Fine-Grained Causal Dynamics Learning with Quantization for Improving Robustness in Reinforcement Learning

Inwoo Hwang[1] Yunhyeok Kwak[1] Suhyung Choi[1] Byoung-Tak Zhang[1] Sanghack Lee[1 2]

Abstract

Causal dynamics learning has recently emerged as a promising approach to enhancing robustness in reinforcement learning (RL). Typically, the goal is to build a dynamics model that makes predictions based on the causal relationships among the entities. Despite the fact that causal connections often manifest only under certain contexts, existing approaches overlook such fine-grained relationships and lack a detailed understanding of the dynamics. In this work, we propose a novel dynamics model that infers fine-grained causal structures and employs them for prediction, leading to improved robustness in RL. The key idea is to jointly learn the dynamics model with a discrete latent variable that quantizes the state-action space into subgroups. This leads to recognizing meaningful context that displays sparse dependencies, where causal structures are learned for each subgroup throughout the training. Experimental results demonstrate the robustness of our method to unseen states and locally spurious correlations in downstream tasks where fine-grained causal reasoning is crucial. We further illustrate the effectiveness of our subgroup-based approach with quantization in discovering fine-grained causal relationships compared to prior methods.

1. Introduction

Model-based reinforcement learning (MBRL) has showcased its capability of solving various sequential decision making problems (Kaiser et al., 2020; Schrittwieser et al., 2020). Since learning an accurate and robust dynamics model is crucial in MBRL, recent works incorporate the causal relationships between the environmental variables, such as objects and the agent, into dynamics learning (Wang et al., 2022; Ding et al., 2022). Unlike the traditional dense models that employ the whole state and action variables to predict the future state, causal dynamics models infer the causal structure of the transition dynamics and make predictions based on it. Consequently, they are more robust to unseen states by discarding spurious dependencies.

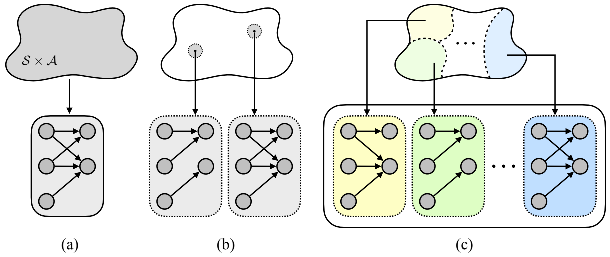


Figure 1. (a) Previous causal dynamics models infer the global causal structure of the transition dynamics. (b) Existing approaches to discovering fine-grained relationships examine individual samples. (c) Our approach quantizes the state-action space into subgroups and infers causal relationships specific to each subgroup.

Our motivation stems from the observation that causal connections often manifest only under certain contexts in many practical scenarios. Consider autonomous driving, where recognizing the traffic signal is crucial for its safety (e.g., stops at red lights). However, in the presence of a pedestrian on the road, it must stop, even with a green light, ignoring the signal, i.e., the traffic signal becomes *locally spurious*. Therefore, such fine-grained causal reasoning will be crucial to the robustness of MBRL for its real-world deployment.

Fine-grained causal relationships can be understood with local independence between the variables, which holds under certain contexts but does not hold in general (Boutilier et al., 2013). Our goal is to incorporate them into dynamics modeling by capturing meaningful contexts that exhibit more sparse dependencies than the entire domain. Unfortunately, prior causal dynamics models examining global independence (Fig. 1-(a)) cannot harness them. On the other hand, existing methods for discovering fine-grained relationships have focused on examining sample-specific dependencies (Pitis et al., 2020; Hwang et al., 2023) (Fig. 1-(b)). However,

[1]AI Institute, Seoul National University [2]Graduate School of Data Science, Seoul National University. Correspondence to: Sanghack Lee <sanghack@snu.ac.kr>, Byoung-Tak Zhang <btzhang@snu.ac.kr>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

it is unclear under which circumstances the inferred dependencies hold, making them hard to interpret and challenging to generalize to unseen states.

In this work, we propose a dynamics model that infers fine-grained causal structures and employs them for prediction, leading to improved robustness in MBRL. For this, we establish a principled way to examine fine-grained causal relationships based on the quantization of the state-action space. Importantly, this provides a clear understanding of meaningful contexts displaying sparse dependencies (Fig. 1-(c)). However, this involves the optimization of the regularized maximum likelihood score over the quantization which is generally intractable. To this end, we present a practical differentiable method that jointly learns the dynamics model and a discrete latent variable that decomposes the state-action space into subgroups by utilizing vector quantization (Van Den Oord et al., 2017). Theoretically, we show that joint optimization leads to identifying meaningful contexts and fine-grained causal structures.

We evaluate our method on both discrete and continuous control environments where fine-grained causal reasoning is crucial. Experimental results demonstrate the superior robustness of our approach to locally spurious correlations and unseen states in downstream tasks compared to prior causal/non-causal approaches. Finally, we illustrate that our method infers fine-grained relationships in a more effective and robust manner compared to sample-specific approaches.

Our contributions are summarized as follows.

- We establish a principled way to examine fine-grained causal relationships based on the quantization of the state-action space which offers an identifiability guarantee and better interpretability.
- We present a theoretically grounded and practical approach to dynamics learning that infers fine-grained causal relationships by utilizing vector quantization.
- We empirically demonstrate that the agent capable of fine-grained causal reasoning is more robust to locally spurious correlations and generalizes well to unseen states compared to past causal/non-causal approaches.

2. Preliminaries

We first introduce the notations and terminologies. Then, we examine related works on causal dynamics learning for RL and fine-grained causal relationships.

2.1. Background

Structural causal model. We adopt a framework of a structural causal model (SCM) (Pearl, 2009) to understand the relationship among variables. An SCM $\mathcal{M}$ is defined as a tuple $\left\langle \mathbf{V}, \mathbf{U}, \mathbf{F}, P(\mathbf{U}) \right\rangle$, where $\mathbf{V} = \{X_1, \cdots, X_d\}$ is a set of endogenous variables and $\mathbf{U}$ is a set of exogenous variables. A set of functions $\mathbf{F} = \{f_1, \cdots, f_d\}$ determine how each variable is generated; $X_j = f_j(Pa(j), \mathbf{U}_j)$ where $Pa(j) \subseteq \mathbf{V} \setminus \{X_j\}$ is parents of X_j and $\mathbf{U}_j \subseteq \mathbf{U}$. An SCM $\mathcal{M}$ induces a directed acyclic graph (DAG) $\mathcal{G} = (V, E)$, i.e., a causal graph (CG) (Peters et al., 2017), where $V = \{1, \cdots, d\}$ and $E \subseteq V \times V$ are the set of nodes and edges, respectively. Each edge denotes a direct causal relationship from X_i to X_j. An SCM entails the conditional independence relationship of each variable (namely, local Markov property): $X_i \perp\!\!\!\perp ND(X_i) \mid Pa(X_i)$, where $ND(X_i)$ is a non-descendant of X_i, which can be read off from the corresponding causal graph.

Factored Markov Decision Process. A Markov Decision Process (MDP) (Sutton & Barto, 2018) is defined as a tuple $\langle \mathcal{S}, \mathcal{A}, T, r, \gamma \rangle$ where $\mathcal{S}$ is a state space, $\mathcal{A}$ is an action space, $T : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{P}(\mathcal{S})$ is a transition dynamics, r is a reward function, and γ is a discount factor. We consider a factored MDP (Kearns & Koller, 1999) where the state and action spaces are factorized as $\mathcal{S} = \mathcal{S}_1 \times \cdots \times \mathcal{S}_N$ and $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_M$. A transition dynamics is factorized as $p(s' \mid s, a) = \prod_j p(s'_j \mid s, a)$ where $s = (s_1, \cdots, s_N)$ and $a = (a_1, \cdots, a_M)$.

Assumptions and notations. We are concerned with an SCM associated with the transition dynamics in a factored MDP where the states are fully observable. To properly identify the causal relationships, we make assumptions standard in the field, namely, Markov property (Pearl, 2009), faithfulness (Peters et al., 2017), causal sufficiency (Spirtes et al., 2000), and that causal connections only appear within consecutive time steps. With these assumptions, we consider a bipartite causal graph $\mathcal{G} = (V, E)$ which consists of the set of nodes $V = \mathbf{X} \cup \mathbf{Y}$ and the set of edges $E \subseteq \mathbf{X} \times \mathbf{Y}$, where $\mathbf{X} = \{S_1, \cdots, S_N, A_1, \cdots, A_M\}$ and $\mathbf{Y} = \{S'_1, \cdots, S'_N\}$. $Pa(j)$ denotes parent variables of S'_j. The conditional independence

$$S'_j \perp\!\!\!\perp \mathbf{X} \setminus Pa(j) \mid Pa(j), \tag{1}$$

entailed by the causal graph $\mathcal{G}$ represents the causal structure of the transition dynamics $p(s' \mid s, a) = \prod_j p(s'_j \mid Pa(j))$.

Dynamics modeling. The traditional way is to use dense dependencies for dynamics modeling: $\prod_j p(s'_j \mid s, a)$. Causal dynamics models (Wang et al., 2021; 2022; Ding et al., 2022) examine the causal structure $\mathcal{G}$ to employ only relevant dependencies: $p(s' \mid s, a; \mathcal{G}) = \prod_j p(s'_j \mid Pa(j))$ (Fig. 1-(a)). Consequently, they are more robust to spurious correlations and unseen states.

2.2. Related Work

Causal dynamics models in RL. There is a growing body of literature on the intersection of causality and RL

(De Haan et al., 2019; Buesing et al., 2019; Zhang et al., 2020a; Sontakke et al., 2021; Schölkopf et al., 2021; Zholus et al., 2022; Zhang et al., 2020b). One focus is dynamics learning, which involves the causal structure of the transition dynamics (Li et al., 2020; Yao et al., 2022; Bongers et al., 2018; Wang et al., 2022; Ding et al., 2022; Feng et al., 2022; Huang et al., 2022) (more broad literature on causal reasoning in RL is discussed in Appendix A.1). Recent works proposed causal dynamics models that make robust predictions based on the causal dependencies (Fig. 1-(a)), utilizing conditional independence tests (Ding et al., 2022) or conditional mutual information (Wang et al., 2022) to infer the causal graph in a factored MDP. However, prior methods cannot harness fine-grained causal relationships that provide a more detailed understanding of the dynamics. In contrast, our work aims to discover and incorporate them into dynamics modeling, demonstrating that fine-grained causal reasoning leads to improved robustness in MBRL.

Discovering fine-grained causal relationships. In the context of RL, a fine-grained structure of the environment dynamics has been leveraged in various ways, e.g., with data augmentation (Pitis et al., 2022), efficient planning (Hoey et al., 1999; Chitnis et al., 2021), or exploration (Seitzer et al., 2021). For this, previous works often exploited domain knowledge (Pitis et al., 2022) or true dynamics model (Chitnis et al., 2021). However, such prior knowledge is often unavailable in the context of dynamics learning. Existing methods for discovering fine-grained relationships examine the gradient (Wang et al., 2023) or attention score (Pitis et al., 2020) of each sample (Fig. 1-(b)). However, such *sample-specific* approaches lack an understanding of under which circumstances the inferred dependencies hold, and it is unclear whether they can generalize to unseen states.

In the field of causality, fine-grained causal relationships have been widely studied, e.g., context-specific independence (Boutilier et al., 2013; Zhang & Poole, 1999; Poole, 1998; Dal et al., 2018; Tikka et al., 2019; Jamshidi et al., 2023) (see Appendix A.2 for the background). Recently, Hwang et al. (2023) proposed an auxiliary network that examines local independence for *each sample*. However, it also does not explicitly capture the context where the local independence holds. In contrast to existing approaches relying on sample-specific inference (Löwe et al., 2022; Pitis et al., 2020; Hwang et al., 2023), we propose to examine causal dependencies at a subgroup level through quantization (Fig. 1-(c)), providing a more robust and principled way of discovering fine-grained causal relationships with a theoretical guarantee.

3. Fine-Grained Causal Dynamics Learning

In this section, we first describe a brief background on local independence and intuition of our approach (Sec. 3.1). We then provide a principled way to examine fine-grained causal relationships (Sec. 3.2). Based on this, we propose a theoretically grounded and practical method for fine-grained causal dynamics modeling (Sec. 3.3). Finally, we provide a theoretical analysis with discussions (Sec. 3.4). All omitted proofs are provided in Appendix B.

3.1. Preliminary

Analogous to the conditional independence explaining the causal relationship between the variables (i.e., Eq. (1)), their fine-grained relationships can be understood with local independence (Hwang et al., 2023). This is written as:

$$S'_j \perp\!\!\!\perp \mathbf{X} \setminus Pa(j;\mathcal{D}) \mid Pa(j;\mathcal{D}), \mathcal{D}, \tag{2}$$

where $\mathcal{D} \subseteq \mathcal{X} = \mathcal{S} \times \mathcal{A}$ is a local subset of the joint state-action space, which we say *context*, and $Pa(j;\mathcal{D}) \subseteq \mathbf{X}$ is a set of state and action variables locally relevant for predicting S'_j under $\mathcal{D}$. We provide a formal definition and detailed background of local independence in Appendix A.3.

For example, consider a mobile home robot interacting with various objects ($Pa(j)$). Under the context of the *door closed* ($\mathcal{D}$), only objects within the same room ($Pa(j;\mathcal{D})$) become relevant. On the other hand, all objects remain relevant under the context of the *door opened*. We say that a context is *meaningful* if it displays sparse dependencies: $Pa(j;\mathcal{D}) \subsetneq Pa(j)$, e.g., *door closed*. We are concerned with the subgraph of the (global) causal graph $\mathcal{G}$ as a graphical representation of such local dependencies.

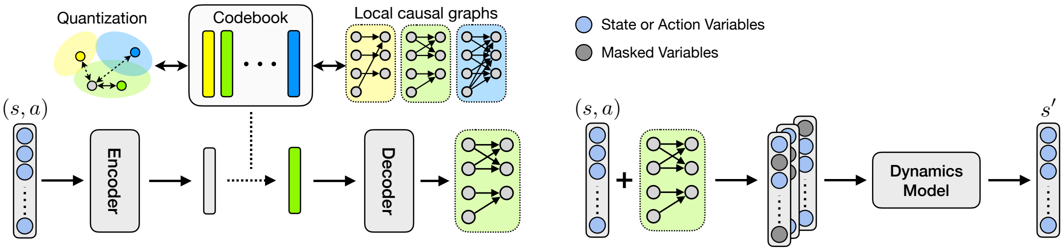
Definition 1. *Local subgraph of the causal graph*[1] *(LCG) on $\mathcal{D} \subseteq \mathcal{X}$ is $\mathcal{G}_{\mathcal{D}} = (V, E_{\mathcal{D}})$ where $E_{\mathcal{D}} = \{(i,j) \mid i \in Pa(j;\mathcal{D})\}$.*

LCG $\mathcal{G}_{\mathcal{D}}$ represents a causal structure of the transition dynamics specific to a certain context $\mathcal{D}$. It is useful for our approach to fine-grained dynamics modeling, e.g., it is sufficient to consider only objects in the same room when the door is closed. In contrast, prior causal dynamics models consider all objects under any circumstances (Fig. 1-(a)).

Importantly, such information (e.g., $\mathcal{D}$ and $\mathcal{G}_{\mathcal{D}}$) is not known in advance, and it is our goal to discover them. For this, existing sample-specific approaches have focused on inferring LCG directly from individual samples (Pitis et al., 2020; Hwang et al., 2023) (Fig. 1-(b)). However, it is unclear under which context the inferred dependencies hold.

Our approach is to quantize the state-action space into subgroups and examine causal structures on *each subgroup* (Fig. 1-(c)). This now makes it clear that each inferred LCG will represent fine-grained causal relationships under the corresponding subgroup. We now proceed to describe a principled way to discover LCGs based on quantization.

[1] For brevity, we will henceforth denote it as *local causal graph*.



(a) Local Causal Graph Inference

(b) Masked Prediction with LCG

Figure 2. Overall framework. (a) For each sample (s, a), our method determines the subgroup to which the sample belongs through quantization and infers the local causal graph (LCG) that represents fine-grained causal relationships specific to the corresponding subgroup. (b) The dynamics model predicts the future state based on the inferred LCG. All components (e.g., dynamics model and codebook) are jointly learned throughout the training in an end-to-end manner.

3.2. Score for Decomposition and Graphs

Let us consider *arbitrary* decomposition $\{\mathcal{E}_z\}_{z=1}^K$ of the state-action space $\mathcal{X}$, where K is the degree of the quantization. The transition dynamics can be decomposed as:

$$\begin{aligned} p(s'_j \mid s, a) &= \sum_z p(s'_j \mid s, a, z)p(z \mid s, a) \\ &= \sum_z p(s'_j \mid Pa(j; \mathcal{E}_z), z)p(z \mid s, a), \end{aligned} \quad (3)$$

where $p(z \mid s, a) = 1$ if $(s, a) \in \mathcal{E}_z$. This illustrates our approach to fine-grained dynamics modeling, employing only locally relevant dependencies according to $\mathcal{G}_{\mathcal{E}_z}$ on each subgroup $\mathcal{E}_z$. We now aim to learn each LCG $\mathcal{G}_{\mathcal{E}_z}$ based on Eq. (3). Specifically, we consider the regularized maximum likelihood score $\mathcal{S}(\{\mathcal{G}_z, \mathcal{E}_z\}_{z=1}^K)$ of the graphs $\{\mathcal{G}_z\}_{z=1}^K$ and decomposition $\{\mathcal{E}_z\}_{z=1}^K$ which is defined as:

$$\sup_\phi \mathbb{E}_{p(s,a,s')} \left[\log \hat{p}(s' \mid s, a; \{\mathcal{G}_z, \mathcal{E}_z\}, \phi) - \lambda |\mathcal{G}_z| \right], \quad (4)$$

where ϕ is the parameters of the dynamics model $\hat{p}$ which employs the graph $\mathcal{G}_z$ for prediction on corresponding subgroup $\mathcal{E}_z$. We now show that graphs that maximize the score faithfully represent causal dependencies on each subgroup.

Theorem 1 (Identifiability of LCGs)**.** *With Assumptions 1 to 4, let* $\{\hat{\mathcal{G}}_z\} \in \operatorname{argmax} \mathcal{S}(\{\mathcal{G}_z, \mathcal{E}_z\}_{z=1}^K)$ *for* $\lambda > 0$ *small enough. Then, each* $\hat{\mathcal{G}}_z$ *is true LCG on* $\mathcal{E}_z$*, i.e.,* $\hat{\mathcal{G}}_z = \mathcal{G}_{\mathcal{E}_z}$*.*

Given the subgroups, corresponding LCGs can be recovered by score maximization. Therefore, it provides a principled way to discover LCGs, which is valid for *any* quantization.

