactly write down our formulas, we slightly abuse the notations by representing s^t, a^t as the joint states and actions at timestep t, while using s_i, a_i to denote the i-th dimension of factorized states or actions. Without the loss of generality, we implement our reward in a deterministic way, which only involves $r(s, g)$ in its notation, we will slightly generalize to some state-action reward function for our analysis as well.

A.1 Formulation of Latent Variable Models

A.1.1 Derivation of Equation (1)

The ELBO of the likelihood of the trajectory is obtained by

$$
\begin{aligned}
\log p(\tau|s^*) &= \log \int p(\tau|\mathcal{G}, s^*) p(\mathcal{G}|s^*) d\mathcal{G} \\
&= \log \int q(\mathcal{G}|\tau) \frac{p(\tau|\mathcal{G}, s^*) p(\mathcal{G}|s^*)}{q(\mathcal{G}|\tau)} d\mathcal{G} \\
&\geq \int q(\mathcal{G}|\tau) \log \frac{p(\tau|\mathcal{G}, s^*) p(\mathcal{G}|s^*)}{q(\mathcal{G}|\tau)} d\mathcal{G} \\
&= \int q(\mathcal{G}|\tau) \left(\log p(\tau|\mathcal{G}, s^*) + \log \frac{p(\mathcal{G}|s^*)}{q(\mathcal{G}|\tau)} \right) d\mathcal{G} \\
&= \int q(\mathcal{G}|\tau) \log p(\tau|\mathcal{G}, s^*) d\mathcal{G} + \int q(\mathcal{G}|\tau) \log \frac{p(\mathcal{G}|s^*)}{q(\mathcal{G}|\tau)} d\mathcal{G} \\
&= \mathbb{E}_{q(\mathcal{G}|\tau)}[\log p(\tau|\mathcal{G}, s^*)] - \mathbb{D}_{\mathrm{KL}}[q(\mathcal{G}|\tau)||p(\mathcal{G})]
\end{aligned}
\tag{9}
$$

where the third line is obtained by Jensen's inequality and the last line is because the prior of the causal graph $\mathcal{G}$ is independent of the achieved goal s^*.

A.1.2 Derivation of Equation (2)

According to the decomposition of state-action trajectory

$$p(\tau) = p(s^0) \sum_{t=0}^{T-1} p(s^{t+1}|s^t, a^t) p(a^t|s^t) \tag{10}$$

we can get the following

$$\begin{aligned} \log p(\tau|\mathcal{G}, s^*) &= \log(s^0, a^0, s^1, a^1, \cdots, a^{T-1}, s^T|\mathcal{G}, s^*) \\ &= \log p(s^0|\mathcal{G}, s^*) + \sum_{t=0}^{T-1} \log p(s^{t+1}|s^t, a^t, \mathcal{G}, s^*) + \sum_{t=0}^{T-1} \log p(a^t|s^t, \mathcal{G}, s^*) \\ &= \log p(s^0) + \sum_{t=0}^{T-1} \log p(s^{t+1}|s^t, a^t, \mathcal{G}) + \sum_{t=0}^{T-1} \log p(a^t|s^t, \mathcal{G}, s^*) \end{aligned} \tag{11}$$

A.2 Transition Model Learning

The optimization in the model learning step can be described below:

$$\arg\max_{\theta} [\sum_{t} \log p_\theta(s_{t+1}|s_t, a_t, \mathcal{G})] \tag{12}$$

where $\tau = [s_1, a_1, \cdots, s_T]$ is the trajectory in data buffer, and $\mathcal{G}$ is the given causal graph.

Here below, we show some necessary definitions and propositions to prove the Lemma 1.

Definition 4 (Structural Hamming Distance (SHD))**.** *For any two DAGs $\mathcal{G}, \mathcal{H}$ with identical vertices set V, we define the following function SHD: $\mathcal{G} \times \mathcal{H} \to \mathbb{R}$,*

$$\textit{SHD}(\mathcal{G}, \mathcal{H}) = \#\{(i,j) \in V^2 \mid \mathcal{G} \textit{ and } \mathcal{H} \textit{ have different edges } e_{ij}\} \tag{13}$$

Definition 5 (Respect the graph)**.** *For any given transition model with specific causal graph $\mathcal{G}$, the transition model respects the graph if the distribution $p(s_{t+1}|a_t, s_t, \mathcal{G})$ can be factorized as:*

$$p(s'|s, a, \mathcal{G}) = \prod_{i \in [M]} p(s'_i|\boldsymbol{PA}(s'_i), \mathcal{G}) \tag{14}$$

where M is the total number of factorized states, $\boldsymbol{PA}(\cdot)$ represents the parents in the causal graph.

Proposition 3 (GRU model respects the graph)**.** *As the parameterized transition model $p_\theta(s'|s, a, \mathcal{G})$ reaches the steady state, it respects the graph.*

Proof of Proposition 3. The GRU modules with parameter $\theta = [W, U]$ can be rewritten as a message passing process, where $AGG.(\cdot)$ is the iterative aggregation function.

$$\begin{aligned} \text{Node Encoder}: h_j^{(0)} &= f_{\text{encoder}}(x_j) \\ \text{Aggregation}: h_j^{(\ell)} &= AGG_{i \in \mathcal{N}(j)}(f_\theta(x_j^{(\ell-1)}, h_i^{(\ell-1)})) \\ \text{Node Decoder}: \boldsymbol{x}_i^{(\ell)} &= f_{\text{decoder}}(h_j^{(\ell-1)}) \end{aligned} \tag{15}$$

As an iterated process of message passing, where the input causal graph controls the information flow between different entities, this GRU model can be rewritten as a fixed point iteration [92]:

$$\boldsymbol{x}_i^{(\ell)} = F_\theta(\mathbf{PA}(\boldsymbol{x}_i)^{(\ell-1)}, \boldsymbol{x}_i^{(\ell-1)}) \tag{16}$$

With proper initialization and some sufficient conditions provided by [92], F has a unique equilibrium point, where

$$\boldsymbol{x}_i^{\infty} = F_\theta(\mathbf{PA}(\boldsymbol{x}_i)^{\infty}, \boldsymbol{x}_i^{\infty}) \tag{17}$$

In our bipartite graph, when GRU reaches the equilibrium point, we can get a structural causal model:

$$s'_i = F_\theta(\mathbf{PA}(s'_i), s_i), \quad \text{where } s'_i \in \mathcal{S}', \mathbf{PA}(s'_i) \in \mathcal{S} \cup \mathcal{A} \tag{18}$$

Based on the SCM derivation 18, we can then factorize the transition model as:

$$p_\theta(s'|s, a, \mathcal{G}) = \prod_{i \in [M]} p_\theta(s'_i|\mathbf{PA}(s'_i), \mathcal{G}) \tag{19}$$

We denote the ground truth causal graph as $G^* = (V, E^*)$, and $\mathbf{PA}^*(s'_i)$ as the true parents of s'_i in G^*. ■

Definition 6 (Causal optimality at equlibrium point)**.** *For any $G' \neq G^*$ with at least one pair of flawed parental relationship $\boldsymbol{PA}'(s'_i) \neq \boldsymbol{PA}^*(s'_i)$, the following inequality holds:*

$$p_\theta(s'_i|\boldsymbol{PA}'(s'_i), \mathcal{G}) \leq p_\theta(s'_i|\boldsymbol{PA}^*(s'_i), \mathcal{G}) \tag{20}$$

Lemma 5 (Local monotonicity)**.** *Given one state variable s_i and its any parental relationship $\boldsymbol{PA}^1(s_i), \boldsymbol{PA}^2(s_i)$, if $\#(\boldsymbol{PA}^1(s_i) \cup \boldsymbol{PA}^*(s_i)) \geq \#(\boldsymbol{PA}^2(s_i) \cup \boldsymbol{PA}^*(s_i))$, then at steady state, the SCM derived in 18 will miss part of the message provided from the true parents, therefore $p_\theta(s'_i|\boldsymbol{PA}^1(s'_i), \mathcal{G}_1) \geq p_\theta(s'_i|\boldsymbol{PA}^2(s'_i), \mathcal{G}_2)$*

Proof of Lemma 1. Based on the factorization defined in 19, we denote the parental relationship in $\mathcal{G}_1$ as $\mathbf{PA}^1(\cdot)$,

$$p_\theta(s'|a, s, \mathcal{G}^*) = \prod_{i \in [M]} p_\theta(s'_i|\mathbf{PA}^*(s'_i), \mathcal{G}) \geq \prod_{i \in [M]} p_\theta(s'_i|\mathbf{PA}^1(s'_i), \mathcal{G}) = p_\theta(s'|a, s, \mathcal{G}_1) \tag{21}$$

For $\mathcal{G}_1$ and $\mathcal{G}_2$, suppose the only different edges e has a target node s'_j, with $\text{SHD}(\mathcal{G}_1, \mathcal{G}^*) < \text{SHD}(\mathcal{G}_2, \mathcal{G}^*)$, based on Lemma 5:

$$\begin{aligned}
p_\theta(s'|a, s, \mathcal{G}_1) &= p_\theta(s'_j|\mathbf{PA}^1(s'_j), \mathcal{G}_1) \prod_{i \in [M] \setminus j} p_\theta(s'_i|\mathbf{PA}^1(s'_i), \mathcal{G}_1) \\
&\geq p_\theta(s'_j|\mathbf{PA}^2(s'_j), \mathcal{G}_2) \prod_{i \in [M] \setminus j} p_\theta(s'_i|\mathbf{PA}^1(s'_i), \mathcal{G}_2) \\
&= p_\theta(s'_j|\mathbf{PA}^2(s'_j), \mathcal{G}_2) \prod_{i \in [M] \setminus j} p_\theta(s'_i|\mathbf{PA}^2(s'_i), \mathcal{G}_2) \\
&= p_\theta(s'|a, s, \mathcal{G}_2).
\end{aligned} \tag{22}$$

Based on the inequality derived in 21 and 22,

$$\log p_\theta(s'|s, a, \mathcal{G}^*) \geq \log p_\theta(s'|s, a, \mathcal{G}_1) \geq \log p_\theta(s'|s, a, \mathcal{G}_2). \tag{23}$$

the monotonicity of likelihood in Lemma 1 is proved. ■

A.3 Derivation of Planner Module

The optimization in the planning part is:

$$\max_\pi \sum_{t=0}^{T-1} \log \pi_\theta(a^t|s^t, \mathcal{G}, s^*) = \max_{[a_0, \cdots, a^{T-1}]} \sum_{t=0}^{T-1} \log \hat{Q}(s^t, a^t) \tag{24}$$

Ideally, given the access to real dynamics $p(s'|s, a)$ and goal distribution $p(g)$. We first define the expected goal-conditioned state-action reward $r(s, a, g) = \mathbb{E}_{s' \sim p(\cdot|s,a)} r(s', g)$, and the expected state-action reward $r(s, a) = \mathbb{E}_{g \sim p_g(\cdot)} r(s, a, g)$. In practice, due to the inaccuracy of transition model, we can only query the following reward estimation at certain state-action pair: $r(s, a, g) = \mathbb{E}_{s' \sim p_\theta(\cdot|s,a,\mathcal{G})} r(s', g), r(s, a) = \mathbb{E}_{g \sim p_g(\cdot)} r(s, a, g)$.

Then we consider the distribution of the goal $g \sim p_g(\cdot)$, which is supported on the state space $\mathcal{S}$. Based on Algorithm 1, our interventional data is collected by the MPC that maximizes the expected discounted cumulative reward from learned dynamics. Thus, we could denote the interventional distribution of state (depending on the current policy π) in the data buffer as $p_{\mathcal{I}^s_\pi}$, $s \sim p_{\mathcal{I}^s_\pi}(\cdot)$ which is also supported on the state space $\mathcal{S}$.

Proof of Lemma 3. Assume our planning algorithm has an infinite planning horizon, with the optimal transition dynamics and optimal policy, the action-value function Q^* can be expressed as:

$$Q^*(s^t, a^t) \stackrel{\text{def}}{=} \mathbb{E}_{s \sim p_{\mathcal{I}^s_{\pi^*}}(\cdot), a \sim \pi^*(s)} \left[\sum_{t'=t}^{\infty} \gamma^{t'-t} r(s^{t'}, a^{t'}) \mid s^t, a^t \right] \tag{25}$$

The estimation of action value function $\hat{Q}(s^t, a^t) = Q^{\hat{\pi}}_{\theta,\mathcal{G}}(s^t, a^t)$ can be written as:

$$\begin{aligned} \hat{Q}(s^t, a^t) &\stackrel{\text{def}}{=} \mathbb{E}_{s \sim p_{\mathcal{I}^s_{\hat{\pi}}}(\cdot), a \sim \hat{\pi}(s)} \left[\sum_{t'=t}^{\infty} \gamma^{t'-t} r(s^{t'}, a^{t'}) \mid s^t, a^t \right] \\ &= \mathbb{E}_{a \sim \hat{\pi}(s)} \left[\sum_{t'=t}^{\infty} \gamma^{t'-t} \mathbb{E}_{g \sim p_g(\cdot), s \sim p_{\mathcal{I}^s_{\hat{\pi}}}(\cdot)} (1 - \mathbb{1}(s' = g)) \mid s^t, a^t \right] \end{aligned} \tag{26}$$

The policy by MPC in algorithm 1 can be deducted by: $\hat{\pi}(s^t) = \arg\max_{a^t \in \mathcal{A}} \hat{Q}(s^t, a^t)$, let $s^0 = s$, and we could derive value function under the MPC policy as follows:

$$\begin{aligned} V(s) &= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{p_g} \mathbb{E}_{p_{\mathcal{I}^s_{\hat{\pi}}}} \mathbb{1}(s = g) = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{p_g} \mathbb{E}_{p_{\mathcal{I}^s_{\hat{\pi}}}} [1 - \mathbb{1}(s \neq g)] \\ &= \sum_{t=0}^{\infty} \gamma^t - \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{p_g} \mathbb{E}_{p_{\mathcal{I}^s_{\hat{\pi}}}} \mathbb{1}(s \neq g) \\ &\leq \frac{1 - \mathbb{D}_{TV}(p_{\mathcal{I}^s_{\hat{\pi}}}, p_g)}{1 - \gamma} \end{aligned} \tag{27}$$

where $\mathbb{D}_{\text{TV}}(p_{\mathcal{I}^s_{\pi}}, p_g)$ is the total variation distance between the marginal state distribution $p_{\mathcal{I}^s_{\pi}}$ in the data buffer, as well as the goal distribution p_g, both of which share the same support. ■

In addition, we could define a more general form of goal-conditioned reward as based on the distance: $r(s, g) = 1 - d(s', g)$. Where $\mathbb{D}$ is a (normalized) distance measure between two vectors in the state space, s.t. $\forall s, g \in \mathcal{S}$, $0 \leq d(s, g) \leq 1$. For instance, if we pick $d(s, g) = \mathbb{1}(s \neq g)$, the derived reward under this distance measure will go back to the reward function defined in section 2.1. By defining a (normalized) ℓ_p distance between s' and g, $d(s, g) = \frac{\|s-g\|_p}{\max_{s_1, s_2 \in \mathcal{S}} \|s_1 - s_2\|_p}$, we can also shape a continuous form of goal-conditioned step reward $r(s, g)$ between 0 and 1. Notice that all the Euclidean-based distances are all valid metrics with symmetry, non-negativity, the identity of indiscernibles, and the triangle inequality. With such a definition, the estimated value function will fit in with:

$$V(s) \leq \frac{1 - \mathcal{W}(p_{\mathcal{I}^s_{\pi}}, p_g)}{1 - \gamma} \tag{28}$$

where $\mathcal{W}$ is some Wasserstein distance between marginal state distribution and goal distribution. Therefore, optimizing the Q value is equivalent to minimizing an upper bound for some types of the statistical distance between goal distribution and target distribution.