Unfortunately, not all quantization is useful for fine-grained dynamics modeling, e.g., by dividing into *lights on* and *lights off*, it still needs to consider all objects under both circumstances. Thus, it is crucial for quantization to capture *meaningful* contexts displaying sparse dependencies. Such useful quantization will allow more sparse dynamics modeling, i.e., the higher score of Eq. (4). Therefore, the decomposition is now also a learning objective towards maximizing Eq. (4), i.e., $\{\mathcal{G}_z^*, \mathcal{E}_z^*\} \in \operatorname{argmax} \mathcal{S}(\{\mathcal{G}_z, \mathcal{E}_z\}_{z=1}^K)$. However, a naive optimization with respect to decomposition is generally intractable. Thus, we devise a practical method allowing joint training with the dynamics model.

3.3. Fine-Grained Causal Dynamics Learning with Quantization

We propose a practical differentiable method that allows joint optimization of Eq. (4) over dynamics model $\hat{p}$, decomposition $\{\mathcal{E}_z\}$, and graphs $\{\mathcal{G}_z\}$, in an end-to-end manner. The key component is a discrete latent codebook $C = \{e_z\}$ where each code e_z represents the pair of a subgroup $\mathcal{E}_z$ and a graph $\mathcal{G}_z$. The codebook learning is differentiable, and these pairs will be learned throughout the training with the dynamics model. The overall framework is shown in Fig. 2.

Quantization. The encoder g_{enc} maps each sample (s, a) into a latent embedding h, which is then quantized to the nearest prototype vector e (i.e., code) in the codebook $C = \{e_1, \cdots, e_K\}$, following Van Den Oord et al. (2017):

$$e = e_z, \quad \text{where } z = \operatorname*{argmin}_{j \in [K]} \|h - e_j\|_2. \quad (5)$$

This entails the subgroups since each sample corresponds to exactly one of the codes, i.e., each code e_z represents the subgroup $\mathcal{E}_z = \{(s, a) \mid e = e_z\}$. Thus, this corresponds to the term $p(z \mid s, a)$ in Eq. (3). In other words, the codebook C serves as a proxy for decomposition $\{\mathcal{E}_z\}_{z=1}^K$.

Local causal graphs. Quantized embedding e is then decoded to an adjacency matrix $A \in \{0, 1\}^{(N+M) \times N}$. The output of the decoder g_{dec} is the parameters of Bernoulli distributions from which the matrix is sampled: $A \sim g_{\text{dec}}(e)$. In other words, each code e_z corresponds to the matrix A_z that represents the graph $\mathcal{G}_z$. To properly backpropagate gradients, we adopt Gumbel-Softmax reparametrization trick (Jang et al., 2017; Maddison et al., 2017).

Dynamics learning. The dynamics model $\hat{p}$ employs the matrix A for prediction: $\sum_j \log \hat{p}(s'_j \mid s, a; A^{(j)})$, where $A^{(j)} \in \{0,1\}^{(N+M)}$ is the j-th column of A. Each entry of $A^{(j)}$ indicates whether the corresponding state or action variable will be used to predict the next state s'_j. This corresponds to the term $p(s'_j \mid Pa(j; \mathcal{E}_z), z)$ in Eq. (3). For the implementation, we mask out the features of unused variables according to A. We found that this is more stable compared to the input masking (Brouillard et al., 2020).

Training objective. We employ a regularization loss $\lambda \cdot \|A\|_1$ to induce a sparse LCG, where λ is a hyperparameter. To update the codebook, we use a quantization loss (Van Den Oord et al., 2017). The training objective is as follows:

$$\mathcal{L}_{\texttt{total}} = \underbrace{-\log \hat{p}(s' \mid s, a; A) + \lambda \cdot \|A\|_1}_{\mathcal{L}_{\texttt{pred}}} + \underbrace{\|\text{sg}\,[h] - e\|_2^2 + \beta \cdot \|h - \text{sg}\,[e]\,\|_2^2}_{\mathcal{L}_{\texttt{quant}}}. \quad (6)$$

Here, $\mathcal{L}_{\texttt{pred}}$ is the masked prediction loss with regularization. $\mathcal{L}_{\texttt{quant}}$ is the quantization loss where sg $[\cdot]$ is a stop-gradient operator and β is a hyperparameter. Specifically, $\|\text{sg}\,[h] - e\|_2^2$ moves each code toward the center of the embeddings assigned to it and $\beta \cdot \|h - \text{sg}\,[e]\,\|_2^2$ encourages the encoder to output the embeddings close to the codes. This allows us to jointly train the dynamics model and the codebook in an end-to-end manner. Intuitively, vector quantization clusters the samples under a similar context and reconstructs the LCGs for each clustering. The rationale is that any error in the graph $\mathcal{G}_z$ or clustering $\mathcal{E}_z$ would lead to the prediction error of the dynamics model. We provide the details of our model in Appendix C.4.

We note that prior works on learning a discrete latent codebook have mostly focused on the reconstruction of the observation (Van Den Oord et al., 2017; Ozair et al., 2021). To the best of our knowledge, our work is the first to utilize vector quantization for discovering diverse causal structures.

Discussion on the codebook collapsing. It is well known that training a discrete latent codebook with vector quantization often suffers from the codebook collapsing, where many codes learn the same output and converge to a trivial solution. For this, we employ exponential moving averages (EMA) to update the codebook, following Van Den Oord et al. (2017). In practice, we found that the training was relatively stable for any choice of the codebook size $K > 2$. In our experiments, we simply fixed it to 16 across all environments since they all performed comparably well, which we will demonstrate in Sec. 4.2.

3.4. Theoretical Analysis and Discussions

So far, we have described how our method learns the decomposition and LCGs through the discrete latent codebook C as a proxy. Our method can be viewed as a practical approach towards the maximization of $\mathcal{S}(\{\mathcal{G}_z, \mathcal{E}_z\}_{z=1}^K)$ since $\mathcal{L}_{\texttt{pred}}$ corresponds to Eq. (4) and $\mathcal{L}_{\texttt{quant}}$ is a mean squared error in the latent space which can be minimized to 0. In this section, we provide its implications and discussions.

Proposition 1. *Let $\{\mathcal{G}_z^*, \mathcal{E}_z^*\} \in \operatorname{argmax} \mathcal{S}(\{\mathcal{G}_z, \mathcal{E}_z\}_{z=1}^K)$ for $\lambda > 0$ small enough, with Assumptions 1 to 5. Then, (i) each $\mathcal{G}_z^*$ is true LCG on $\mathcal{E}_z^*$, and (ii) $\mathbb{E}\left[|\mathcal{G}_z^*|\right] \le \mathbb{E}\left[|\mathcal{G}_z|\right]$ where $\{\mathcal{G}_z\}$ are LCGs on arbitrary decomposition $\{\mathcal{E}_z\}_{z=1}^K$.*

In other words, the decomposition that maximizes the score is optimal in terms of $\mathbb{E}\left[|\mathcal{G}_z|\right] = \sum_z p(\mathcal{E}_z)|\mathcal{G}_z|$. This is an important property involving the contexts which are more likely (i.e., large $p(\mathcal{E})$) and more meaningful (i.e., sparse $\mathcal{G}_\mathcal{E}$). Therefore, Prop. 1 implies that score maximization would lead to the fine-grained understanding of the dynamics *at best* it can achieve given the quantization degree K.

We now illustrate how the optimal decomposition $\{\mathcal{E}_z^*\}_{z=1}^K$ in Prop. 1 with sufficient quantization degree identifies important context $\mathcal{D}$ (e.g., *door closed*) that displays fine-grained causal relationships. We say the context $\mathcal{D}$ is *canonical* if $\mathcal{G}_\mathcal{F} = \mathcal{G}_\mathcal{D}$ for any $\mathcal{F} \subset \mathcal{D}$.

Theorem 2 (Identifiability of contexts)**.** *Let $\{\mathcal{G}_z^*, \mathcal{E}_z^*\} \in \operatorname{argmax} \mathcal{S}(\{\mathcal{G}_z, \mathcal{E}_z\}_{z=1}^K)$ for $\lambda > 0$ small enough, with Assumptions 1 to 5. Suppose $\mathcal{X} = \cup_{m \in [H]} \mathcal{D}_m$ where $\mathcal{G}_{\mathcal{D}_m}$ is distinct for all $m \in [H]$, and $\mathcal{D}_1, \cdots, \mathcal{D}_H$ are disjoint and canonical. Suppose $K \ge H$. Then, for all $m \in [H]$, there exists $I_m \subset [K]$ such that $\mathcal{D}_m = \bigcup_{z \in I_m} \mathcal{E}_z^*$ almost surely.*

In other words, the joint optimization of Eq. (4) over the quantization and dynamics model with sufficient quantization degree perfectly captures meaningful contexts that exist in the system (Thm. 2) and recovers corresponding LCGs (Prop. 1-(i)), thereby leading to a fine-grained understanding of the dynamics. Our method described in the previous section serves as a practical approach toward this goal.

Discussion on the codebook size. Thm. 2 also implies that the identification of the meaningful contexts is agnostic to the quantization degree K, as long as $K \ge H$. In Sec. 4.2, we demonstrate that our method works reasonably well for various quantization degrees in practice. We note that determining a minimal and sufficient number of quantization H is *not* our primary focus. This is because over-parametrization of quantization incurs only a small memory cost for additional codebook vectors in practice. Note that even if $K < H$, Prop. 1-(ii) guarantees that it would still discover meaningful fine-grained causal relationships, optimal in terms of $\mathbb{E}\left[|\mathcal{G}_z|\right]$.

Relationship to past approaches. To better understand our approach, we draw connections to (i) prior causal dynamics models and (ii) sample-specific approaches to discovering fine-grained dependencies. First, our method with the quantization degree $K = 1$ degenerates to prior causal dynamics models (Wang et al., 2022; Ding et al., 2022): it would discover (global) causal dependencies (i.e., special case of Thm. 1) but cannot harness fine-grained relationships. Second, our method without quantization reverts to sample-specific approaches ($K \to \infty$), e.g., the auxiliary network that infers local independence directly from each sample (Hwang et al., 2023). As described earlier, it is unclear under which context the inferred dependencies hold. In Sec. 4.2, we demonstrate that this makes their inferences often inconsistent within the same context and prone to overfitting, while our approach with quantization infers fine-grained causal relationships in a more effective and robust manner.

4. Experiments

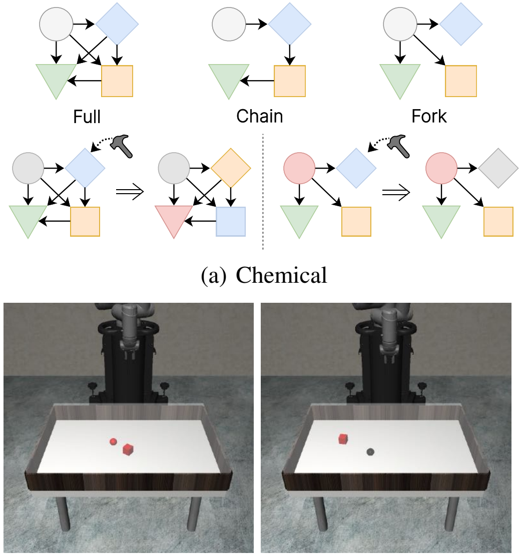
In this section, we evaluate our method, coined Fine-Grained Causal Dynamics Learning (**FCDL**), to investigate the following questions: (1) Does our method improve robustness in MBRL (Tables 1 and 2)? (2) Does our method discover fine-grained causal relationships and capture meaningful contexts (Figs. 5 to 7)? (3) Is our method more effective and robust compared to sample-specific approaches (Figs. 6 and 7)? (4) How does the degree of quantization affect performance (Fig. 7 and Table 3)?

4.1. Experimental Setup

The environments are designed to exhibit fine-grained causal relationships under a particular context $\mathcal{D}$. The state variables (e.g., position, velocity) are fully observable, following prior works (Ding et al., 2022; Wang et al., 2022; Seitzer et al., 2021; Pitis et al., 2020; 2022). Experimental details are provided in Appendix C.[2]

4.1.1. Environments

Chemical (Ke et al., 2021)**.** It is a widely used benchmark for systematic evaluation of causal reasoning in RL. There are 10 nodes, each colored with one of 5 colors. According to the underlying causal graph, an action changes the colors of the intervened node's descendants as depicted in Fig. 3(a). The task is to match the colors of each node to the given target. We designed two settings, named *full-fork* and *full-chain*. In both settings, the underlying CG is both *full*. When the color of the root node is red ($\mathcal{D}$), the colors change according to *fork* or *chain*, respectively ($\mathcal{G}_\mathcal{D}$). For example, in *full-fork*, all other parent nodes except the root become irrelevant under this context. Otherwise ($\mathcal{D}^c$), the transition respects the graph *full* (i.e., $\mathcal{G}_{\mathcal{D}^c} = \mathcal{G}$). During the test, the root color is set to red, and LCG (*fork* or *chain*) is activated. Here, the agent receives a noisy observation for some nodes, and the task is to match the colors of other clean nodes, as depicted in Appendix C.1 (Fig. 8). The agent capable of fine-grained causal reasoning would generalize well since corrupted nodes are locally spurious to predict other nodes.



(a) Chemical

(b) Magnetic

Figure 3. Illustrations for each environment. (a) In Chemical, colors change by the action according to the underlying causal graph. (b) In Magnetic, the red object exhibits magnetism.

Magnetic. We designed a robot arm manipulation environment based on the Robosuite framework (Zhu et al., 2020). There is a moving ball and a box on the table, colored red or black (Fig. 3(b)). Red color indicates that the object is *magnetic*, and attracts the other magnetic object. For example, when both are red, magnetic force will be applied, and the ball will move toward the box. Otherwise, i.e., under the non-magnetic context, the box would have no influence on the ball. The color and position of the objects are randomly initialized for each episode, i.e., each episode is under the magnetic or non-magnetic context during training. The task is to reach the ball, predicting its trajectory. In this environment, non-magnetic context $\mathcal{D}$ displays sparse dependencies ($\mathcal{G}_\mathcal{D} \subsetneq \mathcal{G}$) because the box no longer influences the ball under this context. In contrast, all causal dependencies remain the same under the magnetic context $\mathcal{D}^c$, i.e., $\mathcal{G}_{\mathcal{D}^c} = \mathcal{G}$. CG and LCGs are shown in Appendix C.1 (Fig. 9). During the test, one of the objects is black, and the box is located at an unseen position. Under this non-magnetic context, the box becomes locally spurious, and thus, the agent aware of fine-grained causal relationships would generalize well to unseen out-of-distribution (OOD) states.

4.1.2. Experimental details

Baselines. We first consider dense models, i.e., a monolithic network implemented as MLP which learns $p(s' \mid s, a)$, and a modular network having a separate network for each variable: $\prod_j p(s'_j \mid s, a)$. We also include a graph neural

[2] Our code is publicly available at https://github.com/iwhwang/Fine-Grained-Causal-RL.

Table 1. Average episode reward on training and downstream tasks in each environment. In Chemical, n denotes the number of noisy nodes in downstream tasks.

Methods	Chemical (*full-fork*)				Chemical (*full-chain*)				Magnetic	
	Train ($n=0$)	Test ($n=2$)	Test ($n=4$)	Test ($n=6$)	Train ($n=0$)	Test ($n=2$)	Test ($n=4$)	Test ($n=6$)	Train	Test
MLP	19.00±0.83	6.49±0.48	5.93±0.71	6.84±1.17	17.91±0.87	7.39±0.65	6.63±0.58	6.78±0.93	8.37±0.74	0.86±0.45
Modular	18.55±1.00	6.05±0.70	5.65±0.50	6.43±1.00	17.37±1.63	6.61±0.63	7.01±0.55	7.04±1.07	8.45±0.80	0.88±0.52
GNN (Kipf et al., 2020)	18.60±1.19	6.61±0.92	6.15±0.74	6.95±0.78	16.97±1.85	6.89±0.28	6.38±0.28	6.56±0.53	8.53±0.83	0.92±0.51
NPS (Goyal et al., 2021a)	7.71±1.22	5.82±0.83	5.75±0.57	5.54±0.80	8.20±0.54	6.92±1.03	6.88±0.79	6.80±0.39	3.13±1.00	0.91±0.69
CDL (Wang et al., 2022)	18.95±1.40	9.37±1.33	8.23±0.40	9.50±1.18	17.95±0.83	8.71±0.55	8.65±0.38	10.23±0.50	**8.75**±0.69	1.10±0.67
GRADER (Ding et al., 2022)	18.65±0.98	9.27±1.31	8.79±0.65	10.61±1.31	17.71±0.54	8.69±0.56	8.75±0.80	10.14±0.33	-	-
Oracle	**19.64**±1.18	7.83±0.87	8.04±0.62	9.66±0.21	17.79±0.76	8.47±0.69	8.85±0.78	10.29±0.37	8.42±0.86	0.95±0.55
NCD (Hwang et al., 2023)	19.30±0.95	10.95±1.63	9.11±0.63	10.32±0.93	**18.27**±0.27	9.60±1.52	8.86±0.23	10.32±0.37	8.48±0.70	1.31±0.77
FCDL (Ours)	19.28±0.87	**15.27**±2.53	**14.73**±1.68	**13.62**±2.56	17.22±0.61	**13.36**±3.60	**12.35**±3.23	**12.00**±1.21	8.52±0.74	**4.81**±3.01

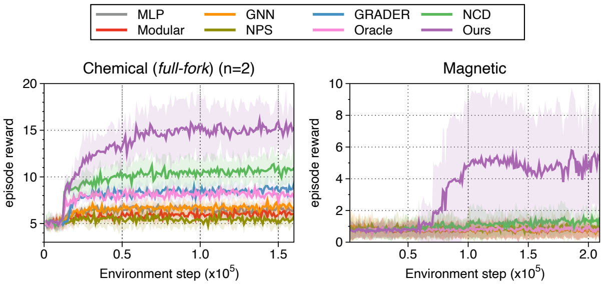


Figure 4. Learning curves on downstream tasks as measured on the average episode reward. Lines and shaded areas represent the mean and standard deviation, respectively.

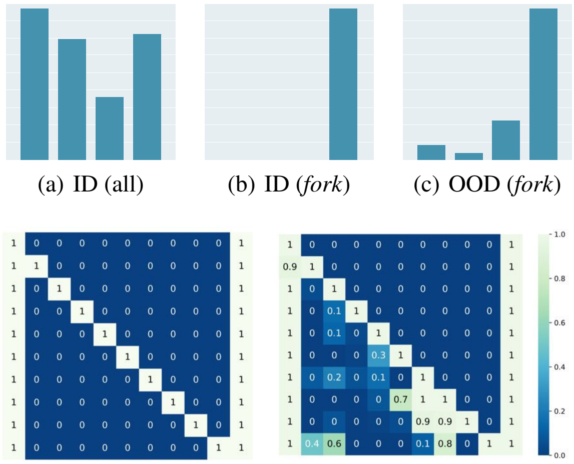
network (GNN) (Kipf et al., 2020), which learns the relational information, and NPS (Goyal et al., 2021a), which learns sparse and modular dynamics. Causal models, including CDL (Wang et al., 2022) and GRADER (Ding et al., 2022), infer causal structure for dynamics learning: $\prod_j p(s'_j \mid Pa(j))$. We also consider an *oracle* model, which leverages the ground truth (global) causal graph. Finally, we compare to NCD (Hwang et al., 2023), a sample-specific approach that examines local independence for each sample.