For the term related to policy in (3), we can define the goal-conditioned policy distribution as:

$$\pi(a^t | s^t, g) \propto \exp(Q(s^t, a^t)) \tag{29}$$

As a result, $\arg\max_\pi \sum_{t=0}^{T-1} \log \pi_\theta(a^t | s^t, s^*, \mathcal{G}) = \arg\max_\pi \sum_{t=0}^{T-1} Q(s^t, a^t)$ However, the real $Q(s^t, a^t)$ is intractable, so we alternatively optimize the $\hat{Q}(s^t, a^t)$ at each time step. Next, we'll start to derive a bound between $\hat{Q}(s, \hat{\pi}(s))$ and $Q(s, \pi^*(s))$

Proof of Lemma 2. For simplicity, we denote the learned transition function $\hat{p}(s'|s, a) = p_\theta(s'|s, a, \mathcal{G})$, which is ϵ_m-approximate dynamics, $\mathbb{D}_{\text{TV}}(\hat{p}, p) = \|\hat{p}(s'|s, a) - p(s'|s, a)\|_\infty \leq \epsilon_m$,

Firstly, we show by value iteration that the estimated value function $\hat{V}(s)$ will converge to $V(s)$: Assume exists $K > 0$, s.t. $\forall k > K$, $\|\hat{p}(s'|, s, a) - p(s, |s, a)\|_\infty \leq \epsilon_m$

$$\hat{V}^{(k+1)}(s) = r(s, \pi(s)) + \gamma \sum_{s'} p(s'|s, \pi(s)) \hat{V}^{(k)}(s). \tag{30}$$

Given the result of the Bellman Contraction,

$$\|\hat{V}^{(k+1)}(s) - V^{\pi^*}(s)\|_\infty \leq \gamma \|\hat{V}^{(k)}(s) - V^{\pi^*}(s)\|_\infty, \quad \lim_{k\to\infty} \|\hat{V}^{(k+1)}(s) - V^{\pi^*}(s)\|_\infty = 0. \tag{31}$$

Based on the definition of greedy policy in planning: $\hat{\pi}(s) = \arg\max_{a\in\mathcal{A}} \hat{Q}(s,a)$, we can derive the inequality:

$$\begin{aligned}
& r(s,\hat{\pi}(s)) + \gamma \sum_{s'} \hat{p}(s'|s,\hat{\pi}(s))\hat{V}(s') \geq r(s,\pi^*(s)) + \gamma \sum_{s'} \hat{p}(s'|s,\pi^*(s))\hat{V}(s') \\
\implies \quad & r(s,\pi^*(s)) - r(s,\hat{\pi}(s)) \leq \gamma \left[\sum_{s'} \hat{p}(s'|s,\hat{\pi}(s))\hat{V}(s') - \sum_{s'} \hat{p}(s'|s,\pi^*(s))\hat{V}(s') \right] \\
\overset{\|\hat{V}(s)-V^{\pi^*}(s)\|_\infty \to 0}{\implies} \quad & r(s,\pi^*(s)) - r(s,\hat{\pi}(s)) \leq \gamma \left[\sum_{s'} \hat{p}(s'|s,\hat{\pi}(s))V^{\pi^*}(s') - \sum_{s'} \hat{p}(s'|s,\pi^*(s))V^{\pi^*}(s') \right].
\end{aligned} \tag{32}$$

Then let s be the state with the largest value error.

$$\begin{aligned}
V^{\pi^*}(s) - V^{\hat{\pi}}(s) =& r(s,\pi^*(s)) - r(s,\hat{\pi}(s)) \\
& + \gamma \Big[\sum_{s'} p(s'|s,\pi^*(s))V^{\pi^*}(s') - \sum_{s'} p(s'|s,\hat{\pi}(s))V^{\hat{\pi}}(s') \Big] \\
\leq & \gamma \sum_{s'} \Big[\hat{p}(s'|s,\hat{\pi}(s))V^{\pi^*}(s') - \hat{p}(s'|s,\pi^*(s))V^{\pi^*}(s') \Big] \\
& + \gamma \sum_{s'} \Big[p(s'|s,\pi^*(s))V^{\pi^*}(s') - p(s'|s,\hat{\pi}(s))V^{\hat{\pi}}(s') \Big] \\
= & \gamma \sum_{s'} \Big[p(s'|s,\pi^*(s)) - \hat{p}(s'|s,\pi^*(s)) \Big] V^{\pi^*}(s') \\
& - \gamma \sum_{s'} \Big[p(s'|s,\hat{\pi}(s)) - \hat{p}(s'|s,\hat{\pi}(s)) \Big] V^{\pi^*}(s') \\
& + \gamma \sum_{s'} p(s'|s,\hat{\pi}(s)) \Big[V^{\pi^*}(s) - V^{\hat{\pi}}(s) \Big]
\end{aligned} \tag{33}$$

Since $r(s,g) \in [0,1]$, the value function $V(s) \in [0, \frac{1}{1-\gamma}]$, also by $\|\hat{p}(s'|s,\hat{\pi}(s)) - p(s'|s,\hat{\pi}(s))\| \leq \epsilon$, we have

$$\begin{aligned}
V^{\pi^*}(s) - V^{\hat{\pi}}(s) &\leq \gamma\epsilon_m(V_{\max} - V_{\min}) + \gamma \sum_{s'} p(s'|s,\hat{\pi}(s)) \Big[V^{\pi^*}(s) - V^{\hat{\pi}}(s) \Big] \\
&= \frac{\gamma\epsilon_m}{1-\gamma} + \gamma \sum_{s'} p(s'|s,\hat{\pi}(s)) \left[V^{\pi^*}(s') - V^{\hat{\pi}}(s') \right].
\end{aligned} \tag{34}$$

We already analyzed the state s with the largest value error, and it's sufficient to show:

$$\begin{aligned}
\|V^{\pi^*}(s) - V^{\hat{\pi}}(s)\|_\infty &\leq \frac{\gamma\epsilon_m}{1-\gamma} + \gamma \sum_{s'} p(s'|s,\hat{\pi}) \|V^{\hat{\pi}}(s) - V^{\pi^*}(s)\|_\infty \\
&= \frac{\gamma\epsilon_m}{1-\gamma} + \gamma \|V^{\pi^*}(s) - V^{\hat{\pi}}(s)\|_\infty
\end{aligned} \tag{35}$$

By combining $\|V^{\pi^*}(s) - V^{\hat{\pi}}(s)\|_\infty$ on both sides, we have

$$\|V^{\pi^*}(s) - V^{\hat{\pi}}(s)\|_\infty \leq \frac{\gamma}{(1-\gamma)^2}\epsilon_m \tag{36}$$

■

A.4 Causal Discovery

A.4.1 Assumptions of Causality

Assumption 2 (Markov property)**.** *Given a DAG $\mathcal{G}$ and a joint distribution $P_{\boldsymbol{X}}$, this distribution is said to satisfy*

- *(i) the global Markov property with respect to the DAG $\mathcal{G}$ if*

$$\boldsymbol{A} \perp\!\!\!\perp_{\mathcal{G}} \boldsymbol{B} | \boldsymbol{C} \Rightarrow \boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} | \boldsymbol{C} \tag{37}$$

  for all disjoint vertex sets $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$. The symbol $independent_{\mathcal{G}}$ denotes d-separation.

- *(ii) the local Markov property with respect to the DAG $\mathcal{G}$ if each variable is independent of its non-descendants (without its parents) given its parents, and*

- *(iii) the Markov factorization property with respect to the DAG $\mathcal{G}$ if*

$$p(\boldsymbol{x}) = p(x_1, \ldots, x_d) = \prod_{j=1}^{d} p(x_j | \boldsymbol{PA}^{\mathcal{G}}(x_j)) \tag{38}$$

  where we assume that $P_{\boldsymbol{X}}$ has a density p.

Assumption 3 (Faithfulness)**.** *Consider a distribution $P_{\boldsymbol{X}}$ and a DAG $\mathcal{G}$, $P_{\boldsymbol{X}}$ is faithful to the DAG $\mathcal{G}$ if we know*

$$\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} | \boldsymbol{C} \Rightarrow \boldsymbol{A} \perp\!\!\!\perp_{\mathcal{G}} \boldsymbol{B} | \boldsymbol{C} \tag{39}$$

for all disjoint vertex sets $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$.