Planning algorithm. For all baselines and our method, we use a model predictive control (Camacho & Alba, 2013) which selects the actions based on the prediction of the learned dynamics model. Specifically, we use the cross-entropy method (CEM) (Rubinstein & Kroese, 2004), which iteratively generates and optimizes action sequences.

Implementation. For our method, we set the hyperparameters $K = 16, \lambda = 0.001$, and $\beta = 0.25$ in all experiments. All methods have a similar model capacity for a fair comparison. For the evaluation, we ran 10 test episodes for every 40 training episodes. The results are averaged over eight different runs. All learning curves are shown in Appendix C.5.

4.2. Results

Downstream task performance (Table 1, Fig. 4). All methods show similar performance on in-distribution (ID) states in training. However, dense models suffer from OOD states in the downstream tasks. Causal models are generally more robust compared to dense models, as they infer the causal graph and discard spurious dependencies. NCD, a sample-specific approach to infer fine-grained dependencies, performs better than causal models on a few downstream tasks, but not always. In contrast, our method consistently outperforms the baselines across all downstream tasks. This empirically validates our hypothesis that fine-grained causal reasoning leads to improved robustness in MBRL.



(a) ID (all) (b) ID (*fork*) (c) OOD (*fork*)

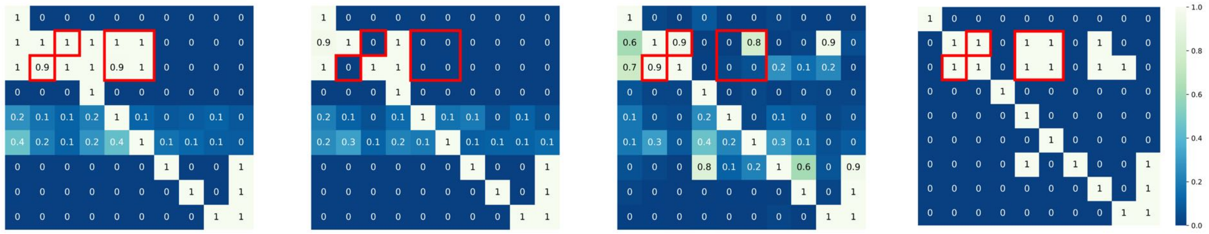
(d) True LCG (*fork*) (e) Learned LCG (*fork*)

Figure 5. (**Top**) Codebook histogram of the sample allocations to each of the codes on (a) all ID states, (b) ID states under *fork*, and (c) OOD states under *fork*. (**Bottom**) (d) True LCG (*fork*). (e) Learned LCG corresponding to the most frequently allocated code in (b) and (c).

Prediction accuracy (Table 2). To better understand the robustness of our method in downstream tasks, we investigate the prediction accuracy on ID and OOD states over the clean nodes in Chemical. As described earlier, noisy nodes are irrelevant for predicting the clean nodes under the LCG (i.e., *fork* or *chain*); thus, they are *locally spurious* on OOD states in downstream tasks. While all methods perform reasonably well on ID states, dense models show a significant performance drop under the presence of noisy

Table 2. Prediction accuracy on ID ($n = 0$) and OOD ($n = 2, 4, 6$) states in Chemical environment.

Setting / n		MLP	Modular	GNN	NPS	CDL	GRADER	Oracle	NCD	FCDL (Ours)
full-fork	($n = 0$)	88.31±1.58	89.24±1.52	88.81±1.44	58.34±2.08	89.22±1.67	87.75±1.64	89.63±1.62	**90.07**±1.22	89.46±1.40
	($n = 2$)	31.11±1.69	26.53±3.45	36.29±3.45	40.56±4.61	35.59±1.85	37.93±1.06	33.87±1.34	41.60±5.08	**66.44**±12.22
	($n = 4$)	30.44±2.28	24.73±5.61	25.80±3.48	26.81±4.37	35.82±1.40	38.94±1.63	36.48±1.80	37.47±2.13	**58.49**±10.20
	($n = 6$)	32.39±1.76	26.73±8.31	21.58±3.44	23.02±4.27	42.22±1.39	45.74±2.25	42.47±0.75	42.27±1.82	**49.09**±4.77
full-chain	($n = 0$)	84.38±1.31	85.92±1.15	85.41±1.84	58.48±2.81	**86.85**±1.47	84.24±1.22	85.76±1.56	85.63±1.01	86.07±1.62
	($n = 2$)	28.66±3.65	25.24±4.68	29.22±3.39	38.73±2.63	34.90±1.59	36.82±3.12	34.63±1.78	40.04±6.21	**60.34**±12.10
	($n = 4$)	26.52±4.26	24.94±4.81	23.28±4.98	27.69±4.28	36.52±1.72	37.41±2.84	38.31±2.48	37.47±2.98	**56.64**±9.40
	($n = 6$)	24.15±4.17	25.09±5.91	20.53±6.96	24.45±3.84	42.06±1.29	43.48±4.14	42.87±2.08	41.19±1.66	**53.29**±6.63



(a) FCDL (CG)

(b) FCDL (LCG)

(c) NCD (LCG, ID)

(d) NCD (LCG, OOD)

Figure 6. Red boxes indicate edges included in global CG, but not in LCG under the non-magnetic context. (a) CG inferred by our method. (b-d) LCG under the non-magnetic context inferred by (b) our method, and NCD on (c) ID and (d) OOD state.

variables, merely above 20% which is an expected accuracy of random guessing. As expected, causal dynamics models tend to be more robust compared to dense models, but they still suffer from OOD states. NCD is more robust than causal models when $n = 2$, but eventually becomes similar to them as the number of noisy nodes increases. In contrast, our method outperforms baselines by a large margin across all downstream tasks, which demonstrates its effectiveness and robustness in fine-grained causal reasoning.

Recognizing important contexts and fine-grained causal relationships (Fig. 5). To illustrate the fine-grained causal reasoning of our method, we closely examine the behavior of our model with the quantization degree $K = 4$ in Chemical (*full-fork*, $n = 2$). Recall each code corresponds to the pair of a subgroup and LCG, Fig. 5(a) shows how ID samples in the batch are allocated to one of the four codes. Interestingly, ID samples corresponding to LCG *fork* are all allocated to the last code (Fig. 5(b)), i.e., the subgroup corresponding to the last code identifies this context. Furthermore, LCG decoded from this code (Fig. 5(e)) accurately captures the true *fork* structure (Fig. 5(d)). This demonstrates that our method successfully recognizes meaningful context and fine-grained causal relationships. Notably, Fig. 5(c) shows that most of the OOD samples under *fork* are correctly allocated to the last code. This illustrates the robustness of our method, i.e., its inference is consistent between ID and OOD states. Additional examples, including the visualization of all the learned LCGs from all codes, are provided in Appendix C.5.

Inferred LCGs compared to sample-specific approach (Fig. 6). We investigated the effectiveness and robustness of our method in fine-grained causal reasoning compared to the sample-specific approach. For this, we examine the inferred LCGs in Magnetic, where true LCGs and CG are shown in Appendix C.1 (Fig. 9). First, our method accurately learns LCG under the non-magnetic context (Fig. 6(b)). On the other hand, the LCG inferred by NCD is rather inaccurate (Fig. 6(c)), including some locally spurious dependencies (3 among 6 red boxes). Furthermore, its inference is inconsistent between ID and OOD states in the same non-magnetic context and completely fails on OOD states (Fig. 6(d)). This demonstrates that our approach is more effective and robust in discovering fine-grained causal relationships.

Evaluation of local causal discovery (Fig. 7). We evaluate our method and NCD using structural hamming distance (SHD) in Magnetic. For each sample, we compare the inferred LCG with the true LCG based on the magnetic/non-magnetic context, and the SHD scores are averaged over the data samples in the evaluation batch. As expected, our method infers fine-grained relationships more accurately and maintains better performance on OOD states across various quantization degrees, which validates its effectiveness and robustness compared to NCD. Lastly, we note that our method with the quantization degree $K = 1$ would learn only a single CG over the entire data domain, as shown in Fig. 6(a). This explains its mean SHD score of 6 in non-magnetic samples in Fig. 7, since CG includes six redundant edges in non-magnetic context (i.e., red boxes in Fig. 9).

Ablation on the quantization degree (Table 3). Finally, we observe that our method works reasonably well across various quantization degrees on all downstream tasks in

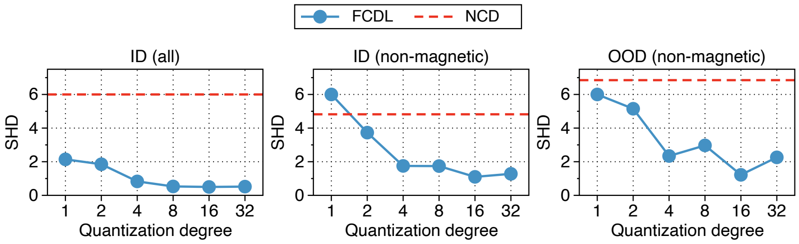


Figure 7. Evaluation of local causal discovery in Magnetic environment.

Table 3. Ablation on the quantization degree.

	Chemical (*full-fork*)		
Methods	$(n = 2)$	$(n = 4)$	$(n = 6)$
CDL	9.37±1.33	8.23±0.40	9.50±1.18
NCD	10.95±1.63	9.11±0.63	10.32±0.93
FCDL $(K = 2)$	13.44±5.41	12.86±5.58	12.99±5.27
FCDL $(K = 4)$	15.73±4.13	16.50±3.40	12.40±2.81
FCDL $(K = 8)$	14.95±1.16	15.03±2.61	13.42±2.67
FCDL $(K = 16)$	15.27±2.53	14.73±1.68	13.62±2.56
FCDL $(K = 32)$	16.12±1.43	14.35±1.37	14.79±2.13

Chemical (*full-fork*). Our method consistently outperforms the prior causal dynamics model (CDL) and sample-specific approach (NCD), which corroborates the results in Fig. 7. During our experiments, we found that the training was relatively stable for any quantization degree of $K > 2$. We also found that instability often occurs under $K = 2$, where the samples frequently fluctuate between two prototype vectors and result in the codebook collapsing. This is also shown in Table 3 where the performance of $K = 2$ is worse compared to other choices of K. We speculate that over-parametrization of quantization could alleviate such fluctuation in general.

5. Discussions and Future Works

High-dimensional observation. The factorization of the state space is natural in many real-world domains (e.g., healthcare, recommender system, social science, economics) where discovering causal relationships is an important problem. Extending our framework to the image would require extracting causal factors from pixels (Schölkopf et al., 2021), which is orthogonal to ours and could be combined with.

Scalability and stability in training. Vector quantization (VQ) is a well-established component in generative models where the quantization degree is usually very high (e.g., $K = 512, 1024$), yet effectively captures diverse visual features. Its scalability is further showcased in complex large-scale datasets (Razavi et al., 2019). In this sense, we believe our framework could extend to complex real-world environments. For training stability, techniques have been recently proposed to prevent codebook collapsing, such as codebook reset (Williams et al., 2020) and stochastic quantization (Takida et al., 2022). We consider that such techniques and tricks could be incorporated into our framework.

Conditional independence test (CIT). A CIT is an effective tool for understanding causal relationships, although often computation-costly. Our method may utilize it to further calibrate the learned LCGs, e.g., applying CIT on each subgroup after the training, which we defer to future work.

Domain knowledge. Our method could leverage prior information on important contexts displaying sparse dependencies, if available. While our method does not rely on such domain knowledge, it would still be useful for discovering fine-grained relationships more efficiently (e.g., Thm. 1).

Implications to real-world scenarios. We believe our work has potential implications in many practical applications since context-dependent causal relationships are prevalent in real-world scenarios. For example, in healthcare, a dynamic treatment regime is a task of determining a sequence of decision rules (e.g., treatment type, drug dosage) based on the patient's health status where it is known that many pathological factors involve fine-grained causal relationships (Barash & Friedman, 2001; Edwards & Toma, 1985). Our experiments illustrate that existing causal/non-causal RL approaches could suffer from locally spurious correlations and fail to generalize in downstream tasks. We believe our work serves as a stepping stone for further investigation into fine-grained causal reasoning of RL systems and their robustness in real-world deployment.

6. Conclusion

We present a novel approach to dynamics learning that infers fine-grained causal relationships, leading to improved robustness of MBRL. We provide a principled way to examine fine-grained dependencies under certain contexts. As a practical approach, our method learns a discrete latent variable that represents the pairs of a subgroup and local causal graphs (LCGs), allowing joint optimization with the dynamics model. Consequently, our method infers fine-grained causal structures in a more effective and robust manner compared to prior approaches. As one of the first steps towards fine-grained causal reasoning in sequential decision-making systems, we hope our work stimulates future research toward this goal.

Impact Statement

In real-world applications, model-based RL requires a large amount of data. As a large-scale dataset may contain sensitive information, it would be advisable to discreetly evaluate the models within simulated environments before their real-world deployment.

Acknowledgements

We would like to thank Sujin Jeon and Hyundo Lee for the useful discussions. We also thank anonymous reviewers for their constructive comments. This work was partly supported by the IITP (RS-2021-II212068-AIHub/10%, RS-2021-II211343-GSAI/10%, 2022-0-00951-LBA/10%, 2022-0-00953-PICA/20%), NRF (RS-2023-00211904/20%, RS-2023-00274280/10%, RS-2024-00353991/10%), and KEIT (RS-2024-00423940/10%) grant funded by the Korean government.

References

Acid, S. and de Campos, L. M. Searching for bayesian network structures in the space of restricted acyclic partially directed graphs. *Journal of Artificial Intelligence Research*, 18:445–490, 2003.

Barash, Y. and Friedman, N. Context-specific bayesian clustering for gene expression data. In *Proceedings of the fifth annual international conference on Computational biology*, pp. 12–21, 2001.

Bareinboim, E., Forney, A., and Pearl, J. Bandits with unobserved confounders: A causal approach. *Advances in Neural Information Processing Systems*, 28, 2015.

Bica, I., Jarrett, D., and van der Schaar, M. Invariant causal imitation learning for generalizable policies. *Advances in Neural Information Processing Systems*, 34:3952–3964, 2021.

Bongers, S., Blom, T., and Mooij, J. M. Causal modeling of dynamical systems. *arXiv preprint arXiv:1803.08784*, 2018.

Boutilier, C., Friedman, N., Goldszmidt, M., and Koller, D. Context-specific independence in bayesian networks. *CoRR*, abs/1302.3562, 2013.

Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., and Drouin, A. Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems*, 33:21865–21877, 2020.

Buesing, L., Weber, T., Zwols, Y., Heess, N., Racaniere, S., Guez, A., and Lespiau, J.-B. Woulda, coulda, shoulda: Counterfactually-guided policy search. In *International Conference on Learning Representations*, 2019.

Camacho, E. F. and Alba, C. B. *Model predictive control*. Springer science & business media, 2013.

Chitnis, R., Silver, T., Kim, B., Kaelbling, L., and Lozano-Perez, T. Camps: Learning context-specific abstractions for efficient planning in factored mdps. In *Conference on Robot Learning*, pp. 64–79. PMLR, 2021.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Dal, G. H., Laarman, A. W., and Lucas, P. J. Parallel probabilistic inference by weighted model counting. In *International Conference on Probabilistic Graphical Models*, pp. 97–108. PMLR, 2018.

De Haan, P., Jayaraman, D., and Levine, S. Causal confusion in imitation learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Ding, W., Lin, H., Li, B., and Zhao, D. Generalizing goal-conditioned reinforcement learning with variational causal reasoning. In *Advances in Neural Information Processing Systems*, 2022.

Edwards, D. and Toma, H. A fast procedure for model search in multidimensional contingency tables. *Biometrika*, 72:339–351, 1985.

Feng, F. and Magliacane, S. Learning dynamic attribute-factored world models for efficient multi-object reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 19117–19144, 2023.

Feng, F., Huang, B., Zhang, K., and Magliacane, S. Factored adaptation for non-stationary reinforcement learning. In *Advances in Neural Information Processing Systems*, 2022.

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

Goyal, A., Lamb, A., Gampa, P., Beaudoin, P., Blundell, C., Levine, S., Bengio, Y., and Mozer, M. C. Factorizing declarative and procedural knowledge in structured, dynamical environments. In *International Conference on Learning Representations*, 2021b.

Goyal, A., Lamb, A., Hoffmann, J., Sodhani, S., Levine, S., Bengio, Y., and Schölkopf, B. Recurrent independent mechanisms. In *International Conference on Learning Representations*, 2021c.

Hoey, J., St-Aubin, R., Hu, A., and Boutilier, C. Spudd: stochastic planning using decision diagrams. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pp. 279–288, 1999.

Huang, B., Lu, C., Leqi, L., Hernández-Lobato, J. M., Glymour, C., Schölkopf, B., and Zhang, K. Action-sufficient state representation learning for control with structural constraints. In *International Conference on Machine Learning*, pp. 9260–9279. PMLR, 2022.

Hwang, I., Kwak, Y., Song, Y.-J., Zhang, B.-T., and Lee, S. On discovery of local independence over continuous variables via neural contextual decomposition. In *Conference on Causal Learning and Reasoning*, pp. 448–472. PMLR, 2023.

Jamshidi, F., Akbari, S., and Kiyavash, N. Causal imitability under context-specific independence relations. In *Advances in Neural Information Processing Systems*, volume 36, pp. 26810–26830, 2023.

Jang, E., Gu, S., and Poole, B. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017.

Kaiser, Ł., Babaeizadeh, M., Miłos, P., Osiński, B., Campbell, R. H., Czechowski, K., Erhan, D., Finn, C., Kozakowski, P., Levine, S., et al. Model based reinforcement learning for atari. In *International Conference on Learning Representations*, 2020.

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. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Kearns, M. and Koller, D. Efficient reinforcement learning in factored mdps. In *IJCAI*, volume 16, pp. 740–747, 1999.

Killian, T. W., Ghassemi, M., and Joshi, S. Counterfactually guided policy transfer in clinical settings. In *Conference on Health, Inference, and Learning*, pp. 5–31. PMLR, 2022.

Kipf, T., van der Pol, E., and Welling, M. Contrastive learning of structured world models. In *International Conference on Learning Representations*, 2020.