A.4.2 KL Divergence as Sparsity Regularization

Similar to the assumption in Factorized MDP, the existence of a causal relationship between two arbitrary entities among x is can also be treated as independent. Therefore, we construct the prior distribution $p(\mathcal{G})$ as independent Bernoulli Distribution in the transition causal graph.

$$p(\mathcal{G}) = \prod_{i \in [M+N], j \in [M]} p(\mathcal{G}_{ij}) = \prod_{i \in [M+N], j \in [M]} p_{ij} \tag{40}$$

where $\mathcal{G}_{ij}$ represents the edge from i-th node in source node set $\mathcal{U} = \{\mathcal{A} \cup \mathcal{S}\}$ to the j-th node in target node set $\mathcal{V} = \{\mathcal{S}'\}$ in the bipartite transition causal graph.

On the other hand, for the variational posterior $q(\mathcal{G}|\tau)$, for the discovered transition causal graph, it needs to satisfy two constraints: (i) $q(\mathcal{G}|\tau)$ needs to be a DAG, denoted $\mathcal{Q}_{DAG}$, and more specifically, a bipartite graph. We denote such subset of DAG as $\mathcal{Q}_{Bi}$, (ii) $q(\mathcal{G}|\tau)$ needs to be as sparse as possible.

Common score-based causal discovery works use two regularization terms, DAGNess and ℓ_1 regularization to constrain the discovered causal graph in the constraint set, while in our work, we explicitly constrain the posterior variational distribution $q(\mathcal{G}|\tau) \in \mathcal{Q}_{Bi} \subset \mathcal{Q}_{DAG}$. We then show in the following section that by defining a certain independent Bernoulli prior $p(\mathcal{G})$, the KL divergence between variational posterior $q(\mathcal{G}|\tau)$ and $p(\mathcal{G})$ can be equivalent to a sparsity regularization.

According to our constraint-based causal reasoning modules, $\mathcal{Q}_{Bi}$ consists of $M(M+N)$ independent binary classifiers (that form a DAG) parameterized by our kernel-based independent testing modules ϕ, i.e.

$$q_\phi(\mathcal{G}|\tau) = \prod_{i \in [M+N], j \in [M]} q_\phi(\mathcal{G}_{ij}|\tau) \triangleq \prod_{i \in [M+N], j \in [M]} q_{ij} \tag{41}$$

Proof of Proposition 1. Let the prior $p_{ij} = \epsilon_{\mathcal{G}} \in (0, \frac{1}{2}], \forall i \in [M+N], j \in [M]$, based on the definition above, the KL divergence term in (3) can be expanded as follows:

$$\begin{aligned}
\mathbb{D}_{\mathrm{KL}}(q_\phi(\mathcal{G}|\tau)\|p(\mathcal{G})) &= \sum_{q\in\mathcal{Q}_{Bi}} \prod_{i,j} q_{ij} \log \frac{\prod_{i,j} q_{ij}}{\prod_{i,j} p_{ij}} = \sum_{i,j} \left[q_{ij} \log \frac{q_{ij}}{p_{ij}} + (1-q_{ij}) \log \frac{1-q_{ij}}{1-p_{ij}} \right] \\
&= \sum_{i,j} \left[q_{ij} \log \frac{q_{ij}}{\epsilon_{\mathcal{G}}} + (1-q_{ij}) \log \frac{1-q_{ij}}{1-\epsilon_{\mathcal{G}}} \right] \\
&= \sum_{i,j} \left[q_{ij} \log q_{ij} + (1-q_{ij}) \log(1-q_{ij}) - q_{ij} \log \epsilon_{\mathcal{G}} - (1-q_{ij}) \log(1-\epsilon_{\mathcal{G}}) \right]
\end{aligned} \tag{42}$$

Since $q_{ij} \in \{0,1\}$, $\lim_{q_{ij}\to 0} q_{ij} \log q_{ij} = \lim_{q_{ij}\to 1} (1-q_{ij}) \log(1-q_{ij}) = 0$,

$$\begin{aligned}
\mathbb{D}_{\mathrm{KL}}(q_\phi(\mathcal{G}|\tau)\|p(\mathcal{G})) &= \sum_{i,j} \left[-q_{ij} \log(\epsilon_{\mathcal{G}}) - (1-q_{ij}) \log(1-\epsilon_{\mathcal{G}}) \right] \\
&= \sum_{i,j} \left[-\mathbb{I}(q_{ij}=1) \log \epsilon_{\mathcal{G}} - \mathbb{I}(q_{ij}=0) \log(1-\epsilon_{\mathcal{G}}) \right] \\
&= \sum_{i,j} \left[-(1-\mathbb{I}(q_{ij}=1)) \log(1-\epsilon_{\mathcal{G}}) - \mathbb{I}(q_{ij}=1) \log \epsilon_{\mathcal{G}} \right] \\
&= \log \frac{1-\epsilon_{\mathcal{G}}}{\epsilon_{\mathcal{G}}} \sum_{i,j} \mathbb{I}(q_{ij}=1)) - \sum_{i,j} \log(1-\epsilon_{\mathcal{G}}) \\
&= \log \left(\frac{1-\epsilon_{\mathcal{G}}}{\epsilon_{\mathcal{G}}} \right) |q_\phi(\mathcal{G}|\tau)|_1 + \text{const} \\
&\triangleq \eta |q_\phi(\mathcal{G}|\tau)|_1 + \text{const}
\end{aligned} \tag{43}$$

Therefore, the KL divergence term is equivalent to an ℓ_1 sparsity regularizer in score-based causal discovery [54]. The strength of this regularizer $\eta = \log\left(\frac{1-\epsilon_{\mathcal{G}}}{\epsilon_{\mathcal{G}}}\right) \in [0, \infty)$. The larger $\epsilon_{\mathcal{G}}$ in prior Bernoulli distribution indicates the smaller strength of this sparsity regularizer (e.g. when $\epsilon_{\mathcal{G}} = \frac{1}{2}, \eta = 0$). In the implementation of data-efficient causal discovery, we adjust the parameter of the classifier to set the strength of this sparsity constraint. ■

A.4.3 Unique Identifiability of Causal Graph

We construct our causal model based on the Factorized MDP in Assumption 1. According to the definition, the causal graph is a directed bipartite graph, with s^t, a^t on the source side, and s^{t+1} on the target side. For the theoretical analysis part in Section A.4.3 and A.2, we denote s^{t+1} as s', s_t as s, a_t as a and $\boldsymbol{x} = \{\mathcal{A} \cup \mathcal{S} \cup \mathcal{S}'\}$, $\boldsymbol{x} \in \mathbb{R}^{2M+N}$ for simplicity.

Definition 7 (Interventional Family $\mathcal{I}$)**.** *For any DAG $\mathcal{G}$, we define the interventional family $\mathcal{I} = (I_1, I_2, \cdots, I_K)$. Here $I_1 := \emptyset$ corresponds to the pure observational setting. The joint distribution for the interventional family can be rewritten as:*

$$p^{(k)}(x_1, \cdots, x_{[2M+N]}) = \prod_{j \notin I_k} p_j^{(1)}(x_j | \boldsymbol{PA}^{\mathcal{G}}(x_j)) \prod_{j \in I_k} p_j^{(k)}(x_j | \boldsymbol{PA}^{\mathcal{G}}(x_j)) \tag{44}$$

Definition 8. *For a specific DAG $\mathcal{G}$, we define $\mathcal{M}(\mathcal{G})$ to be the set of strictly positive densities $p : \mathbb{R}^{2|\mathcal{S}|+|\mathcal{A}|} \to \mathbb{R}$ which satisfies:*

$$p(x_1, \cdots, x_{[2M+N]}) = \prod_{j \in [2M+N]} p_j(x_j | \boldsymbol{PA}^{\mathcal{G}}(x_j)) \tag{45}$$

where $\int_{\mathcal{X}_j} f_j(x_j | \boldsymbol{PA}^{\mathcal{G}}(x_j)) dx_j = 1$ for all $\boldsymbol{PA}^{\mathcal{G}}(x_j) \in \mathcal{X}_j$ and all $j \in [2M+N]$.