Kumor, D., Zhang, J., and Bareinboim, E. Sequential causal imitation learning with unobserved confounders. In *Advances in Neural Information Processing Systems*, volume 34, pp. 14669–14680, 2021.

Lee, S. and Bareinboim, E. Structural causal bandits: Where to intervene? In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Lee, S. and Bareinboim, E. Characterizing optimal mixed policies: Where to intervene and what to observe. In *Advances in Neural Information Processing Systems*, volume 33, pp. 8565–8576, 2020.

Li, M., Zhang, J., and Bareinboim, E. Causally aligned curriculum learning. In *The Twelfth International Conference on Learning Representations*, 2024.

Li, Y., Torralba, A., Anandkumar, A., Fox, D., and Garg, A. Causal discovery in physical systems from videos. *Advances in Neural Information Processing Systems*, 33: 9180–9192, 2020.

Löwe, S., Madras, D., Zemel, R., and Welling, M. Amortized causal discovery: Learning to infer causal graphs from time-series data. In *Conference on Causal Learning and Reasoning*, pp. 509–525. PMLR, 2022.

Lu, C., Schölkopf, B., and Hernández-Lobato, J. M. Deconfounding reinforcement learning in observational settings. *arXiv preprint arXiv:1812.10576*, 2018.

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*, volume 2021, 2021.

Maddison, C. J., Mnih, A., and Teh, Y. W. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2017.

Madumal, P., Miller, T., Sonenberg, L., and Vetere, F. Explainable reinforcement learning through a causal lens. In *Proceedings of the AAAI conference on artificial intelligence*, pp. 2493–2500, 2020.

Mesnard, T., Weber, T., Viola, F., Thakoor, S., Saade, A., Harutyunyan, A., Dabney, W., Stepleton, T. S., Heess, N., Guez, A., et al. Counterfactual credit assignment in model-free reinforcement learning. In *International Conference on Machine Learning*, pp. 7654–7664. PMLR, 2021.

Mutti, M., De Santi, R., Rossi, E., Calderon, J. F., Bronstein, M., and Restelli, M. Provably efficient causal model-based reinforcement learning for systematic generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 9251–9259, 2023.

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.

Oberst, M. and Sontag, D. Counterfactual off-policy evaluation with gumbel-max structural causal models. In *International Conference on Machine Learning*, pp. 4881–4890. PMLR, 2019.

Ozair, S., Li, Y., Razavi, A., Antonoglou, I., Van Den Oord, A., and Vinyals, O. Vector quantized models for planning.

In *International Conference on Machine Learning*, pp. 8302–8313. PMLR, 2021.

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

Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Pitis, S., Creager, E., and Garg, A. Counterfactual data augmentation using locally factored dynamics. *Advances in Neural Information Processing Systems*, 33, 2020.

Pitis, S., Creager, E., Mandlekar, A., and Garg, A. MocoDA: Model-based counterfactual data augmentation. In *Advances in Neural Information Processing Systems*, 2022.

Poole, D. Context-specific approximation in probabilistic inference. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pp. 447–454, 1998.

Ramsey, J., Spirtes, P., and Zhang, J. Adjacency-faithfulness and conservative causal inference. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pp. 401–408, 2006.

Razavi, A., van den Oord, A., and Vinyals, O. Generating diverse high-fidelity images with vq-vae-2. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Rezende, D. J., Danihelka, I., Papamakarios, G., Ke, N. R., Jiang, R., Weber, T., Gregor, K., Merzic, H., Viola, F., Wang, J., et al. Causally correct partial models for reinforcement learning. *arXiv preprint arXiv:2002.02836*, 2020.

Rubinstein, R. Y. and Kroese, D. P. *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation, and machine learning*, volume 133. Springer, 2004.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634, 2021.

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839): 604–609, 2020.

Seitzer, M., Schölkopf, B., and Martius, G. Causal influence detection for improving efficiency in reinforcement learning. In *Advances in Neural Information Processing Systems*, 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.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Takida, Y., Shibuya, T., Liao, W., Lai, C.-H., Ohmura, J., Uesaka, T., Murata, N., Takahashi, S., Kumakura, T., and Mitsufuji, Y. Sq-vae: Variational bayes on discrete representation with self-annealed stochastic quantization. In *International Conference on Machine Learning*, pp. 20987–21012. PMLR, 2022.

Tikka, S., Hyttinen, A., and Karvanen, J. Identifying causal effects via context-specific independence relations. *Advances in Neural Information Processing Systems*, 32: 2804–2814, 2019.

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.

Van Den Oord, A., Vinyals, O., et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

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.

Wang, Z., Xiao, X., Xu, Z., Zhu, Y., and Stone, P. Causal dynamics learning for task-independent state abstraction. In *International Conference on Machine Learning*, pp. 23151–23180. PMLR, 2022.

Wang, Z., Hu, J., Stone, P., and Martín-Martín, R. ELDEN: Exploration via local dependencies. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Williams, W., Ringer, S., Ash, T., MacLeod, D., Dougherty, J., and Hughes, J. Hierarchical quantized autoencoders. *Advances in Neural Information Processing Systems*, 33: 4524–4535, 2020.

Yang, K., Katcoff, A., and Uhler, C. Characterizing and learning equivalence classes of causal dags under interventions. In *International Conference on Machine Learning*, pp. 5541–5550. PMLR, 2018.

Yao, W., Chen, G., and Zhang, K. Learning latent causal dynamics. *arXiv preprint arXiv:2202.04828*, 2022.

Yoon, J., Wu, Y.-F., Bae, H., and Ahn, S. An investigation into pre-training object-centric representations for reinforcement learning. In *International Conference on Machine Learning*, pp. 40147–40174. PMLR, 2023.

Zadaianchuk, A., Seitzer, M., and Martius, G. Self-supervised visual reinforcement learning with object-centric representations. In *International Conference on Learning Representations*, 2021.

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, J., Kumor, D., and Bareinboim, E. Causal imitation learning with unobserved confounders. *Advances in neural information processing systems*, 33:12263–12274, 2020b.

Zhang, N. L. and Poole, D. On the role of context-specific independence in probabilistic inference. In *16th International Joint Conference on Artificial Intelligence, IJCAI 1999, Stockholm, Sweden*, volume 2, pp. 1288, 1999.

Zholus, A., Ivchenkov, Y., and Panov, A. Factorized world models for learning causal relationships. In *ICLR2022 Workshop on the Elements of Reasoning: Objects, Structure and Causality*, 2022.

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

A. Appendix for Preliminary

A.1. Extended Related Work

Recently, incorporating causal reasoning into RL has gained much attention in the community in various aspects. For example, causality has been shown to improve off-policy evaluation (Buesing et al., 2019; Oberst & Sontag, 2019), goal-directed tasks (Nair et al., 2019), credit assignment (Mesnard et al., 2021), robustness (Lyle et al., 2021; Volodin et al., 2020), policy transfer (Killian et al., 2022), explainability (Madumal et al., 2020), and policy learning with counterfactual data augmentation (Lu et al., 2020; Pitis et al., 2020; 2022). Causality has also been integrated with bandits (Bareinboim et al., 2015; Lee & Bareinboim, 2018; 2020), curriculum learning (Li et al., 2024) or imitation learning (Bica et al., 2021; De Haan et al., 2019; Zhang et al., 2020b; Kumor et al., 2021; Jamshidi et al., 2023) to handle the unobserved confounders and learn generalizable policies. Another line of work focused on causal reasoning over the high-dimensional visual observation (Lu et al., 2018; Rezende et al., 2020; Feng et al., 2022; Feng & Magliacane, 2023), e.g., learning sparse and modular dynamics (Goyal et al., 2021c;b;a), where the representation learning is crucial (Zhang et al., 2019; Sontakke et al., 2021; Tomar et al., 2021; Schölkopf et al., 2021; Zadaianchuk et al., 2021; Yoon et al., 2023).

Our work falls into the category of incorporating causality into dynamics learning in RL (Mutti et al., 2023), where recent works have focused on conditional independences between the variables and their global causal relationships (Wang et al., 2021; 2022; Ding et al., 2022). On the contrary, our work incorporates fine-grained causal relationships into dynamics learning, which is underexplored in prior works.

A.2. Background on Local Independence Relationship

In this subsection, we provide the background on the local independence relationship. We first describe context-specific independence (CSI) (Boutilier et al., 2013), which denotes a variable being conditionally independent of others given a particular context, not the full set of parents in the graph.

Definition 2 (Context-Specific Independence (CSI) (Boutilier et al., 2013), reproduced from Hwang et al. (2023))**.** *Y is said to be* **contextually independent** *of $\mathbf{X}_B$ given the context $\mathbf{X}_A = \mathbf{x}_A$ if $P\left(y \mid \mathbf{x}_A, \mathbf{x}_B\right) = P\left(y \mid \mathbf{x}_A\right)$, holds for all $y \in \mathcal{Y}$ and $\mathbf{x}_B \in \mathcal{X}_B$ whenever $P\left(\mathbf{x}_A, \mathbf{x}_B\right) > 0$. This will be denoted by $Y \perp\!\!\!\perp \mathbf{X}_B \mid \mathbf{X}_A = \mathbf{x}_A$.*

CSI has been widely studied especially for discrete variables with low cardinality, e.g., binary variables. Context-set specific independence (CSSI) generalizes the notion of CSI allowing continuous variables.

Definition 3 (Context-Set Specific Independence (CSSI) (Hwang et al., 2023))**.** *Let $\mathbf{X} = \{X_1, \cdots, X_d\}$ be a non-empty set of the parents of Y in a causal graph, and $\mathcal{E} \subseteq \mathcal{X}$ be an event with a positive probability. $\mathcal{E}$ is said to be a* **context set** *which induces* **context-set specific independence (CSSI)** *of $\mathbf{X}_{A^c}$ from Y if $p\left(y \mid \mathbf{x}_{A^c}, \mathbf{x}_A\right) = p\left(y \mid \mathbf{x}'_{A^c}, \mathbf{x}_A\right)$ holds for every $\left(\mathbf{x}_{A^c}, \mathbf{x}_A\right), \left(\mathbf{x}'_{A^c}, \mathbf{x}_A\right) \in \mathcal{E}$. This will be denoted by $Y \perp\!\!\!\perp \mathbf{X}_{A^c} \mid \mathbf{X}_A, \mathcal{E}$.*

Intuitively, it denotes that the conditional distribution $p(y \mid x) = p(y \mid x_{A^c}, x_A)$ is the same for different values of $\mathbf{x}_{A^c}$, for all $x = (\mathbf{x}_{A^c}, \mathbf{x}_A) \in \mathcal{E}$. In other words, only a subset of the parent variables is sufficient for modeling $p(y \mid x)$ on $\mathcal{E}$.

A.3. Fine-Grained Causal Relationships in Factored MDP

As mentioned in Sec. 2, we consider factored MDP where the causal graph is directed bipartite and make standard assumptions in the field to properly identify the causal relationships in MBRL (Ding et al., 2022; Wang et al., 2021; 2022; Seitzer et al., 2021; Pitis et al., 2020; 2022).

Assumption 1. *We assume Markov property (Pearl, 2009), faithfulness (Peters et al., 2017), and causal sufficiency (Spirtes et al., 2000).*

Recall that $\mathbf{X} = \{S_1, \cdots, S_N, A_1, \cdots, A_M\}$, $\mathbf{Y} = \{S'_1, \cdots, S'_N\}$, and $Pa(j)$ is parent variables of S'_j. Now, we formally define local independence by adapting CSSI to our setting.

Definition 4 (Local Independence)**.** *Let $\mathbf{T} \subseteq Pa(j)$ and $\mathcal{E} \subseteq \mathcal{X}$ with $p(\mathcal{E}) > 0$. We say the local independence $S'_j \perp\!\!\!\perp \mathbf{X} \setminus \mathbf{T} \mid \mathbf{T}, \mathcal{E}$ holds on $\mathcal{E}$ if $p(s'_j \mid \mathbf{x}_{T^c}, \mathbf{x}_T) = p(s'_j \mid \mathbf{x}'_{T^c}, \mathbf{x}_T)$ holds for every $\left(\mathbf{x}_{T^c}, \mathbf{x}_T\right), \left(\mathbf{x}'_{T^c}, \mathbf{x}_T\right) \in \mathcal{E}$.*[3]

It implies that only a subset of the parent variables ($\mathbf{x}_T$) is locally relevant on $\mathcal{E}$, and any other remaining variables ($\mathbf{x}_{T^c}$) are locally irrelevant, i.e., $p(s'_j \mid \mathbf{x})$ is a function of $\mathbf{x}_T$ on $\mathcal{E}$. Local independence generalizes conditional independence

[3] T denotes an index set of $\mathbf{T}$.

in the sense that if $S'_j \perp\!\!\!\perp \mathbf{X} \setminus \mathbf{T} \mid \mathbf{T}$ holds, then $S'_j \perp\!\!\!\perp \mathbf{X} \setminus \mathbf{T} \mid \mathbf{T}, \mathcal{E}$ holds for any $\mathcal{E} \subseteq \mathcal{X}$. Throughout the paper, we are concerned with the events with the positive probability, i.e., $p(\mathcal{E}) > 0$.

Definition 5. *$Pa(j; \mathcal{E})$ is a subset of $Pa(j)$ such that $S'_j \perp\!\!\!\perp \mathbf{X} \setminus Pa(j; \mathcal{E}) \mid Pa(j; \mathcal{E}), \mathcal{E}$ holds and $S'_j \not\perp\!\!\!\perp \mathbf{X} \setminus \mathbf{T} \mid \mathbf{T}, \mathcal{E}$ for any $\mathbf{T} \subsetneq Pa(j; \mathcal{E})$.*

In other words, $Pa(j; \mathcal{E})$ is a minimal subset of $Pa(j)$ in which the local independence on $\mathcal{E}$ holds. Clearly, $Pa(j; \mathcal{X}) = Pa(j)$, i.e., local independence on $\mathcal{X}$ is equivalent to the (global) conditional independence.

LCG (Def. 1) describes fine-grained causal relationships specific to $\mathcal{E}$. LCG is always a subgraph of the (global) causal graph, i.e., $\mathcal{G}_\mathcal{D} \subseteq \mathcal{G}$, because if a dependency (i.e., edge) does not exist under the whole domain, it cannot exist under any context. Note that $\mathcal{G}_\mathcal{X} = \mathcal{G}$, i.e., local independence and LCG under $\mathcal{X}$ are equivalent to conditional independence and CG, respectively.

Analogous to the faithfulness assumption (Peters et al., 2017) that no conditional independences other than ones entailed by CG are present, we introduce a similar assumption for LCG and local independence.

Assumption 2 ($\mathcal{E}$-Faithfulness)**.** *For any $\mathcal{E}$, no local independences on $\mathcal{E}$ other than the ones entailed by $\mathcal{G}_\mathcal{E}$ are present, i.e., for any j, there does not exists any $\mathbf{T}$ such that $Pa(j; \mathcal{E}) \setminus \mathbf{T} \neq \emptyset$ and $S'_j \perp\!\!\!\perp \mathbf{X} \setminus \mathbf{T} \mid \mathbf{T}, \mathcal{E}$.*

Regardless of $\mathcal{E}$-faithfulness assumption, LCG always exists because $Pa(j; \mathcal{E})$ always exists. However, such LCG may not be unique without this (see Hwang et al. (2023, Example. 2) for this example). Assumption 2 implies the uniqueness of $Pa(j; \mathcal{E})$ and $\mathcal{G}_\mathcal{E}$, and thus it is required to properly identify fine-grained causal relationships between the variables.

Such fine-grained causal relationships are prevalent in the real world. *Physical law*; To move a static object, a force exceeding frictional resistance must be exerted. Otherwise, the object would not move. *Logic*; Consider $A \vee B \vee C$. When A is true, any changes of B or C no longer affect the outcome. *Biology*; In general, smoking has a causal effect on blood pressure. However, one's blood pressure becomes independent of smoking if a ratio of alpha and beta lipoproteins is larger than a certain threshold (Edwards & Toma, 1985).

B. Appendix for Method and Theoretical Analysis

B.1. Fine-Grained Dynamics Modeling

With the arbitrary decomposition $\{\mathcal{E}_z\}_{z=1}^K$, true transition dynamics $p(s' \mid s, a)$ can be written as:

$$p(s' \mid s, a) = \sum_z p(s' \mid s, a, z) p(z \mid s, a) = \sum_z \prod_j p(s'_j \mid Pa(j; \mathcal{E}_z), z) \mathbb{1}_{\{(s,a) \in \mathcal{E}_z\}}, \tag{7}$$

where $p(z \mid s, a) = 1$ if $(s, a) \in \mathcal{E}_z$. This illustrates our approach to dynamics modeling based on fine-grained causal dependencies: $p(s'_j \mid s, a)$ is a function of $Pa(j, \mathcal{E}_z)$ on $\mathcal{E}_z$, and our dynamics model employs locally relevant dependencies $Pa(j, \mathcal{E}_z)$ for predicting S'_j. Our dynamics modeling with some graphs $\{\mathcal{G}_z\}_{z=1}^K$ is:

$$\hat{p}(s' \mid s, a; \{\mathcal{G}_z, \mathcal{E}_z\}, \phi) = \sum_z \hat{p}(s' \mid s, a; \mathcal{G}_z, \phi_z) \mathbb{1}_{\{(s,a) \in \mathcal{E}_z\}} = \sum_z \prod_j \hat{p}_j(s'_j \mid Pa^{\mathcal{G}_z}(j); \phi_z^{(j)}) \mathbb{1}_{\{(s,a) \in \mathcal{E}_z\}}, \tag{8}$$

where $\phi_z^{(j)}$ takes $Pa^{\mathcal{G}_z}(j)$ as an input and outputs the parameters of the density function $\hat{p}_j$ and $\phi := \{\phi_z^{(j)}\}$. We denote $\hat{p}_{\{\mathcal{G}_z, \mathcal{E}_z\}, \phi} := \hat{p}(s' \mid s, a; \{\mathcal{G}_z, \mathcal{E}_z\}, \phi)$ and $\hat{p}_{\mathcal{G}_z, \phi_z} := \hat{p}(s' \mid s, a; \mathcal{G}_z, \phi_z)$. In other words, $\hat{p}_{\{\mathcal{G}_z, \mathcal{E}_z\}, \phi}(s' \mid s, a) = \hat{p}_{\mathcal{G}_z, \phi_z}(s' \mid s, a)$ if $(s, a) \in \mathcal{E}_z$.