Definition 9. *For a specific DAG $\mathcal{G}$ and an interventional family $\mathcal{I}$, we define*

$$\mathcal{M}_{\mathcal{I}}(\mathcal{G}) := \{ [p^{(k)}]_{k\in[K]} \mid \forall k \in [K], p^{(k)} \in \mathcal{M}(\mathcal{G}), \forall j \notin I_k, p_j^{(k)}(x | \boldsymbol{PA}^{\mathcal{G}}(x)) = p_j^{(1)}(x | \boldsymbol{PA}^{\mathcal{G}}(x)) \} \tag{46}$$

Such set of functions is conherent with condition of strictly positive densities in (45) as well as factorization of interventional distribution in (44).

Definition 10 ($\mathcal{I}$-Markov Equivalence Class, $\mathcal{I}$-MEC)**.** *Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are $\mathcal{I}$-Markov equivalent iff $\mathcal{M}_{\mathcal{I}}(\mathcal{G}_1) = \mathcal{M}_{\mathcal{I}}(\mathcal{G}_2)$. We denote by $\mathcal{I} - MEC(\mathcal{G}_1)$ the set of all DAGs which are $\mathcal{I}$-Markov equivalent to $\mathcal{G}_1$, this is the $\mathcal{I}$-Markov equivalence class of $\mathcal{G}_1$.*

Lemma 6 (Sufficient and Necessary Conditions for $\mathcal{I}$-MEC, Yang el. al. [93])**.** *Suppose the interventional family $\mathcal{I}$ is such that $\mathcal{I}_1 := \emptyset$. Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are $\mathcal{I}$-Markov equivalent iff their I-DAGs $\mathcal{G}_{\mathcal{I}_1}$ and $\mathcal{G}_{\mathcal{I}_2}$ share the same skeleton and v-structures.*

Proof of Proposition 2. In the bipartite graph $(\mathcal{U}, \mathcal{V}, E)$, for the discovered graph $\hat{\mathcal{G}}$ that is in the $\mathcal{I}$-Markov equivalence class of the ground truth causal graph, $\hat{\mathcal{G}}$ is unique.

Based on the Lemma 6, all possible $\hat{\mathcal{G}}$ that are $\mathcal{I}$-Markov equivalent will share an identical skeleton with G^*, so we consider only graphs obtained by reversing edges in $\hat{\mathcal{G}}$.

Due to the bipartite nature of the transition causal graph defined in Definition 3, for all the v-structured colliders $c \in \mathcal{C}$, we know that $c \in \mathcal{S}'$, therefore, reversing any edge of $\hat{\mathcal{G}}$ will harm the immorality of $\hat{\mathcal{G}}$, and the new graph will no longer be an $\mathcal{I}$-MEC to $\mathcal{G}^*$ Therefore, $\hat{\mathcal{G}}$ is the only graph in the $\mathcal{I}$-MEC of $\mathcal{G}^*$, i.e. $\hat{\mathcal{G}} = \mathcal{G}^*$. ■

A.4.4 Causal Discovery Benefits from Policy Learning

In this section, we would like to show how the learned GCRL model could aid the performance of causal discovery. Before we put the formal proof, we first list several assumptions which is quite common in causal discovery literatures [94].

Assumption 4 (Oracle Conditional Independence Test)**.** *The conditional independent test could tell the independence between any two variables in the causal graph.*

Proof of Lemma 4. Given the oracle conditional independence test in appendix 4, if the state distribution covers all the support of goal distribution, with abundant actions from action space, we can cover all the connections between the current state and the next states.

When $\mathbb{D}_{\text{TV}}(p_{\mathcal{I}^s_\pi}, p_g) < \epsilon_g$, it is sufficient to derive that $p_{\mathcal{I}^s_\pi}(s) > 0, \forall s \in \text{Supp}_g$, where Supp_g is the support set of the goal distribution p_g.

We then discuss the three possible circumstances under such condition:

- Case 1: Both source state and target state distribution in the buffer are fully supported on Supp_g. In this case, given our Assumption 4 and abundant samples, our causal discovery $q_\phi(\mathcal{G}|\tau)$ will correctly classify all the edges in the transition causal graph.
- Case 2: Only the target state distribution is supported on Supp_g, while the source side leaves away from the goal node. In this case, the independent tests may not be able to distinguish the (in)dependence relationship between goal nodes and other state nodes, $\text{SHD}(\hat{\mathcal{G}}, \mathcal{G}^*) \leq |\mathcal{S}| - 1$.
- Case 3: Only source state is supported on Supp_g, while the target side leaves away from the goal node. This case corresponds to the case where some initial states hit the goal, while the learned transition model and policy fail to guide the future states to the goal. The causal discovery model ϕ may falsely classify the edges from all the source states (except the source goal state) towards the target goal states. Thus $\text{SHD}(\hat{\mathcal{G}}, \mathcal{G}^*) \leq |\mathcal{S}| - 1$.

In conclusion, for all the three cases that satisfy $\mathbb{D}_{\text{TV}}(p_{\mathcal{I}^s_\pi}, p_g) \leq \epsilon_g$, we have

$$\max_{\hat{\mathcal{G}} \sim q_\phi(\cdot|\tau)} \Big[\text{SHD}(\hat{\mathcal{G}}, \mathcal{G}^*)\Big] \leq |\mathcal{S}| - 1 \tag{47}$$

thus

$$\mathbb{E}_{\hat{\mathcal{G}} \sim q_\phi(\cdot|\tau)} \Big[\text{SHD}(\hat{\mathcal{G}}, \mathcal{G}^*)\Big] \leq \max_{\hat{\mathcal{G}} \sim q_\phi(\cdot|\tau)} \Big[\text{SHD}(\hat{\mathcal{G}}, \mathcal{G}^*)\Big] \leq |\mathcal{S}| - 1 \tag{48}$$

■

A.5 Overall Performance Guarantee of Iterative Optimization

Based on all the derivation from previous sections, we finally give out the proof of the overall performance of our proposed iterative optimization in *GRADER*.

Proof of Theorem 1. Let $d_{max} = \max_{s_1,s_2\in\mathcal{S}} \|s_1 - s_2\|^2$, $d_\theta = \|\hat{s}'(\theta) - s'\|^2$, then the log likelihood term becomes

$$p_\theta(s'|s,a) \propto \exp(d_{max} - d_\theta) \tag{49}$$

In the model learning part, since we take the log space, we have $\log p_\theta(s'|s,a) = (d_{max} - d_\theta) - C$. We neglect the constant term C when deriving the bound. Without the loss of generality, we set $p_\theta(s'|s,a) = \exp(d_{max} - d_\theta), \log p_\theta(s'|s,a) = d_{max} - d_\theta$ in (3). As $d_{max} - d_\theta \geq 0$, we have the Lipchitz $L \leq 1$ of log function,

$$\begin{aligned} \Big\| \log \hat{p}(s'|s,a) - \log p(s'|s,a) \Big\|_\infty &\leq L \Big\| \hat{p}(s'|s,a) - p(s'|s,a) \Big\|_\infty \\ &\leq \Big\| \hat{p}(s'|s,a) - p(s'|s,a) \Big\|_\infty \leq \epsilon_m \end{aligned} \tag{50}$$

Based on Lemma 2 that is derived in Appendix A.3, we have the policy learning term

$$\begin{aligned} \Big\| \log \hat{\pi}(a|s,g) - \log \pi^*(a|s,g) \Big] \Big\|_\infty &= \Big\| \hat{Q}(s,\hat{\pi}(s)) - Q(s,\pi^*(s)) \Big\|_\infty \\ &= \Big\| Q(s,\hat{\pi}(s)) - Q(s,\pi^*(s)) \Big\|_\infty \\ &= \Big\| V^{\hat{\pi}}(s) - V^{\pi^*}(s) \Big\|_\infty \\ &\leq \frac{\gamma}{(1-\gamma)^2} \epsilon_m \end{aligned} \tag{51}$$

For the KL divergence term, if the goal distribution satisfies $\epsilon_g > \frac{\gamma}{1-\gamma}\epsilon_m$, the following conditions hold:

$$V^{\hat{\pi}}(s) > V^{\pi^*}(s) - \frac{\gamma}{(1-\gamma)^2}\epsilon_m \tag{52}$$

According to Lemma 3 and the condition that $V(s) \in [0, \frac{1}{1-\gamma}]$,

$$\begin{aligned} \mathbb{D}_{\text{TV}}(p_{\mathcal{I}_\pi^s}, p_g) &\leq 1 - (1-\gamma)V^{\hat{\pi}}(s) \\ &< 1 - (1-\gamma)V^{\pi^*}(s) + \frac{\gamma}{1-\gamma}\epsilon_m \\ &= (1-\gamma^{t^*-1}) + \epsilon_g \overset{t^*=1}{=} \epsilon_g \end{aligned} \tag{53}$$

where t^* is the shortest time step to reach the goal. Here we assume $t^* = 1$ for optimal policy in the theoretical design part, while in practice, the bound may get loosened when larger t^* or smaller γ.