Now, we revisit the score function in Eq. (4):

$$\begin{aligned}
\mathcal{S}(\{\mathcal{G}_z, \mathcal{E}_z\}_{z=1}^K) &:= \sup_\phi \mathbb{E}_{p(s,a,s')} \left[\log \hat{p}(s' \mid s, a; \{\mathcal{G}_z, \mathcal{E}_z\}, \phi) - \lambda |\mathcal{G}_z| \right], && (9) \\
&= \sup_\phi \mathbb{E}_{p(s,a)} \mathbb{E}_{p(s' \mid s,a)} \left[\log \hat{p}(s' \mid s, a; \{\mathcal{G}_z, \mathcal{E}_z\}, \phi) - \lambda |\mathcal{G}_z| \right], && (10) \\
&= \sup_\phi \sum_z \int_{(s,a) \in \mathcal{E}_z} p(s, a) \left(\mathbb{E}_{p(s' \mid s,a)} \log \hat{p}(s' \mid s, a; \mathcal{G}_z, \phi_z) - \lambda |\mathcal{G}_z| \right), && (11) \\
&= \sup_\phi \sum_z \left[\int_{(s,a) \in \mathcal{E}_z} p(s, a) \left(\mathbb{E}_{p(s' \mid s,a)} \log \hat{p}(s' \mid s, a; \mathcal{G}_z, \phi_z) \right) - \lambda \cdot p(\mathcal{E}_z) \cdot |\mathcal{G}_z| \right], && (12)
\end{aligned}$$

where $\hat{p}(s' \mid s,a;\mathcal{G}_z,\phi_z)=\prod_j \hat{p}_j(s'_j \mid Pa^{\mathcal{G}_z}(j);\phi_z^{(j)})$.

B.2. Proof of Thm. 1

Due to the nature of factored MDP where the causal graph is directed bipartite, each Markov equivalence class (MEC) constrained under temporal precedence contains a *unique* causal graph (i.e., a skeleton determines a unique causal graph since temporal precedence fully orients the edges). Given this background, it is known that the causal graph is *uniquely identifiable* with oracle conditional independence test (Ding et al., 2022) or score-based method (Brouillard et al., 2020).

We will now show that LCG is also uniquely identifiable via score maximization. Our proof techniques are built upon Brouillard et al. (2020). It is worth noting that they provide the identifiability of (global) CG up to $\mathcal{I}$-MEC (Yang et al., 2018) by utilizing observational and interventional data. In contrast, our analysis is on the identifiability of **LCGs** by utilizing only observational data. We start by adopting some assumptions from Brouillard et al. (2020).

Assumption 3. *The ground truth density $p(s' \mid s,a)\in\mathcal{H}(\{\mathcal{G}_z^*,\mathcal{E}_z\})$ for any decomposition $\{\mathcal{E}_z\}$ with corresponding true LCGs $\{\mathcal{G}_z^*\}$, where $\mathcal{H}(\{\mathcal{G}_z^*,\mathcal{E}_z\}):=\{p \mid \exists\phi, p=\hat{p}_{\{\mathcal{G}_z^*,\mathcal{E}_z\},\phi}\}$. We assume the density $\hat{p}_{\{\mathcal{G}_z^*,\mathcal{E}_z\},\phi}$ is strictly positive for all ϕ.*

Definition 6. *For a graph $\mathcal{G}$ and $\mathcal{E}\subset\mathcal{X}$, let $\mathcal{F}_{\mathcal{E}}(\mathcal{G})$ be a set of conditional densities f such that $f(s' \mid s,a)=\prod_j f_j(s'_j \mid Pa^{\mathcal{G}}(j))$ for all $(s,a)\in\mathcal{E}$ where each f_j is a conditional density.*

Assumption 4. $|\mathbb{E}_{p(s,a,s')}\log p(s' \mid s,a)|<\infty$.

Assumption 3 states that the model parametrized by neural network has sufficient capacity to represent the ground truth density. Assumption 4 is a technical tool for handling the score differences as we will see later.

Lemma 1. *Let $\mathcal{G}_z^*$ be a true LCG on $\mathcal{E}_z$ for all z. Then, $\mathcal{S}(\{\mathcal{G}_z^*,\mathcal{E}_z\}_{z=1}^K)=\mathbb{E}_{p(s,a,s')}\log p(s' \mid s,a)-\lambda\cdot\mathbb{E}\left[|\mathcal{G}_z^*|\right]$.*

Proof. First,

$$0\le D_{KL}(p \parallel \hat{p}_{\{\mathcal{G}_z^*,\mathcal{E}_z\},\phi})=\mathbb{E}_{p(s,a,s')}\log p(s' \mid s,a)-\mathbb{E}_{p(s,a,s')}\log\hat{p}(s' \mid s,a;\{\mathcal{G}_z^*,\mathcal{E}_z\},\phi), \tag{13}$$

where the equality holds because $\mathbb{E}_{p(s,a,s')}\log p(s' \mid s,a)<\infty$ by Assumption 4. Therefore,

$$\sup_\phi \mathbb{E}_{p(s,a,s')}\log\hat{p}(s' \mid s,a;\{\mathcal{G}_z^*,\mathcal{E}_z\},\phi)\le\mathbb{E}_{p(s,a,s')}\log p(s' \mid s,a). \tag{14}$$

On the other hand, by Assumption 3, there exists ϕ^* such that $p=\hat{p}_{\{\mathcal{G}_z^*,\mathcal{E}_z\},\phi^*}$. Hence,

$$\sup_\phi \mathbb{E}_{p(s,a,s')}\log\hat{p}(s' \mid s,a;\{\mathcal{G}_z^*,\mathcal{E}_z\},\phi)\ge\mathbb{E}_{p(s,a,s')}\log\hat{p}(s' \mid s,a;\{\mathcal{G}_z^*,\mathcal{E}_z\},\phi^*)=\mathbb{E}_{p(s,a,s')}\log p(s' \mid s,a). \tag{15}$$

By Eqs. (14) and (15), we have $\sup_\phi \mathbb{E}_{p(s,a,s')}\log\hat{p}(s' \mid s,a;\{\mathcal{G}_z^*,\mathcal{E}_z\},\phi)=\mathbb{E}_{p(s,a,s')}\log p(s' \mid s,a)$. Therefore, we have $\mathcal{S}(\{\mathcal{G}_z^*,\mathcal{E}_z\}_{z=1}^K)=\mathbb{E}_{p(s,a,s')}\log p(s' \mid s,a)-\lambda\cdot\mathbb{E}\left[|\mathcal{G}_z^*|\right]$. □

Corollary 1. *Let $\mathcal{G}_z^*$ be a true LCG on $\mathcal{E}_z$ for all z. Then, $|\mathcal{S}(\{\mathcal{G}_z^*,\mathcal{E}_z\}_{z=1}^K)|<\infty$.*

Proof. By Lemma 1, $\mathcal{S}(\{\mathcal{G}_z^*,\mathcal{E}_z\}_{z=1}^K)=\mathbb{E}_{p(s,a,s')}\log p(s' \mid s,a)-\lambda\cdot\mathbb{E}\left[|\mathcal{G}_z^*|\right]$. Since $|\mathbb{E}_{p(s,a,s')}\log p(s' \mid s,a)|<\infty$ by Assumption 4 and $|\mathcal{G}_z^*|\le N(N+M)$, this concludes that $|\mathcal{S}(\{\mathcal{G}_z^*,\mathcal{E}_z\}_{z=1}^K)|<\infty$. □

Lemma 2. *Let $\mathcal{G}_z^*$ be a true LCG on $\mathcal{E}_z$ for all z. Then,*

$$\mathcal{S}(\{\mathcal{G}_z^*,\mathcal{E}_z\}_{z=1}^K)-\mathcal{S}(\{\mathcal{G}_z,\mathcal{E}_z\}_{z=1}^K)=\inf_\phi D_{KL}(p \parallel \hat{p}_{\{\mathcal{G}_z,\mathcal{E}_z\},\phi})+\lambda\sum_z p(\mathcal{E}_z)(|\mathcal{G}_z|-|\mathcal{G}_z^*|). \tag{16}$$

Proof. First, we can rewrite the score $\mathcal{S}(\{\mathcal{G}_z,\mathcal{E}_z\}_{z=1}^K)$ as:

$$\begin{aligned}\mathcal{S}(\{\mathcal{G}_z,\mathcal{E}_z\}_{z=1}^K)&=\sup_\phi \mathbb{E}_{p(s,a,s')}\log\hat{p}(s' \mid s,a;\{\mathcal{G}_z,\mathcal{E}_z\},\phi)-\lambda\cdot\mathbb{E}\left[|\mathcal{G}_z|\right] &(17)\\ &=-\inf_\phi -\mathbb{E}_{p(s,a,s')}\log\hat{p}(s' \mid s,a;\{\mathcal{G}_z,\mathcal{E}_z\},\phi)-\lambda\cdot\mathbb{E}\left[|\mathcal{G}_z|\right] &(18)\\ &=-\inf_\phi D_{KL}(p \parallel \hat{p}_{\{\mathcal{G}_z,\mathcal{E}_z\},\phi})+\mathbb{E}_{p(s,a,s')}\log p(s' \mid s,a)-\lambda\cdot\mathbb{E}\left[|\mathcal{G}_z|\right] &(19)\end{aligned}$$

The last equality holds by Assumption 4. Subtracting Eq. (19) from Lemma 1, we obtain:

$$\mathcal{S}(\{\mathcal{G}_z^*, \mathcal{E}_z\}_{z=1}^K) - \mathcal{S}(\{\mathcal{G}_z, \mathcal{E}_z\}_{z=1}^K) = \inf_\phi D_{KL}(p \parallel \hat{p}_{\{\mathcal{G}_z, \mathcal{E}_z\}, \phi}) + \lambda \sum_z p(\mathcal{E}_z)(|\mathcal{G}_z| - |\mathcal{G}_z^*|).$$

Note that $|\mathcal{S}(\{\mathcal{G}_z^*, \mathcal{E}_z\}_{z=1}^K)| < \infty$ by Corollary 1, and thus, this score difference is well defined. □

Lemma 3 (Modified from Brouillard et al. (2020), Lemma 16)**.** *If $p \notin \mathcal{F}_{\mathcal{E}_z}(\mathcal{G}_z)$, then*

$$\inf_{\phi_z} \int p_z(s,a) D_{KL}(p(\cdot \mid s,a) \parallel \hat{p}_{\mathcal{G}_z, \phi_z}(\cdot \mid s,a)) > 0, \tag{20}$$

where $p_z(s,a) \coloneqq p(s,a \mid z) = p(s,a)/p(\mathcal{E}_z)$ for all $(s,a) \in \mathcal{E}_z$, i.e., density function of the distribution $P_{S \times A | \mathcal{E}_z}$.

Proof. First, since $\hat{p}_{\mathcal{G}_z, \phi_z} \in \mathcal{F}_{\mathcal{E}_z}(\mathcal{G}_z)$ for all ϕ_z,

$$\inf_{\phi_z} \int p_z(s,a) D_{KL}(p(\cdot \mid s,a) \parallel \hat{p}_{\mathcal{G}_z, \phi_z}(\cdot \mid s,a)) \geq \inf_{f \in \mathcal{F}_{\mathcal{E}_z}(\mathcal{G}_z)} \int p_z(s,a) D_{KL}(p(\cdot \mid s,a) \parallel f(\cdot \mid s,a)). \tag{21}$$

Now, let $\hat{f}(s' \mid s,a) \coloneqq \prod_j p_z(s'_j \mid Pa^{\mathcal{G}_z}(j))$ for all $(s,a) \in \mathcal{E}_z$. Then, for any $f \in \mathcal{F}_{\mathcal{E}_z}(\mathcal{G}_z)$,

$$\begin{aligned}\int p_z(s,a) \int p(s' \mid s,a) \log \frac{\hat{f}(s' \mid s,a)}{f(s' \mid s,a)} &= \int p_z(s,a,s') \sum_j \log \frac{p_z(s'_j \mid Pa^{\mathcal{G}_z}(j))}{f_j(s'_j \mid Pa^{\mathcal{G}_z}(j))} \\ &= \sum_j \int p_z(s,a,s') \log \frac{p_z(s'_j \mid Pa^{\mathcal{G}_z}(j))}{f_j(s'_j \mid Pa^{\mathcal{G}_z}(j))} \\ &= \sum_j \int p_z(Pa^{\mathcal{G}_z}(j)) \int p_z(s'_j \mid Pa^{\mathcal{G}_z}(j)) \log \frac{p_z(s'_j \mid Pa^{\mathcal{G}_z}(j))}{f_j(s'_j \mid Pa^{\mathcal{G}_z}(j))} \geq 0. \end{aligned} \tag{22}$$

Therefore, for any $f \in \mathcal{F}_{\mathcal{E}_z}(\mathcal{G}_z)$,

$$\begin{aligned}\int p_z(s,a) D_{KL}(p(\cdot \mid s,a) \parallel f(\cdot \mid s,a)) &= \int p_z(s,a) \int p(s' \mid s,a) \log \frac{p(s' \mid s,a)}{\hat{f}(s' \mid s,a)} \frac{\hat{f}(s' \mid s,a)}{f(s' \mid s,a)} \\ &= \int p_z(s,a) D_{KL}(p \parallel \hat{f}) + \int p_z(s,a) \int p(s' \mid s,a) \log \frac{\hat{f}(s' \mid s,a)}{f(s' \mid s,a)} \\ &\geq \int p_z(s,a) D_{KL}(p \parallel \hat{f}).\end{aligned}$$

Here, the last inequality holds by Eq. (22). Therefore,

$$\inf_{f \in \mathcal{F}_{\mathcal{E}_z}(\mathcal{G}_z)} \int p_z(s,a) D_{KL}(p(\cdot \mid s,a) \parallel f(\cdot \mid s,a)) = \int p_z(s,a) D_{KL}(p(\cdot \mid s,a) \parallel \hat{f}(\cdot \mid s,a)) > 0. \tag{23}$$

Here, the last inequality holds because $\hat{f} \in \mathcal{F}_{\mathcal{E}_z}(\mathcal{G}_z)$ and $p \notin \mathcal{F}_{\mathcal{E}_z}(\mathcal{G}_z)$ and thus $p \neq \hat{f}$. By Eqs. (21) and (23), the proof is complete. □

Theorem 1 (Identifiability of LCGs)**.** *With Assumptions 1 to 4, let $\{\hat{\mathcal{G}}_z\} \in \operatorname{argmax} \mathcal{S}(\{\mathcal{G}_z, \mathcal{E}_z\}_{z=1}^K)$ for $\lambda > 0$ small enough. Then, each $\hat{\mathcal{G}}_z$ is true LCG on $\mathcal{E}_z$, i.e., $\hat{\mathcal{G}}_z = \mathcal{G}_{\mathcal{E}_z}$.*

Proof. To simplify the notation, let $\mathcal{G}_z^*$ be a true LCG on $\mathcal{E}_z$ for all z, i.e., $\mathcal{G}_z^* \coloneqq \mathcal{G}_{\mathcal{E}_z}$ for brevity. It is enough to show that $\mathcal{S}(\{\mathcal{G}_z^*, \mathcal{E}_z\}_{z=1}^K) > \mathcal{S}(\{\mathcal{G}_z, \mathcal{E}_z\}_{z=1}^K)$ if $\mathcal{G}_z^* \neq \mathcal{G}_z$ for some z.

Now, by Lemma 2,

$$
\begin{aligned}
&\mathcal{S}(\{\mathcal{G}^*_z, \mathcal{E}_z\}_{z=1}^K) - \mathcal{S}(\{\mathcal{G}_z, \mathcal{E}_z\}_{z=1}^K) && (24)\\
&= \inf_{\phi} D_{KL}(p \parallel \hat{p}_{\{\mathcal{G}_z, \mathcal{E}_z\}, \phi}) + \lambda \sum_z p(\mathcal{E}_z)(|\mathcal{G}_z| - |\mathcal{G}^*_z|) && (25)\\
&= \inf_{\phi} \int p(s,a) D_{KL}(p(\cdot \mid s,a) \parallel \hat{p}_{\{\mathcal{G}_z, \mathcal{E}_z\}, \phi}(\cdot \mid s,a)) + \lambda \sum_z p(\mathcal{E}_z)(|\mathcal{G}_z| - |\mathcal{G}^*_z|) && (26)\\
&= \inf_{\phi} \sum_z \int_{(s,a)\in\mathcal{E}_z} p(s,a) D_{KL}(p(\cdot \mid s,a) \parallel \hat{p}_{\mathcal{G}_z, \phi_z}(\cdot \mid s,a)) + \lambda \sum_z p(\mathcal{E}_z)(|\mathcal{G}_z| - |\mathcal{G}^*_z|) && (27)\\
&= \sum_z \inf_{\phi_z} p(\mathcal{E}_z) \int p_z(s,a) D_{KL}(p(\cdot \mid s,a) \parallel \hat{p}_{\mathcal{G}_z, \phi_z}(\cdot \mid s,a)) + \lambda \sum_z p(\mathcal{E}_z)(|\mathcal{G}_z| - |\mathcal{G}^*_z|) && (28)\\
&= \sum_z p(\mathcal{E}_z) \inf_{\phi_z} D_{KL}(p_z \parallel \hat{p}_{\mathcal{G}_z, \phi_z}) + \lambda \sum_z p(\mathcal{E}_z)(|\mathcal{G}_z| - |\mathcal{G}^*_z|) && (29)\\
&= \sum_z p(\mathcal{E}_z) \left[\inf_{\phi_z} D_{KL}(p_z \parallel \hat{p}_{\mathcal{G}_z, \phi_z}) + \lambda(|\mathcal{G}_z| - |\mathcal{G}^*_z|) \right] = \sum_z p(\mathcal{E}_z) \cdot A_z. && (30)
\end{aligned}
$$

For brevity, we denote $D_{KL}(p_z \parallel \hat{p}_{\mathcal{G}_z, \phi_z}) \coloneqq \int p_z(s,a) D_{KL}(p(\cdot \mid s,a) \parallel \hat{p}_{\mathcal{G}_z, \phi_z}(\cdot \mid s,a))$ and $A_z \coloneqq \inf_{\phi_z} D_{KL}(p_z \parallel \hat{p}_{\mathcal{G}_z, \phi_z}) + \lambda(|\mathcal{G}_z| - |\mathcal{G}^*_z|)$. Now, we will show that for all $z \in [K]$, $A_z > 0$ if and only if $\mathcal{G}^*_z \neq \mathcal{G}_z$.

Case 0: $\mathcal{G}^*_z = \mathcal{G}_z$. Clearly, $A_z = 0$ in this case.

Case 1: $\mathcal{G}^*_z \subsetneq \mathcal{G}_z$. Then, $|\mathcal{G}_z| > |\mathcal{G}^*_z|$ and thus $A_z > 0$ since $\lambda(|\mathcal{G}_z| - |\mathcal{G}^*_z|) > 0$.