Since $\mathbb{D}_{\text{TV}}(p_{\mathcal{I}_\pi^s}, p_g) \leq \epsilon_g$, according to Lemma 4 proved in Appendix A.4, we have

$$\begin{aligned} \Big\| \mathbb{D}_{\text{KL}}(q_\phi||p) - \mathbb{D}_{\text{KL}}(q_\phi^*||p) \Big\|_\infty &= \Big\| \log\left(\frac{1-\epsilon_\mathcal{G}}{\epsilon_\mathcal{G}}\right) \|q_\phi(\mathcal{G}|\tau)\|_1 - \log\left(\frac{1-\epsilon_\mathcal{G}}{\epsilon_\mathcal{G}}\right) \|q_\phi^*(\mathcal{G}|\tau)\|_1 \Big\|_\infty \\ &= \log\left(\frac{1-\epsilon_\mathcal{G}}{\epsilon_\mathcal{G}}\right) \Big\| \|q_\phi(\mathcal{G}|\tau)\|_1 - \|q_\phi^*(\mathcal{G}|\tau)\|_1 \Big\|_\infty \\ &\leq \log\left(\frac{1-\epsilon_\mathcal{G}}{\epsilon_\mathcal{G}}\right) \Big\| q_\phi(\mathcal{G}|\tau) - q_\phi^*(\mathcal{G}|\tau) \Big\|_\infty \\ &= \log\left(\frac{1-\epsilon_\mathcal{G}}{\epsilon_\mathcal{G}}\right) \max_\mathcal{G} \Big[\text{SHD}(\mathcal{G}, \mathcal{G}^*) \Big] \\ &\leq \log\left(\frac{1-\epsilon_\mathcal{G}}{\epsilon_\mathcal{G}}\right) \Big(|\mathcal{S}| - 1 \Big) \end{aligned} \tag{54}$$

Finally, we can derive the overall performance guarantee as follows:

$$
\begin{aligned}
\|\mathcal{J}^*(\theta,\phi) - \hat{\mathcal{J}}(\hat{\theta},\hat{\phi})\|_\infty &= \Big\| \sum_{t=0}^{T-1} \Big\{ \Big[\log \hat{p}(s^{t+1}|s^t,a^t) - \log p(s^{t+1}|s^t,a^t) \Big] \\
&\quad + \Big[\log \hat{\pi}(a^t|s^t,s^*) - \log \pi^*(a^t|s^t,s^*) \Big] \Big\} + \Big[\mathbb{D}_{\mathrm{KL}}(\hat{q}_\phi||p) - \mathbb{D}_{\mathrm{KL}}(q^*_\phi||p) \Big] \Big\|_\infty \\
&\le \sum_{t=0}^{T-1} \Big\{ \Big\| \log \hat{p}(s^{t+1}|s^t,a^t) - \log p(s^{t+1}|s^t,a^t) \Big\|_\infty \\
&\quad + \Big\| \log \hat{\pi}(a^t|s^t,s^*) - \log \pi^*(a^t|s^t,s^*) \Big\|_\infty \Big\} + \Big\| \mathbb{D}_{\mathrm{KL}}(\hat{q}_\phi||p) - \mathbb{D}_{\mathrm{KL}}(q^*_\phi||p) \Big\|_\infty \\
&\le \sum_{t=0}^{T-1} \Big(\epsilon_m + \frac{\gamma}{(1-\gamma)^2}\epsilon_m \Big) + \log\Big(\frac{1-\epsilon_{\mathcal{G}}}{\epsilon_{\mathcal{G}}} \Big) \Big(|\mathcal{S}| - 1 \Big) \\
&= \Big[1 + \frac{\gamma}{(1-\gamma)^2} \Big] \epsilon_m T + \log\Big(\frac{1-\epsilon_{\mathcal{G}}}{\epsilon_{\mathcal{G}}} \Big) \Big(|\mathcal{S}| - 1 \Big)
\end{aligned}
\tag{55}
$$

■

B Additional Experiment

B.1 Overall Performance

The overall performance results corresponding to the Table 1 for *Stack* and *Unlock* environments are shown in Figure 7 and Figure 8.

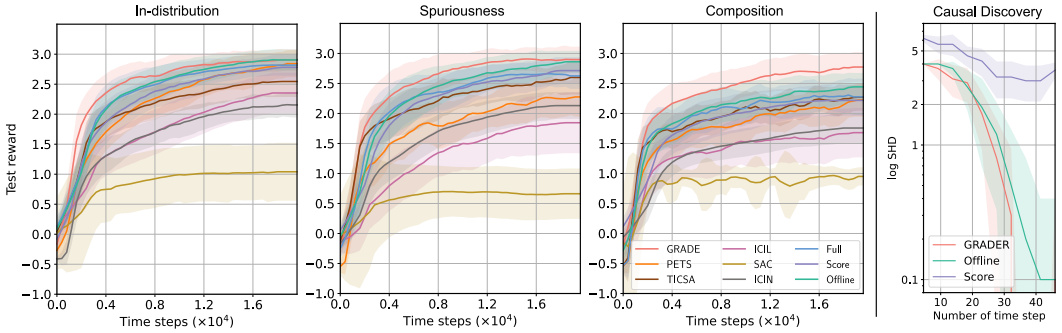

Figure 7: The testing reward and causal discovery results of *Stack* environment.

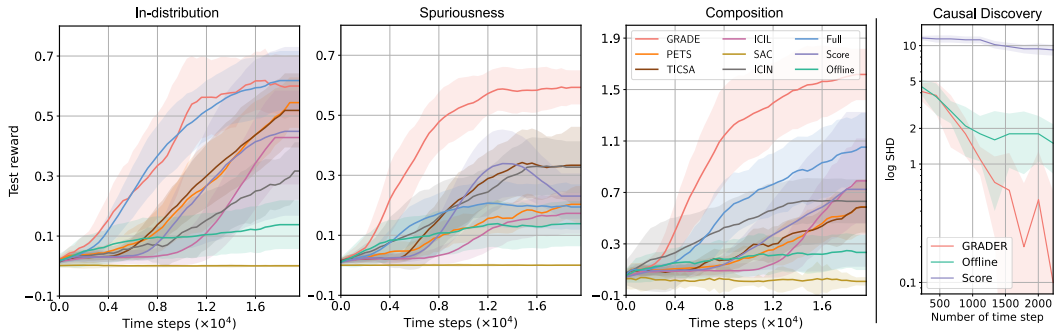

Figure 8: The testing reward and causal discovery results of *Unlock* environment.

In all *Stack* experiments, we find that the advantage of GRADER over other methods is small. The reason is that this task is simple and the true causal graph only contains 7 nodes as shown in

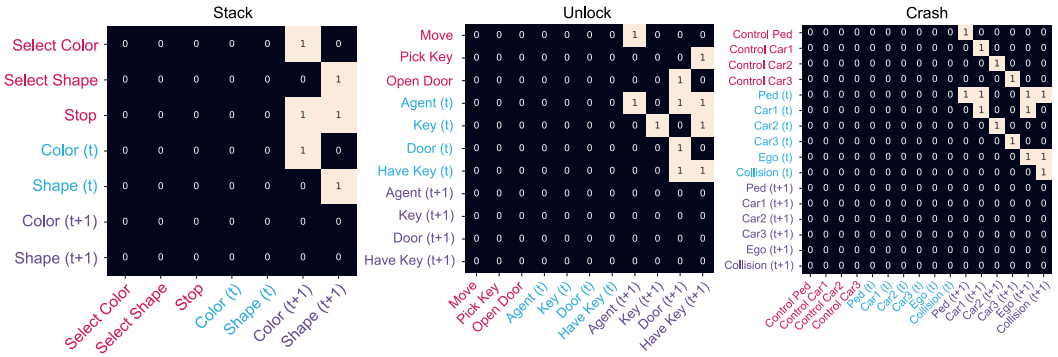

Figure 9: Discovered causal graphs of three environments. Color meaning: Action, State, Next state.