Case 2: $\mathcal{G}^*_z \not\subseteq \mathcal{G}_z$. In this case, there exists $(i \to j) \in \mathcal{G}^*_z$ such that $(i \to j) \notin \mathcal{G}_z$. Thus, $S'_j \perp\!\!\!\perp_{\mathcal{G}_z} X_i \mid \mathbf{X} \setminus \{X_i\}$ and $S'_j \not\perp\!\!\!\perp_{\mathcal{G}^*_z} X_i \mid \mathbf{X} \setminus \{X_i\}$. Therefore, $S'_j \not\perp\!\!\!\perp_p X_i \mid \mathbf{X} \setminus \{X_i\}, \mathcal{E}_z$ by Assumption 2. Thus, $p \notin \mathcal{F}_{\mathcal{E}_z}(\mathcal{G}_z)$ and we have $\inf_{\phi_z} D_{KL}(p_z \parallel \hat{p}_{\mathcal{G}_z, \phi_z}) > 0$ by Lemma 3.

Now, we consider two subcases: (i) $\mathcal{G}_z \in \mathbb{G}^+_z \coloneqq \{\mathcal{G}' \mid \mathcal{G}^*_z \not\subseteq \mathcal{G}', |\mathcal{G}'| \geq |\mathcal{G}^*_z|\}$, and (ii) $\mathcal{G}_z \in \mathbb{G}^-_z \coloneqq \{\mathcal{G}' \mid \mathcal{G}^*_z \not\subseteq \mathcal{G}', |\mathcal{G}'| < |\mathcal{G}^*_z|\}$. Clearly, if $\mathcal{G}_z \in \mathbb{G}^+_z$ then $A_z > 0$. Suppose $\mathcal{G}_z \in \mathbb{G}^-_z$. Then,

$$
\begin{aligned}
&\lambda \leq \eta_z \coloneqq \frac{1}{N(N+M)+1} \min_{\mathcal{G}' \in \mathbb{G}^-_z} \inf_{\phi_z} D_{KL}(p_z \parallel \hat{p}_{\mathcal{G}', \phi_z}) && (31)\\
\implies \quad &\lambda \leq \frac{\inf_{\phi_z} D_{KL}(p_z \parallel \hat{p}_{\mathcal{G}', \phi_z})}{N(N+M)+1} < \frac{\inf_{\phi_z} D_{KL}(p_z \parallel \hat{p}_{\mathcal{G}', \phi_z})}{|\mathcal{G}^*_z| - |\mathcal{G}'|} \quad \text{for } \forall \mathcal{G}' \in \mathbb{G}^-_z && (32)\\
\implies \quad &\inf_{\phi_z} D_{KL}(p_z \parallel \hat{p}_{\mathcal{G}', \phi_z}) + \lambda(|\mathcal{G}_z| - |\mathcal{G}^*_z|) > 0 \quad \text{for } \forall \mathcal{G}' \in \mathbb{G}^-_z. && (33)
\end{aligned}
$$

Here, we use the fact that $|\mathcal{G}^*_z| - |\mathcal{G}'| \leq |\mathcal{G}^*_z| < N(N+M)+1$. Therefore, for $0 < \forall \lambda \leq \eta_z$, we have $A_z > 0$ if $\mathcal{G}^*_z \neq \mathcal{G}_z$. Here, we note that $\eta_z > 0$ for all z, since $\mathbb{G}^-_z$ is finite and $\inf_{\phi_z} D_{KL}(p_z \parallel \hat{p}_{\mathcal{G}', \phi_z}) > 0$ for any $\mathcal{G}' \in \mathbb{G}^-_z$ by Lemma 3.

Consequently, for $0 < \lambda \leq \eta(\{\mathcal{E}_z\}) \coloneqq \min_z \eta_z$, we have $\mathcal{S}(\{\mathcal{G}^*_z, \mathcal{E}_z\}_{z=1}^K) - \mathcal{S}(\{\mathcal{G}_z, \mathcal{E}_z\}_{z=1}^K) > 0$ if $\mathcal{G}^*_z \neq \mathcal{G}_z$ for some z. We also note that $\eta(\{\mathcal{E}_z\}) > 0$ since $\eta_z > 0$ for all z. □

B.3. Proof of Prop. 1

Definition 7. *Let $\mathcal{T} \coloneqq \{\{\mathcal{E}_z\}_{z=1}^K\}$, i.e., a set of all decompositions of size K.*

Definition 8. *Let $\mathcal{T}_\lambda \coloneqq \{\{\mathcal{E}_z\}_{z=1}^K \mid \eta(\{\mathcal{E}_z\}) \geq \lambda\}$.*

Remark 1. *$\mathcal{T}_\lambda \to \mathcal{T} (= \mathcal{T}_0)$ as $\lambda \to 0$.*

Recall that Thm. 1 holds for $0 < \lambda \leq \eta(\{\mathcal{E}_z\})$. Here, $\eta(\{\mathcal{E}_z\})$ is the value corresponding to the specific decomposition $\{\mathcal{E}_z\}$. For the arguments henceforth, we consider the arbitrary decomposition and thus introduce the following assumption.

Assumption 5. $\inf_{\{\mathcal{E}_z\}\in\mathcal{T}} \eta(\{\mathcal{E}_z\}) > 0$.

Note that $\eta(\{\mathcal{E}_z\}) > 0$ for any $\{\mathcal{E}_z\}$, and thus $\inf_{\{\mathcal{E}_z\}\in\mathcal{T}} \eta(\{\mathcal{E}_z\}) \geq 0$. We now take $0 < \lambda \leq \inf_{\{\mathcal{E}_z\}\in\mathcal{T}} \eta(\{\mathcal{E}_z\})$ with Assumption 5, which allows Thm. 1 to hold on any arbitrary decomposition. It is worth noting that this assumption is purely technical because for a small fixed $\lambda > 0$, the arguments henceforth hold for all $\{\mathcal{E}_z\} \in \mathcal{T}_\lambda$, where $\mathcal{T}_\lambda \to \mathcal{T}$ as $\lambda \to 0$.

Proposition 1. *Let $\{\mathcal{G}_z^*, \mathcal{E}_z^*\} \in \operatorname{argmax} \mathcal{S}(\{\mathcal{G}_z, \mathcal{E}_z\}_{z=1}^K)$ for $\lambda > 0$ small enough, with Assumptions 1 to 5. Then, (i) each $\mathcal{G}_z^*$ is true LCG on $\mathcal{E}_z^*$, and (ii) $\mathbb{E}\left[|\mathcal{G}_z^*|\right] \leq \mathbb{E}\left[|\mathcal{G}_z|\right]$ where $\{\mathcal{G}_z\}$ are LCGs on arbitrary decomposition $\{\mathcal{E}_z\}_{z=1}^K$.*

Proof. Let $0 < \lambda \leq \inf_{\{\mathcal{E}_z\}\in\mathcal{T}} \eta(\{\mathcal{E}_z\})$. (i) First, $\{\mathcal{G}_z^*, \mathcal{E}_z^*\}_{z=1}^K \in \operatorname{argmax}_{\{\mathcal{G}_z, \mathcal{E}_z\}} \mathcal{S}(\{\mathcal{G}_z, \mathcal{E}_z\}_{z=1}^K)$ implies that $\{\mathcal{G}_z^*\}_{z=1}^K$ also maximizes the score on the fixed $\{\mathcal{E}_z^*\}_{z=1}^K$, i.e., $\{\mathcal{G}_z^*\}_{z=1}^K \in \operatorname{argmax}_{\{\mathcal{G}_z\}} \mathcal{S}(\{\mathcal{G}_z, \mathcal{E}_z^*\}_{z=1}^K)$. Thus, each $\mathcal{G}_z^*$ is true LCG on $\mathcal{E}_z^*$ by Thm. 1, i.e., $\mathcal{G}_z^* = \mathcal{G}_{\mathcal{E}_z^*}$.

(ii) Also, since $\{\mathcal{E}_z\}_{z=1}^K$ is the arbitrary decomposition, $\mathcal{S}(\{\mathcal{G}_z^*, \mathcal{E}_z^*\}) \geq \mathcal{S}(\{\mathcal{G}_z, \mathcal{E}_z\})$ holds. Since $\{\mathcal{G}_z\}$ is the true LCGs on each $\mathcal{E}_z$, i.e., $\mathcal{G}_z = \mathcal{G}_{\mathcal{E}_z}$, by Lemma 1,

$$\mathcal{S}(\{\mathcal{G}_z, \mathcal{E}_z\}_{z=1}^K) = \mathbb{E}_{p(s,a,s')} \log p(s' \mid s, a) - \lambda \sum_z p(\mathcal{E}_z) \cdot |\mathcal{G}_z|. \tag{34}$$

Similarly, since $\{\mathcal{G}_z^*\}$ is the true LCGs on each $\mathcal{E}_z^*$,

$$\mathcal{S}(\{\mathcal{G}_z^*, \mathcal{E}_z^*\}_{z=1}^K) = \mathbb{E}_{p(s,a,s')} \log p(s' \mid s, a) - \lambda \sum_z p(\mathcal{E}_z^*) \cdot |\mathcal{G}_z^*|. \tag{35}$$

Therefore, $0 \leq \mathcal{S}(\{\mathcal{G}_z^*, \mathcal{E}_z^*\}) - \mathcal{S}(\{\mathcal{G}_z, \mathcal{E}_z\}) = \mathbb{E}\left[|\mathcal{G}_z|\right] - \mathbb{E}\left[|\mathcal{G}_z^*|\right]$ holds, and thus $\mathbb{E}\left[|\mathcal{G}_z^*|\right] \leq \mathbb{E}\left[|\mathcal{G}_z|\right]$. □

B.4. Proof of Thm. 2

We first provide some useful lemma.

Lemma 4 (Hwang et al. (2023), Prop. 4)**.** *$S'_j \perp\!\!\!\perp \mathbf{X} \setminus Pa(j;\mathcal{E}) \mid Pa(j;\mathcal{E}), \mathcal{F}$ holds for any $\mathcal{F} \subseteq \mathcal{E}$.*

Lemma 5 (Monotonicity)**.** *Let $\mathcal{F} \subseteq \mathcal{E}$. Then, $\mathcal{G}_\mathcal{F} \subseteq \mathcal{G}_\mathcal{E}$.*

Proof. Since $S'_j \perp\!\!\!\perp \mathbf{X} \setminus Pa(j;\mathcal{E}) \mid Pa(j;\mathcal{E}), \mathcal{F}$ holds by Lemma 4, $Pa(j;\mathcal{F}) \subseteq Pa(j;\mathcal{E})$ holds by definition; otherwise, $Pa(j;\mathcal{F}) \setminus Pa(j;\mathcal{E}) \neq \emptyset$ which leads to contradiction. Therefore, $Pa(j;\mathcal{F}) \subseteq Pa(j;\mathcal{E})$ for all j and thus $\mathcal{G}_\mathcal{F} \subseteq \mathcal{G}_\mathcal{E}$. □

Now, we provide a proof of Thm. 2.

Definition 9. *The context $\mathcal{D} \subset \mathcal{X}$ is canonical if $\mathcal{G}_\mathcal{F} = \mathcal{G}_\mathcal{D}$ for any $\mathcal{F} \subset \mathcal{D}$.*

Theorem 2 (Identifiability of contexts)**.** *Let $\{\mathcal{G}_z^*, \mathcal{E}_z^*\} \in \operatorname{argmax} \mathcal{S}(\{\mathcal{G}_z, \mathcal{E}_z\}_{z=1}^K)$ for $\lambda > 0$ small enough, with Assumptions 1 to 5. Suppose $\mathcal{X} = \cup_{m\in[H]}\mathcal{D}_m$ where $\mathcal{G}_{\mathcal{D}_m}$ is distinct for all $m \in [H]$, and $\mathcal{D}_1, \cdots, \mathcal{D}_H$ are disjoint and canonical. Suppose $K \geq H$. Then, for all $m \in [H]$, there exists $I_m \subset [K]$ such that $\mathcal{D}_m = \bigcup_{z\in I_m} \mathcal{E}_z^*$ almost surely.*

Proof. Let $\{\mathcal{F}_z\}_{z=1}^K$ be the decomposition such that for all $m \in [H]$, $\bigcup_{z\in J_m} \mathcal{F}_z = \mathcal{D}_m$ for some $J_m \subset [K]$. Note that such decomposition exists since $K \geq H$. Let $\{\mathcal{G}_z\}_{z=1}^K$ be the true LCGs corresponding to each $\mathcal{F}_z$, i.e., $\mathcal{G}_z = \mathcal{G}_{\mathcal{F}_z}$. Recall that $\mathbb{E}\left[|\mathcal{G}_z^*|\right] \leq \mathbb{E}\left[|\mathcal{G}_z|\right]$ holds by Prop. 1, we have

$$\begin{aligned}
0 \leq \mathbb{E}\left[|\mathcal{G}_z|\right] - \mathbb{E}\left[|\mathcal{G}_z^*|\right] &= \sum_i p(\mathcal{F}_i)|\mathcal{G}_i| - \sum_j p(\mathcal{E}_j^*)|\mathcal{G}_j^*| \\
&= \sum_{i,j} p(\mathcal{F}_i \cap \mathcal{E}_j^*)(|\mathcal{G}_i| - |\mathcal{G}_j^*|).
\end{aligned} \tag{36}$$

Suppose $p(\mathcal{F}_i \cap \mathcal{E}_j^*) > 0$ for some i, j. Let $\mathcal{C}_{ij} \coloneqq \mathcal{F}_i \cap \mathcal{E}_j^*$. Since $\mathcal{F}_i \subset \mathcal{D}_m$ for some m and $\mathcal{D}_m$ is canonical, $\mathcal{F}_i$ is also canonical. Therefore, $\mathcal{G}_i = \mathcal{G}_{\mathcal{C}_{ij}}$ since $\mathcal{C}_{ij} \subset \mathcal{F}_i$. Since $\mathcal{C}_{ij} \subset \mathcal{E}_j^*$, we have $\mathcal{G}_{\mathcal{C}_{ij}} \subseteq \mathcal{G}_j^*$ by Lemma 5. Therefore, we have $\mathcal{G}_i \subseteq \mathcal{G}_j^*$. Therefore, $|\mathcal{G}_i| - |\mathcal{G}_j^*| \leq 0$ for any i, j such that $p(\mathcal{F}_i \cap \mathcal{E}_j^*) > 0$. Thus, by Eq. (36), $|\mathcal{G}_i| = |\mathcal{G}_j^*|$ if $p(\mathcal{F}_i \cap \mathcal{E}_j^*) > 0$. Since $\mathcal{G}_i \subseteq \mathcal{G}_j^*$ if $p(\mathcal{F}_i \cap \mathcal{E}_j^*) > 0$, we conclude that

$$\mathcal{G}_i = \mathcal{G}_j^* \quad \text{if} \quad p(\mathcal{F}_i \cap \mathcal{E}_j^*) > 0. \tag{37}$$

Now, for arbitrary $\mathcal{E}_j^*$, suppose there exist $s \neq t$ such that $p(\mathcal{D}_s \cap \mathcal{E}_j^*) > 0$ and $p(\mathcal{D}_t \cap \mathcal{E}_j^*) > 0$. Then, there exist some $\mathcal{F}_i \subset \mathcal{D}_s$ and $\mathcal{F}_k \subset \mathcal{D}_t$ such that $p(\mathcal{F}_i \cap \mathcal{E}_j^*) > 0$ and $p(\mathcal{F}_k \cap \mathcal{E}_j^*) > 0$. By Eq. (37), we have $\mathcal{G}_i = \mathcal{G}_j^* = \mathcal{G}_k$. Also, $\mathcal{G}_{\mathcal{D}_s} = \mathcal{G}_i$ and $\mathcal{G}_{\mathcal{D}_t} = \mathcal{G}_k$ since $\mathcal{D}_s, \mathcal{D}_t$ are canonical. Therefore, we have $\mathcal{G}_{\mathcal{D}_s} = \mathcal{G}_{\mathcal{D}_t}$, which contradicts that $\mathcal{G}_{\mathcal{D}_m}$ is distinct

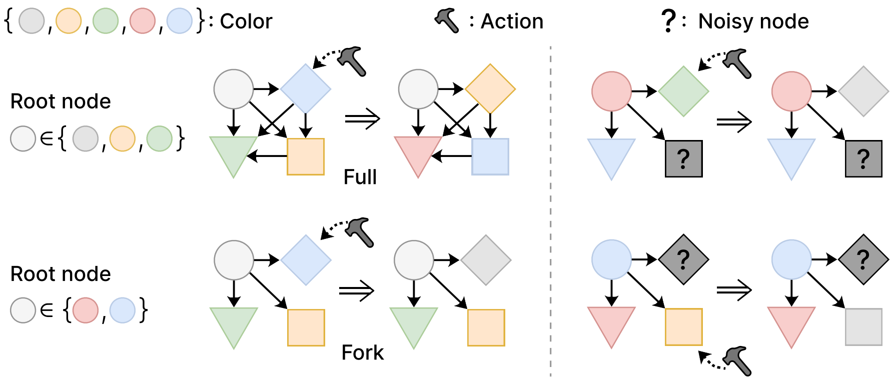


Figure 8. Illustration of Chemical (*full-fork*) environment with 4 nodes. (**Left**) the color of the root node determines the activation of local causal graph *fork*. (**Right**) the noisy nodes are redundant for predicting the colors of other nodes under the local causal graph.

for all m. Therefore, for any $\mathcal{E}_j^*$, there exists a unique $\mathcal{D}_m$ such that $p(\mathcal{D}_m \cap \mathcal{E}_j^*) > 0$, which leads $p(\mathcal{E}_j^* \setminus \mathcal{D}_m) = 0$ since $\{\mathcal{D}_m\}_{m\in[H]}$ is a decomposition of $\mathcal{X}$. Let $I_m = \{j \in [K] \mid p(\mathcal{D}_m \cap \mathcal{E}_j^*) > 0\}$. Here, we have

$$p\left(\bigcup_{z\in I_m} \mathcal{E}_z^* \setminus \mathcal{D}_m\right) = \sum_{z\in I_m} p(\mathcal{E}_z^* \setminus \mathcal{D}_m) = 0. \tag{38}$$

Also, by the definition of I_m and because $\{\mathcal{E}_z^*\}_{z\in[K]}$ is a decomposition of $\mathcal{X}$, we have

$$p\left(\mathcal{D}_m \setminus \bigcup_{z\in I_m} \mathcal{E}_z^*\right) = 0. \tag{39}$$

Therefore, by Eqs. (38) and (39), we have $\mathcal{D}_m = \bigcup_{z\in I_m} \mathcal{E}_z^*$ almost surely for all $m \in [H]$. □

C. Appendix for Experiments

C.1. Environment Details

Table 4. Environment configurations.

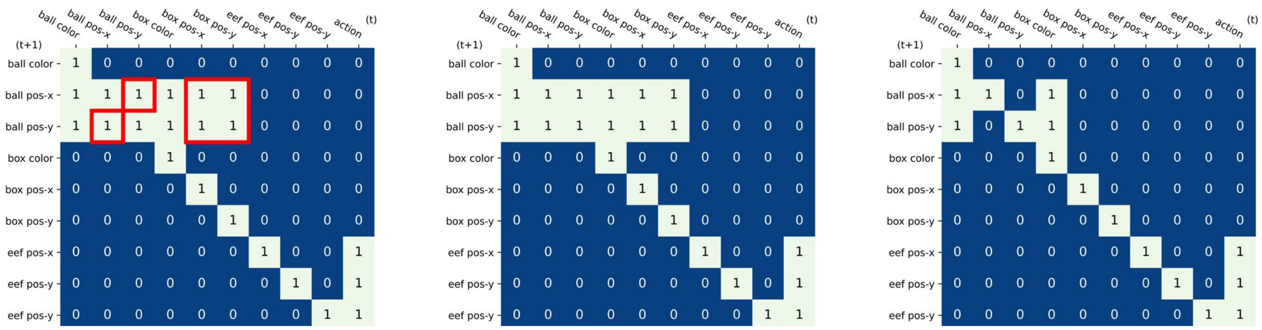
	Chemical		Magnetic
Parameters	*full-fork*	*full-chain*	
Training step	1.5×10^5	1.5×10^5	2×10^5
Optimizer	Adam	Adam	Adam
Learning rate	1e-4	1e-4	1e-4
Batch size	256	256	256
Initial step	1000	1000	2000
Max episode length	25	25	25
Action type	Discrete	Discrete	Continuous

Table 5. CEM parameters.

	Chemical		Magnetic
CEM parameters	*full-fork*	*full-chain*	
Planning length	3	3	1
Number of candidates	64	64	64
Number of top candidates	32	32	32
Number of iterations	5	5	5
Exploration noise	N/A	N/A	1e-4
Exploration probability	0.05	0.05	N/A

C.1.1. Chemical

Here, we describe two settings, namely *full-fork* and *full-chain*, modified from Ke et al. (2021). In both settings, there are 10 state variables representing the color of corresponding nodes, with each color represented as a one-hot encoding. The action variable is a 50-dimensional categorical variable that changes the color of a specific node to a new color (e.g., changing the color of the third node to blue). According to the underlying causal graph and pre-defined conditional probability distributions, implemented with randomly initialized neural networks, an action changes the colors of the intervened object's descendants as depicted in Fig. 8. As shown in Fig. 3(a), the (global) causal graph is *full* in both settings, and the LCG is



(t+1) \ (t)	ball color	ball pos-x	ball pos-y	box color	box pos-x	box pos-y	eef pos-x	eef pos-y	eef pos-y	action
ball color	1	0	0	0	0	0	0	0	0	0
ball pos-x	1	1	1	1	1	1	0	0	0	0
ball pos-y	1	1	1	1	1	1	0	0	0	0
box color	0	0	0	1	0	0	0	0	0	0
box pos-x	0	0	0	0	1	0	0	0	0	0
box pos-y	0	0	0	0	0	1	0	0	0	0
eef pos-x	0	0	0	0	0	0	1	0	0	1
eef pos-y	0	0	0	0	0	0	0	1	0	1
eef pos-y	0	0	0	0	0	0	0	0	1	1

(a)

(t+1) \ (t)	ball color	ball pos-x	ball pos-y	box color	box pos-x	box pos-y	eef pos-x	eef pos-y	eef pos-y	action
ball color	1	0	0	0	0	0	0	0	0	0
ball pos-x	1	1	1	1	1	1	0	0	0	0
ball pos-y	1	1	1	1	1	1	0	0	0	0
box color	0	0	0	1	0	0	0	0	0	0
box pos-x	0	0	0	0	1	0	0	0	0	0
box pos-y	0	0	0	0	0	1	0	0	0	0
eef pos-x	0	0	0	0	0	0	1	0	0	1
eef pos-y	0	0	0	0	0	0	0	1	0	1
eef pos-y	0	0	0	0	0	0	0	0	1	1

(b)

(t+1) \ (t)	ball color	ball pos-x	ball pos-y	box color	box pos-x	box pos-y	eef pos-x	eef pos-y	eef pos-y	action
ball color	1	0	0	0	0	0	0	0	0	0
ball pos-x	1	1	0	1	0	0	0	0	0	0
ball pos-y	1	0	1	1	0	0	0	0	0	0
box color	0	0	0	1	0	0	0	0	0	0
box pos-x	0	0	0	0	1	0	0	0	0	0
box pos-y	0	0	0	0	0	1	0	0	0	0
eef pos-x	0	0	0	0	0	0	1	0	0	1
eef pos-y	0	0	0	0	0	0	0	1	0	1
eef pos-y	0	0	0	0	0	0	0	0	1	1

(c)

Figure 9. (a) Causal graph of Magnetic environment. Red boxes indicate redundant edges under the non-magnetic context. (b) LCG under the magnetic context, which is the same as global CG. (c) LCG under the non-magnetic context.

fork and *chain*, respectively. For example in *full-fork*, the LCG *fork* is activated according to the particular color of the root node, as shown in Fig. 8.

In both settings, the task is to match the colors of each node to the given target. The reward function is defined as:

$$r = \frac{1}{|O|} \sum_{i \in O} \mathbb{1} \left[s_i = g_i \right], \tag{40}$$

where O is a set of the indices of observable nodes, s_i is the current color of the i-th node, and g_i is the target color of the i-th node in this episode. Success is determined if all colors of observable nodes are the same as the target. During training, all 10 nodes are observable, i.e., $O = \{0, \cdots, 9\}$. In downstream tasks, the root color is set to induce the LCG, and the agent receives noisy observations for a subset of nodes, aiming to match the colors of the rest of the observable nodes. As shown in Fig. 8, noisy nodes are spurious for predicting the colors of other nodes under the LCG. Thus, the agent capable of reasoning the fine-grained causal relationships would generalize well in downstream tasks. Note that the transition dynamics of the environment is the same in training and downstream tasks. To create noisy observations, we use a noise sampled from $\mathcal{N}(0, \sigma^2)$, similar to Wang et al. (2022), where the noise is multiplied to the one-hot encoding representing color during the test. In our experiments, we use $\sigma = 100$.

As the root color determines the local causal graph in both settings, the root node is always observable to the agent during the test. The root colors of the initial state and the goal state are the same, inducing the local causal graph. As the root color can be changed by the action during the test, this may pose a challenge in evaluating the agent's reasoning of local causal relationships. This can be addressed by modifying the initial distribution of CEM to exclude the action on the root node and only act on the other nodes during the test. Nevertheless, we observe that restricting the action on the root during the test has little impact on the behavior of any model, and we find that this is because the agent rarely changes the root color as it already matches the goal color in the initial state.

C.1.2. Magnetic

In this environment, there are two objects on a table, a moving ball and a box, colored either red or black, as shown in Fig. 3(b). The red color indicates that the object is *magnetic*. In other words, when they are both colored red, magnetic force will be applied and the ball will move toward the box. If one of the objects is colored black, the ball would not move since the box has no influence on the ball.

The state consists of the color, x, y position of each object, and x, y, z position of the end-effector of the robot arm, where the color is given as the 2-dimensional one-hot encoding. The action is a 3-dimensional vector that moves the robot arm. The causal graph of the Magnetic environment is shown in Fig. 9(a). LCGs under magnetic and non-magnetic context are shown in Figs. 9(b) and 9(c), respectively. The table in our setup has a width of 0.9 and a length of 0.6, with the y-axis defined by the width and the x-axis defined by the length. For each episode, the initial positions of a moving ball and a box are randomly sampled within the range of the table.

The task is to move the robot arm to reach the moving ball. Thus, accurately predicting the trajectory of the ball is crucial. The reward function is defined as:

$$r = 1 - \tanh(5 \cdot \|eef - g\|_1), \tag{41}$$

where the $eef \in \mathbb{R}^3$ is the current position of the end-effector, $g = (b_x, b_y, 0.8) \in \mathbb{R}^3$, and (b_x, b_y) is the current position of the moving ball. Success is determined if the distance is smaller than 0.05. During the test, the color of one of the objects is black and the box is located at the position unseen during the training. Specifically, the box position is sampled from $\mathcal{N}(0, \sigma^2)$ during the test. Note that the box can be located outside of the table, which never happens during the training. In our experiments, we use $\sigma = 100$.

C.2. Experimental Details

To assess the performance of different dynamics models of the baselines and our method, we use a model predictive control (MPC) (Camacho & Alba, 2013) which selects the actions based on the prediction of the learned dynamics model, following prior works (Ding et al., 2022; Wang et al., 2022). Specifically, we use a cross-entropy method (CEM) (Rubinstein & Kroese, 2004), which iteratively generates and refines action sequences through a process of sampling from a probability distribution that is updated based on the performance of these sampled sequences, with a known reward function. We use a random policy for the initial data collection. Environmental configurations and CEM parameters are shown in Tables 4 and 5, respectively. Most of the experiments were processed using a single NVIDIA RTX 3090. For Fig. 7, we use structural hamming distance (SHD) for evaluation, which is a metric used to quantify the dissimilarity between two graphs based on the number of edge additions or deletions needed to make the graphs identical (Acid & de Campos, 2003; Ramsey et al., 2006).

C.3. Implementation of Baselines

For all methods, the dynamics model outputs the parameters of categorical distribution for discrete variables, and the mean and standard deviation of normal distribution for continuous variables. All methods have a similar number of model parameters for a fair comparison. Detailed parameters of each model are shown in Table 6.

MLP and Modular. MLP models the transition dynamics as $p(s' \mid s, a)$. Modular has a separate network for each state variable, i.e., $\prod_j p(s'_j \mid s, a)$, where each network is implemented as an MLP.

GNN, NPS, and CDL. We employ publicly available source codes.[4] For NPS (Goyal et al., 2021a), we search the number of rules $N \in \{4, 15, 20\}$. CDL (Wang et al., 2022) infers the causal structure by estimating conditional mutual information (CMI) and models the dynamics as $\prod_j p(s'_j \mid Pa(j))$. For CDL, we search the initial CMI threshold $\epsilon \in \{0.001, 0.002, 0.005, 0.01, 0.02\}$ and exponential moving average (EMA) coefficient $\tau \in \{0.9, 0.95, 0.99, 0.999\}$. As CDL is a two-stage method, we only report their final performance.

GRADER. We implement GRADER (Ding et al., 2022) based on the code provided by the authors.[5] GRADER relies on the conditional independence test (CIT) to discover the causal structure. In Chemical, we ran the CIT for every 10 episodes, following their default setting. We only report its performance in Chemical due to the poor scalability of the conditional independence test in Magnetic environment, which took about 30 minutes for each test.

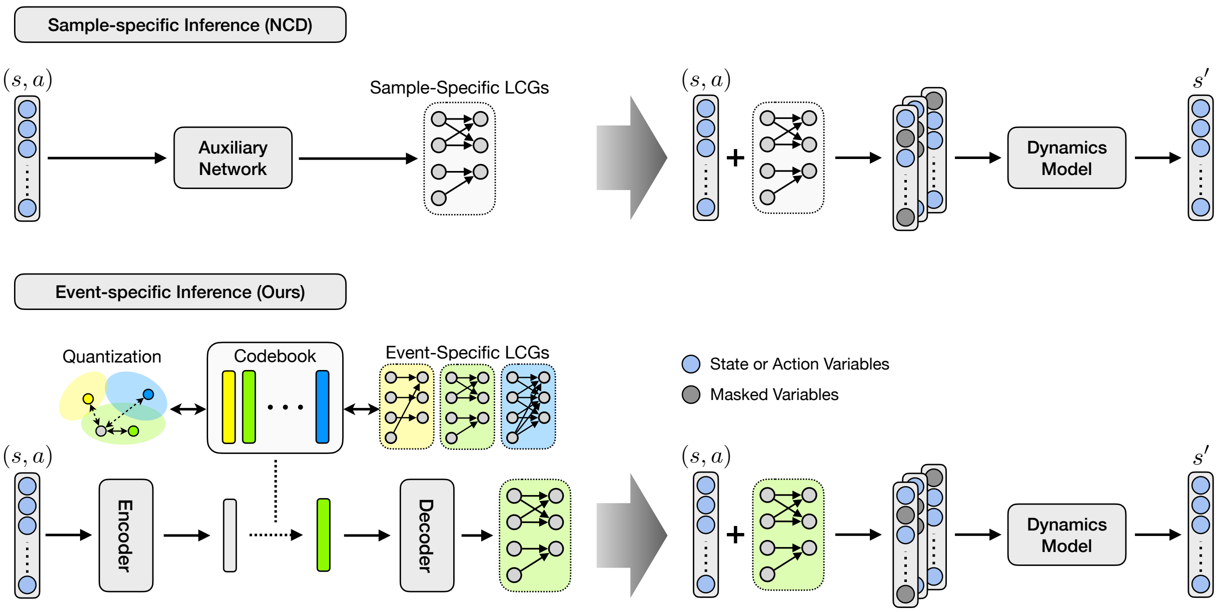
Oracle and NCD. For a fair comparison, we employ the same architecture for the dynamic models of Oracle, NCD, and our method, as their main difference lies in the inference of local causal graphs (LCG). As illustrated in Fig. 10, the key difference is that NCD (Hwang et al., 2023) performs direct inference of the LCG from each individual sample (referred to as *sample-specific* inference), while our method decomposes the data domain and infers the LCGs for each subgroup through quantization. We provide an implementation details of our method in the next subsection.

C.4. Implementation of FCDL

For our method, we use MLPs for the implementation of $g_{\texttt{enc}}, g_{\texttt{dec}}$, and $\hat{p}$, with configurations provided in Table 6. The quantization encoder $g_{\texttt{enc}}$ of our method or the auxiliary network of NCD shares the initial feature extraction layer with the dynamics model $\hat{p}$ as we found that it yields better performance compared to full decoupling of them.

[4] https://github.com/wangzizhao/CausalDynamicsLearning

[5] https://github.com/GilgameshD/GRADER



(a) Local Causal Graph Inference (b) Masked Prediction with LCG

Figure 10. Comparison of the sample-specific inference of NCD **(top)** and quantization-based inference of our method **(bottom)**.

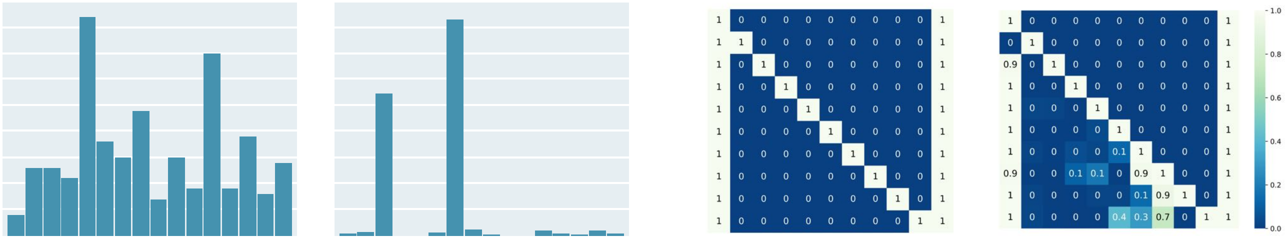
C.4.1. Dynamics Model

Recall our dynamics modeling in Eq. (8) that $\hat{p}(s'_j \mid Pa^{\mathcal{G}_z}(j); \phi_z^{(j)})$ if $(s, a) \in \mathcal{E}_z$, which corresponds to $p(s'_j \mid s, a) = p(s'_j \mid Pa(j; \mathcal{E}_z), z)$ in Eq. (3). Here, each $\phi_z^{(j)}$ is a neural network that takes $Pa^{\mathcal{G}_z}(j)$ as an input and predicts s'_j under $\mathcal{E}_z$. In general, this separate network for each subgroup would allow it to effectively adapt to environments with complex dynamics and learn transition functions separately for each subgroup. However, this requires a total of $K \times N$ separate networks, which could incur a computational burden. Instead, we employ an efficient parameter-sharing mechanism to simplify the model implementation: we let the dynamics model consist of separate networks for each state variable, i.e., $\phi = \{\phi^{(j)}\}$ and each $\phi^{(j)}$ takes $(Pa^{\mathcal{G}_z}(j), z)$ as an input, instead of using separate networks $\phi_z^{(j)}$ for each $\mathcal{E}_z$, which is analogous to $p(s'_j \mid Pa(j; \mathcal{E}_z), z)$. This requires a total of N separate networks, one for each state variable. There are different implementation design choices for z in $(Pa^{\mathcal{G}_z}(j), z)$. We consider two cases: (i) concatenation of $Pa^{\mathcal{G}_z}(j)$ and e_z (i.e., code), and (ii) concatenation of $Pa^{\mathcal{G}_z}(j)$ and one-hot encoding of z (dimension of K). We opt for a simpler choice of the latter. This allows us to model (possibly) different transition functions for each subgroup with a single dynamics model for each state variable. Note that if the subgroups having the same LCG share the same transition function, such labeling of z could be further omitted.

For the implementation of taking $Pa^{\mathcal{G}_z}(j)$ as input for $\hat{p}(s'_j \mid Pa^{\mathcal{G}_z}(j); \phi_z^{(j)})$, we simply mask out the features of unused variables, but other design choices such as Gated Recurrent Unit (Chung et al., 2014; Ding et al., 2022) are also possible. As architectural design is not the primary focus of this work, we leave the exploration of different architectures to future work. Note that all baselines except MLP (e.g., GNN and causal dynamics models) use separate networks for each state variable, and we made sure that all methods have a similar number of model parameters for a fair comparison.