Figure B.2. Due to the simple causal graph, even the Offline random policy can obtain the true causal graph, thus there is almost no difference between the discovery efficiency between GRADER and Offline as shown in the right part of Figure 7.

In the *Unlock-I* experiment, there is no gap between GRADER and Full, which means the causal graph may not have many contributions to solving this task. However, there are large gaps in *Unlock-S* and *Unlock-C* settings since indicating that the causal graph helps the model obtain better generalizable performance. As for the Offline method, since the causal graph is wrongly discovered, the performance is bad in all three settings.

B.2 Causal Graph Analysis

Since the environments we designed have clear and explicit causality, we can get the true causal graph with human analysis. We plot the true causal graphs corresponding to the three environments in Figure 9, where the semantic meanings of all nodes are explained in Appendix C.2.1. We observe that the causal graphs are sparse with very few edges, indicating that non-causal methods that use the full graph may import redundant or even wrong information.

B.3 Distance between Goal and State Distribution

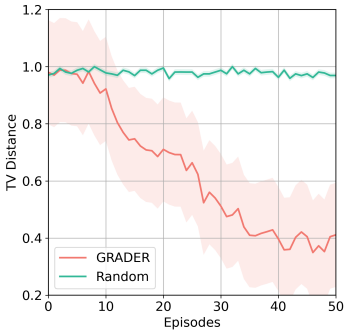

Figure 10: TV distance between goal and state distributions.

In Figure 10, we empirically show the upper bound proved in (27), which describes the TV distance between the goal distribution and the state distribution collected from the GRADER policy. We use 10 trails and plot the mean and standard deviation of the distance. We observe that the distance becomes smaller as the policy gets better in GRADER. This supports our statement that the planning module helps to collect better data samples, which will be used in the causal discovery module. We also plot the distance with a random policy, which is always large since the goal is not easy to be achieved by random actions.

B.4 Task Performance of Chemistry Experiment

In the main context, we only show the discovery results of the Chemistry experiment. We provide the results of task performance in this section. The downstream task is to change the color of nodes to match given colors within maximum steps ($T = 10$). A reward $r = 1$ is received if all colors are matched. Results are reported with 200 episodes. We use planning horizon $H = 5$. We provide the RL downstream task results in Table 3 (ID setting) and Table 4 (OOD setting). The testing reward is shown in Figure 11. The graphs in the ID setting have 10 nodes while those in the OOD setting have 5 nodes. In the ID setting, we randomly sample the target colors in the goal. In the OOD setting, we set the target colors of all nodes to the same color during the training to create spurious correlations, then randomly set the target colors during testing.

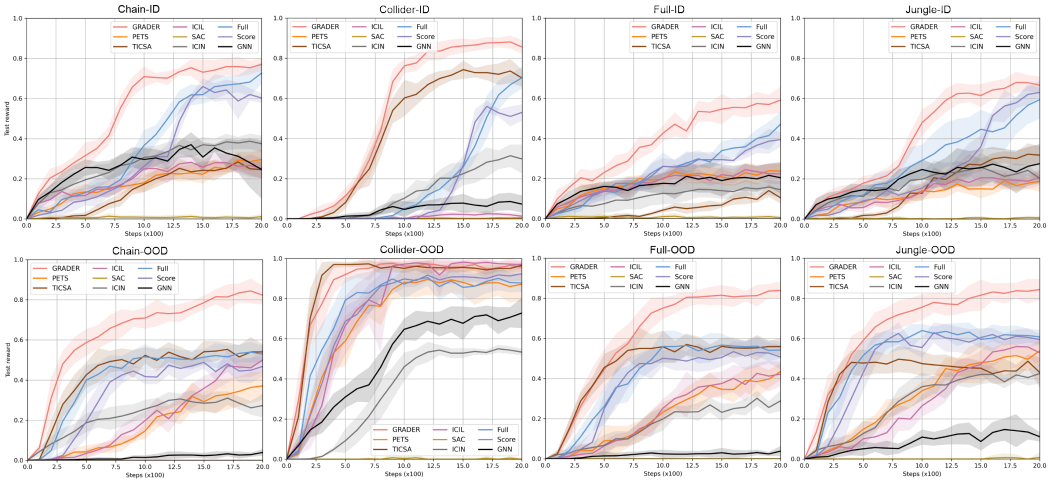

Figure 11: Reward of Chemistry environment under ID and OOD setting.

Table 3: Success rate (%) for Chemistry environment (ID). **Bold** font means the best.

Env	SAC	ICIN	PETS	TICSA	ICIL	GNN	GRADER	Score	Full
Collider	0.0±0.0	29.8±7.2	0.6±0.8	70.1±4.2	1.3±1.3	7.3±4.3	**85.5±3.4**	53.0±4.1	70.3±5.3
Chain	1.1±1.3	37.5±4.0	29.6±4.9	24.5±4.3	25.3±5.1	24.6±14.5	**77.0±3.2**	60.2±2.4	72.7±5.3
Jungle	0.6±0.8	20.2±1.5	18.8±5.2	31.8±4.5	20.6±3.9	27.5±9.8	**69.6±4.3**	63.0±2.3	59.4±9.5
Full	0.5±0.8	4.5±4.0	23.7±4.3	10.4±3.0	22.3±5.1	20.4±7.8	**59.1±6.6**	39.4±3.5	47.1±6.1

C Additional Information

C.1 Details about Conditional Independence Test

In Algorithm 1, we describe the discovery of causal graph with edge inference $e_{ij} \leftarrow q_\phi(\cdot|\mathcal{B}, \eta)$ implemented by conditional independent test. We ignore the details about the test process in the main context and thus provide more details in this section.

For discrete variables, we use Pearson's chi-square test[1], which is a statistical test applied to sets of categorical data to evaluate how likely it is that any observed difference between the sets arose by chance. In our experiment, we use the implementation provided by package Scipy [2].

We first define the null hypothesis, which is true when two random variables are statistically independent. These two variables have samples stored in a contingency table O, which has c columns and r rows. Then, the "theoretical frequency" for a cell is:

$$E_{ij} = N_{p_{i\cdot}p_{\cdot j}}, \quad p_{i\cdot} = \sum_{j=1}^{c} \frac{O_{i,j}}{N}, \quad p_{i\cdot} = \sum_{i=1}^{r} \frac{O_{i,j}}{N} \tag{56}$$

where N is the total sample size in the table, $O_{i,j}$ is the sample size of cell (i, j). Then, we can calculate the value of the test statistic:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \tag{57}$$

Now, we can obtain a p-value (falls in $[0, 1]$) that indicates the significance of this statistic follows the χ^2 distribution from chi-square probability[3]. We compare this p-value with a threshold η and reject

[1] https://en.wikipedia.org/wiki/Pearson%27s_chi-squared_test

[2] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2_contingency.html

[3] https://people.richland.edu/james/lecture/m170/tbl-chi.html

Table 4: Success rate (%) for Chemistry environment (OOD). **Bold** font means the best.

Env	SAC	ICIN	PETS	TICSA	ICIL	GNN	GRADER	Score	Full
Collider	0.0±0.0	53.3±1.6	87.2±8.5	96.6±1.4	97.0±2.0	72.8±7.5	**95.8±2.6**	92.4±3.5	87.8±4.4
Chain	0.0±0.0	27.3±5.9	37.1±7.0	54.0±3.8	50.0±5.8	3.9±1.6	**82.3±4.5**	46.8±5.0	52.9±4.3
Jungle	0.8±2.4	42.6±4.9	53.9±5.5	43.1±7.5	52.9±7.0	11.1±2.4	**84.4±5.1**	59.5±2.7	60.8±3.5
Full	0.0±0.0	28.9±5.0	43.5±4.1	55.9±4.5	42.2±5.9	3.8±2.5	**83.9±4.4**	50.7±6.0	54.2±4.1

the null hypothesis if the p-value is smaller than η. Therefore, the larger we set η, the more likely we find the two variables are dependent. This testing process is summarized in Algorithm 2.

If the two variables are continuous, we cannot use the above statistical test anymore. We turn to a more advanced test method proposed in [40]. The general idea is that if $P(X|Y,Z) = P(X,Y)$, Z is not useful as a feature to predict X. To achieve this, the authors propose to use decision tree regression to predict Y using both X and Z, and also using Z only.

Algorithm 2: Independence Test for Discrete Variables.

Input: A contingency table O with samples for two variables X and Y.
Define Null hypothesis: X and Y are independent.
Calculate $p_{i\cdot} = \sum_{j=1}^{c} \frac{O_{i,j}}{N}$ and $p_{i\cdot} = \sum_{i=1}^{r} \frac{O_{i,j}}{N}$
Calculate expected frequencies $E_{ij} = N_{p_{i\cdot} p_{\cdot j}}$
Calculate the chi-square statistic $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$
Obtain p-value p from chi-square probability
if $p < \eta$ **then**
  Reject the Null hypothesis, i.e., X and Y are dependent.

C.2 Experiment Details

C.2.1 Environment Design

More details about the design of the environments are summarized below:

Stack: Manipulation is important for house-holding and factory assembly. Sometimes, the color of the object is not relevant to the task but may leak information by sharing spuriousness with the task. Also, the goal could compose several previously seen goals such as repeating similar actions. Totally, we have 5 different shapes and 5 different colors and the goals are some combinations of the colors and shapes. At each step, the agent can either stack an object with a chosen color and shape or stop stacking. The state is the colors and shapes of all current objects. The agent receives a positive reward when the goal is achieved and a punishment if it stacks a new object.

Unlock: Collecting specific objects to fulfill required conditions is useful for mobile robots. The causality in this environment exists between the key and the door. The action contains six operations, including four-direction movements (Move), pick key (Pick Key), and open door (Open Door). The state is the position of the agent, the position of the key, and the status of the door. In the first generalization setting, we intentionally create a spurious correlation between the position of the key and the door. If the agent figures out that key can open the door no matter what its position is, it will ignore the spurious correlation. In the second generalization setting, we increase the number of door from one to two. This setting contains two same sub-tasks and can be used to test the compositional generalization.

Crash: The causality in this environment mainly exists between the pedestrian (Ped), the ego vehicle (Ego), and another vehicle (Car 1) [95]. The collision between Ped and Ego only happens when the view of Ego is blocked by Car 1. To make this happen, we design a rule-based AV, which will brake if it detects any obstacles within a certain distance. Therefore, if the pedestrian directly hits the AV, the AV will stop and the crash will not happen. To make this task difficult, we also place two other vehicles (Car 2 and Car 3) on the scene but they will not interrupt the crash scenario. The agent can control the acceleration and steering of Ped, Car 1, Car 2, and Car 3. The state is the position and velocity of all objects plus the status of whether a collision happens. To create a spurious correlation, we fix the initial distance between Ego and Ped to a constant since this creates a shortcut for the

feature extractor. However, remembering this distance is not enough since we change the initial distance in the testing stage.

Chemistry: Please refer to [43] for more details.

Table 5: Environment configurations used in experiments

Parameters	Stack	Unlock	Crash	Chemistry
Max step size	5	15	30	10
State dimension	50	110	22	100
Action dimension	12	8	8	100
Action type	Discrete	Discrete	Continuous	Discrete

C.2.2 Model Structures and Hyper-parameters

Since different nodes have different dimensions, we design a set of encoders E_j and a set of decoders D_j to convert the dimension of features. Thus, the entire model structure is

$$s_j^{t+1} = D_j(f_{\theta_j}(E_j([\mathbf{PA}_j^{\mathcal{G}}]^t, N_j))), \quad \forall j \in [M] \tag{58}$$

We list all important hyper-parameters in the implementation for three environments in Table 6.

C.3 Broader Social Impact and Additional Limitation

C.3.1 Broader Social Impact

We identify several important social impacts of our proposed method, including both positive and potential negative impacts:

1) Incorporating causality into reinforcement learning methods increases both the interpretability and generalizability of artificial intelligence, which helps users easily check the working progress of agents and the source of failures.
2) Insufficient data and training may cause flawed causal graphs, which may lead to a wrong understanding of the causation of the task. This wrong understanding of the task may cause risky and irrational actions of agents.
3) The discovered causal graph could be accessed and modified by users to manipulate the behaviors of agents on purpose. If the task contains private information, the discovered causal graph may cause privacy issues when the graph is interpreted by other users.

To mitigate the potential negative societal impacts mentioned above, we encourage research to follow these instructions:

1) People should always check the convergence of the causal discovery step and verify the discovered causal graph with domain knowledge.
2) The discovered causal graph should be frequently checked and verified with the training data to ensure its correctness. The causal graphs also need to be encrypted and only accessible to algorithms and trustworthy users.

C.3.2 Additional Limitation

Causal discovery methods. The gradient-based discovery method are widely investigated recently for large datasets since they have good scalability. However, these methods also require lots of training data to converge. In Online RL, we don't have enough data at the beginning of training. Thus, constraint-based methods are more suitable for causal RL tasks.

Although our constraint-based causal discovery does not scale as well as the score-based methods, our proposed independence tests achieve a time complexity of $\Omega(|S|(|S|+|A|))$, which is tolerable for most RL problems with lower dimensional state space. Empirical studies also show that our independent tests enjoy better data efficiency.

Table 6: Hyper-parameters of models used in experiments

Models	Parameters	Environment			
		Stack	Unlock	Crash	Chemistry
GRADER	Learning rate	0.001	0.001	0.0001	0.001
	Size of buffer $\mathcal{B}$	4000	10000	10000	4000
	Epoch per iteration	20	5	10	20
	Batch size	256	256	256	256
	Planning horizon H	5	10	20	5
	Planning population	500	100	1000	700
	Reward discount γ	0.99	0.99	0.99	0.99
	ϵ-greedy ratio	0.4	0.4	0.5	0.5
	Causal Discovery η	0.01	0.01	0.01	0.01
	GRU hiddens	32	64	128	32
PETS*	MLP hiddens	32	64	128	32
	MLP layers	2	2	2	2
	Ensemble number	5	5	5	5
TICSA*	Size of buffer $\mathcal{B}$	20000	400000	40000	20000
	Pretrain buffer	200	2000	5000	200
	Initialized mask coef.	1.0	1.0	1.0	1.0
	MLP hiddens	32	64	128	32
	Sparsity regularizer	0.5	1.0	0.2	0.5
ICIL*	Size of buffer $\mathcal{B}$	20000	400000	40000	20000
	Learning rate of MINE	0.0001	0.0001	0.0001	0.0001
	MLP hiddens	32	64	128	32
	MINE hiddens	32	64	128	32
	Env. Numbers	5	3	3	5
SAC	Learning rate	0.001	0.001	0.0001	0.001
	Size of buffer $\mathcal{B}$	4000	10000	10000	4000
	Update step τ	0.005	0.005	0.0001	0.005
	Update iteration	3	3	3	3
	Entropy α	0.2	0.2	0.2	0.2
	Batch size	256	256	256	256
	Reward discount γ	0.99	0.99	0.99	0.99
	MLP hiddens	64	128	256	64
ICIN	Learning rate	0.001	0.001	0.0001	0.001
	Size of buffer $\mathcal{B}$	4000	10000	10000	4000
	Batch size	256	256	256	256
	MLP hiddens	64	128	256	64
	MLP layers	3	3	3	3

* Use the same planning parameters as GRADER.

Assumptions in our theoretical analysis. Faithfulness and Markov properties are commonly used in causal discovery literature such as [54, 39]. It is claimed in [39] that the oracle independent test can be ensured by the satisfaction of Markov property and faithfulness. Recent work [94] also assumes the oracle of conditional independent test in its Assumption 2.1. Practically, the oracle test can be implemented with certain sub-linear sample complexity, as is investigated in [41].

In reinforcement learning tasks, agents interact with the environment by doing interventions, which is achieved by assigning values to action nodes. Then, the intervention results are reflected by the states. Under fully observable Markov settings, the value of these states contains all information about the intervention. Thus, our RL setting usually satisfies the assumptions we use in the theoretical proof.