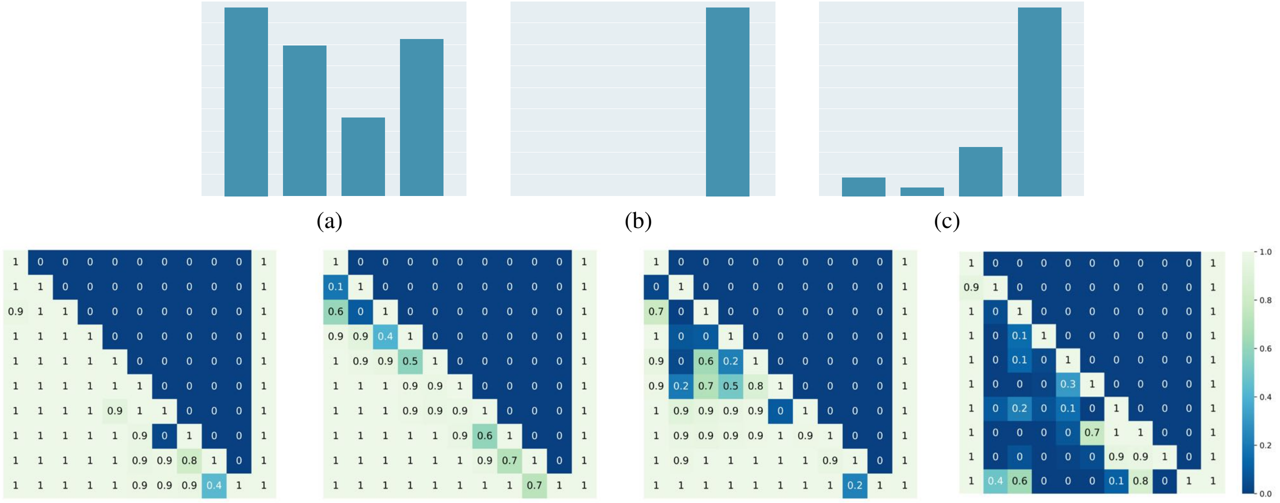
C.4.2. Backpropagation

We now describe how each component of our method is updated by the training objective in Eq. (6). First, $\mathcal{L}_{\texttt{pred}}$ updates the encoder $g_{\texttt{enc}}(s, a)$, decoder $g_{\texttt{dec}}(e)$, and the dynamics model $\hat{p}$. Recall that $A \sim g_{\texttt{dec}}(e)$, backpropagation from A in $\mathcal{L}_{\texttt{pred}}$ updates the quantization decoder $g_{\texttt{dec}}$ through e. During the backward path in Eq. (5), gradients are copied from e (= input of $g_{\texttt{dec}}$) to h (= output of $g_{\texttt{enc}}$), following VQ-VAE (Van Den Oord et al., 2017). By doing so, $\mathcal{L}_{\texttt{pred}}$ also updates the quantization encoder $g_{\texttt{enc}}$ and h. Second, $\mathcal{L}_{\texttt{quant}}$ updates $g_{\texttt{enc}}$ and the codebook C. We note that $\mathcal{L}_{\texttt{pred}}$ also affects the learning of the codebook C since h is updated with $\mathcal{L}_{\texttt{pred}}$. The rationale behind this trick of VQ-VAE is that the gradient $\nabla_e \mathcal{L}_{\texttt{pred}}$ could guide the encoder $g_{\texttt{enc}}$ to change its output $h = g_{\texttt{enc}}(s, a)$ to lower the prediction loss $\mathcal{L}_{\texttt{pred}}$, altering the



(a) (b) (c) (d)

Figure 11. (a,b) Codebook histogram on (a) ID states during training and (b) OOD states during the test in Chemical (*full-fork*). (c) True causal graph of the *fork* structure. (d) Learned LCG corresponding to the most used code in (b).



(d) (e) (f) (g)

Figure 12. Analysis of LCGs learned by our method with quantization degree of 4 in Chemical (*full-fork*) environment. (a-c) Codebook histogram on (a) ID states, (b) ID states on local structure *fork*, and (c) OOD states on local structure. (d-g) Learned LCGs. The descriptions of the histograms are also applied to Figs. 13 to 15, 17 and 18.

quantization (i.e., assignment of the cluster) in the next forward pass. A larger prediction loss (which implies that this sample (s, a) is assigned to the wrong cluster) induces a bigger change on h, and consequently, it would be more likely to cause a re-assignment of the cluster.

C.4.3. Hyperparameters

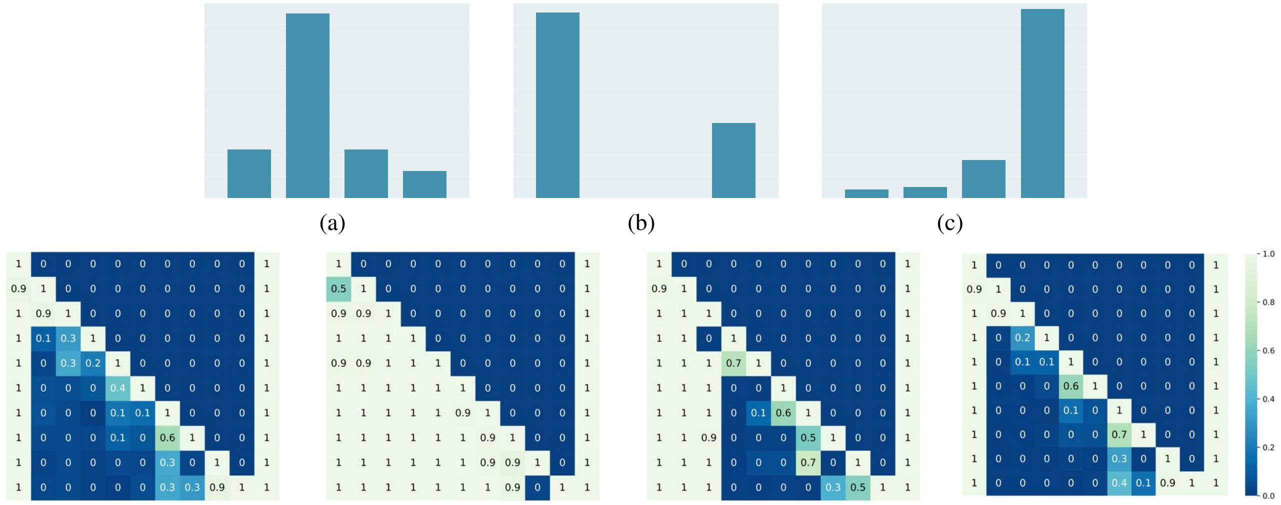
For all experiments, we fix the codebook size $K = 16$, regularization coefficient $\lambda = 0.001$, and commitment coefficient $\beta = 0.25$, as we found that the performance did not vary much for any $K > 2$, $\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}\}$ and $\beta \in \{0.1, 0.25\}$.

C.5. Additional Experimental Results

C.5.1. Detailed analysis of learned LCGs

LCGs learned by our method with a quantization degree of 4 in Chemical are shown in Figs. 12 and 13. Among the 4 codes, one (Fig. 12(b)) or two (Fig. 13(b)) represent the local causal structure *fork*. Our method successfully infers the proper code for most of the OOD samples (Figs. 12(c) and 13(c)). Two sample runs of our method with a quantization degree of 4 in Magnetic are shown in Figs. 14 and 15. Our method successfully learns LCGs correspond to a non-magnetic context (Figs. 14(d), 14(g), 15(d) and 15(f)) and magnetic context (Figs. 14(e), 14(f), 15(e) and 15(g)).

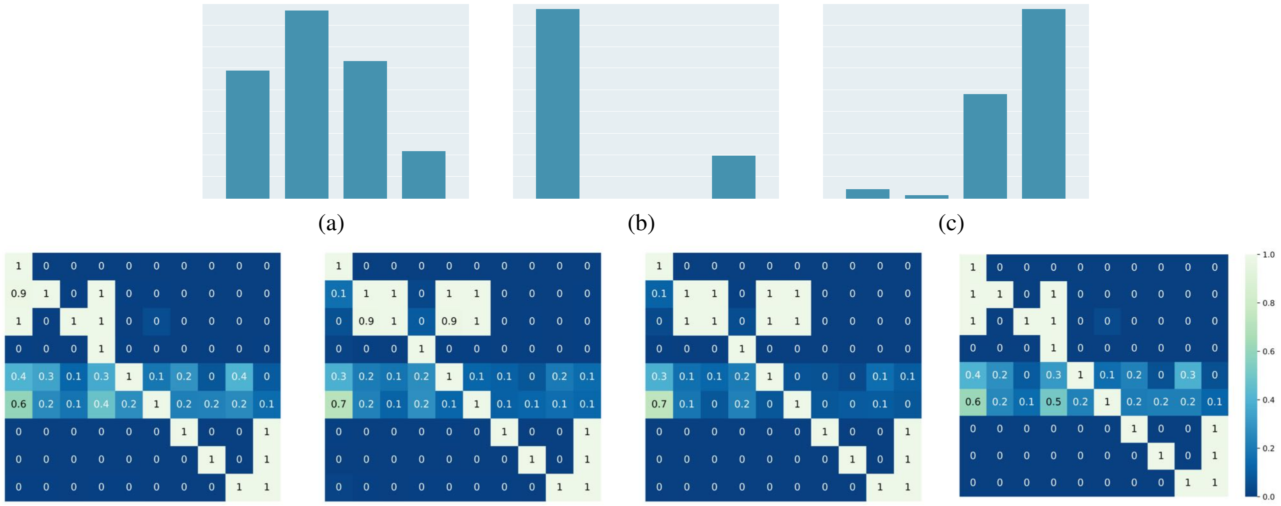
We also observe that our method discovers more fine-grained relationships. Recall that the non-magnetic context is determined when one of the objects is black, the box would have no influence on the ball regardless of the color of the box when the ball is black, and vice versa. As shown in Fig. 16, our method discovers the context where the ball is black (Fig. 16(b)), and the context where the box is black (Fig. 16(a)).



(a) (b) (c)

(d) (e) (f) (g)

Figure 13. Another sample run of our method with quantization degree of 4 in Chemical (*full-fork*).



(a) (b) (c)

(d) (e) (f) (g)

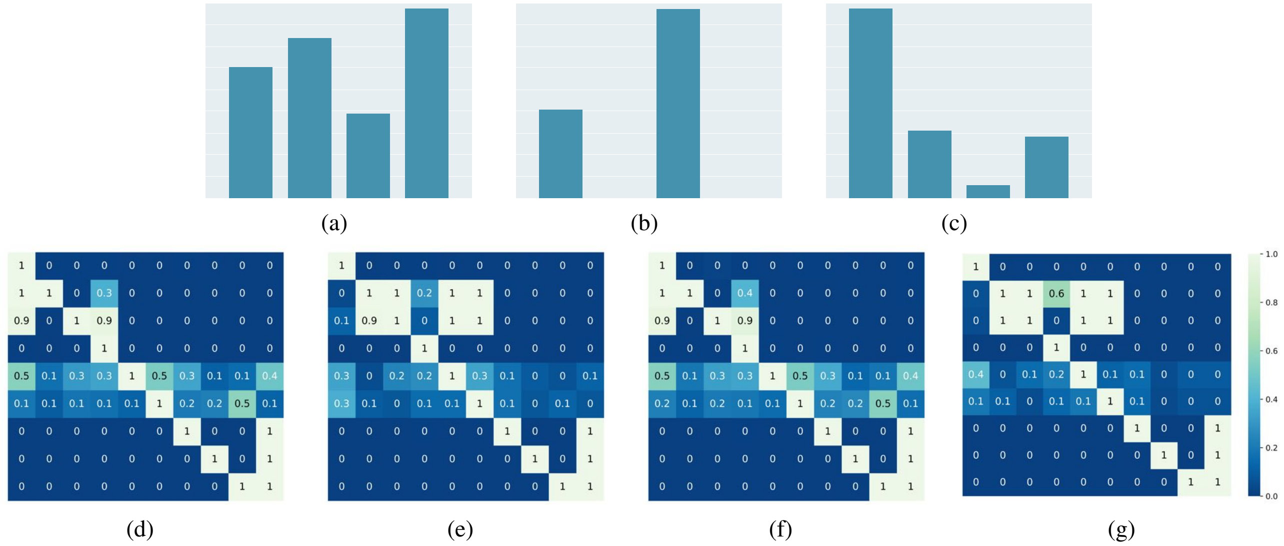
Figure 14. Analysis of LCGs learned by our method with quantization degree of 4 in Magnetic.

We observe that the training of latent codebook with vector quantization is often unstable when $K = 2$. We demonstrate the success (Fig. 17) and failure (Fig. 18) cases of our method with a quantization degree of 2. In a failure case, we observe that the embeddings frequently fluctuate between the two codes, resulting in both codes corresponding to the global causal graph and failing to capture the LCG, as shown in Fig. 18.

C.5.2. Learning curves on all downstream tasks

Fig. 19 shows the learning curves on training in all environments. Figs. 4, 20 and 21 shows the learning curves on all downstream tasks.[6]

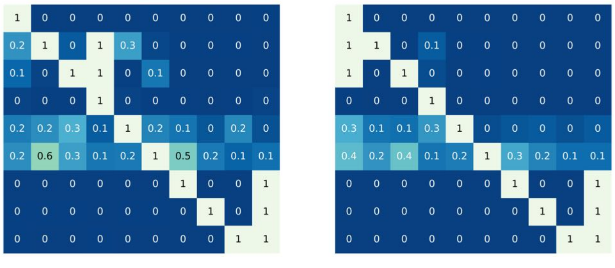
[6] As CDL is a two-stage method that requires searching the best threshold after the first stage training, we only report their final performance.



(a) (b) (c)

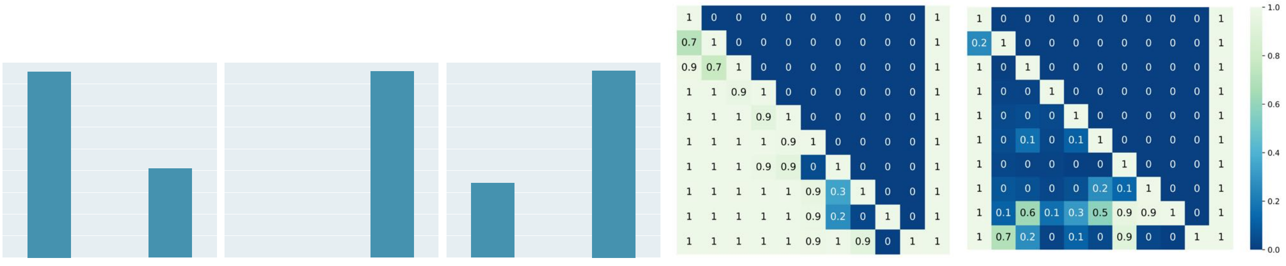
(d) (e) (f) (g)

Figure 15. Another sample run of our method with quantization degree of 4 in Magnetic.



(a) (b)

Figure 16. More fine-grained LCGs learned by our method with quantization degree of 16 in Magnetic.



(a) (b) (c) (d) (e)

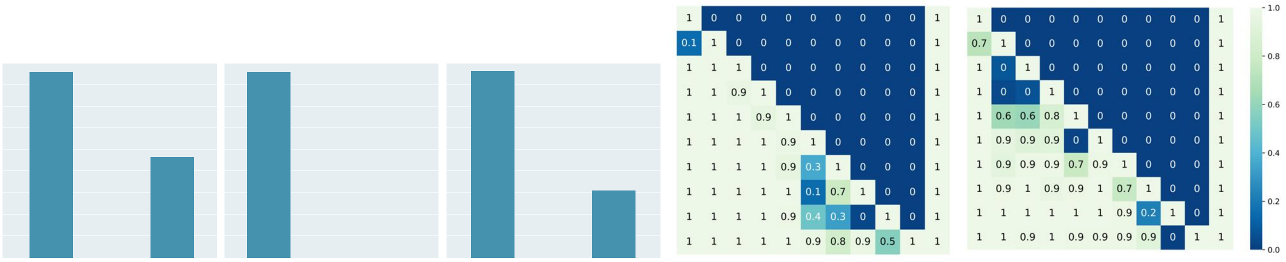
Figure 17. Analysis of LCGs learned by our method with quantization degree of 2 in Chemical (*full-fork*).

D. Additional Discussions

D.1. Difference from Sample-based Inference

Sample-based inference methods, e.g., NCD (Hwang et al., 2023) for LCG or ACD (Löwe et al., 2022) for CG, can be seen as learning causal graphs with gated edges. They learn a function that maps each sample to the adjacency matrix where each entry is the binary variable indicating whether the corresponding edge is on or off under the current state. The critical difference from ours is that LCGs learned from sample-based inference methods are *unbounded* and *blackbox*.

Specifically, it is hard to understand which local structures and contexts are identified since they can only be examined by observing the inference outcome from all samples (i.e., *blackbox*). Also, there is no (practical or theoretical) guarantee that



(a) (b) (c) (d) (e)

Figure 18. Failure case of our method with quantization degree of 2 in Chemical (*full-fork*).

it outputs the same graph from the states within the same context, since the output of the function is *unbounded*. In contrast, our method learns a finite set of LCGs where the contexts are explicitly identified by latent clustering. In other words, the outcome is bounded (infers one of the K graphs) and the contexts are more interpretable.

For the robustness of the model and principled understanding of the fine-grained structures, the practical or theoretical guarantee and interpretability are crucial, and we demonstrate the improved robustness of our method compared to prior sample-based inference methods. On the other hand, sample-based inference or local edge switch methods have strength in their simple design and efficiency, and it is known that the signals from such local edge switch enhance exploration in RL (Seitzer et al., 2021; Wang et al., 2023). For the practitioners, the choice would depend on their purpose, e.g., whether their primary interest is on the robustness and principled understanding of the fine-grained structures.

D.2. Limitations and Future Works

Insufficient or biased data may lead to inaccurate learning of causal relationships, including both CG and LCG. Our work explored the potential of utilizing LCGs to deal with (locally) spurious correlations arising from insufficient or biased data in the context of MBRL. While we assumed causal sufficiency, unobserved variables may also influence the causal relationships. These assumptions are commonly adopted in the field, yet we consider that relaxing these assumptions would be a promising future direction. Another promising future direction is to explore an inherent structure to the quantization that can efficiently handle a large number of contexts.

Table 6. Parameters of each model.

Models	Parameters	Chemical		Magnetic
		full-fork	*full-chain*	
MLP	Hidden dim	1024	1024	512
	Hidden layers	3	3	4
Modular	Hidden dim	128	128	128
	Hidden layers	4	4	4
GNN	Node attribute dim	256	256	256
	Node network hidden dim	512	512	512
	Node network hidden layers	3	3	3
	Edge attribute dim	256	256	256
	Edge network hidden dim	512	512	512
	Edge network hidden layers	3	3	3
NPS	Number of rules	20	20	15
	Cond selector dim	128	128	128
	Rule embedding dim	128	128	128
	Rule selector dim	128	128	128
	Feature encoder hidden dim	128	128	128
	Feature encoder hidden layers	2	2	2
	Rule network hidden dim	128	128	128
	Rule network hidden layers	3	3	3
CDL	Hidden dim	128	128	128
	Hidden layers	4	4	4
	CMI threshold	0.001	0.001	0.001
	CMI optimization frequency	10	10	10
	CMI evaluation frequency	10	10	10
	CMI evaluation step size	1	1	1
	CMI evaluation batch size	256	256	256
	EMA discount	0.9	0.9	0.99
Grader	Feature embedding dim	128	128	N/A
	GRU hidden dim	128	128	N/A
	Causal discovery frequency	10	10	N/A
Oracle	Hidden dim	128	128	128
	Hidden layers	4	4	5
NCD	Hidden dim	128	128	128
	Hidden layers	4	4	5
	Auxiliary network hidden dim	128	128	128
	Auxiliary network hidden layers	2	2	2
Ours	Hidden dim	128	128	128
	Hidden layers	4	4	5
	VQ encoder	[128, 64]	[128, 64]	[128, 64]
	VQ decoder	[32]	[32]	[32]
	Codebook size	16	16	16
	Code dimension	16	16	16

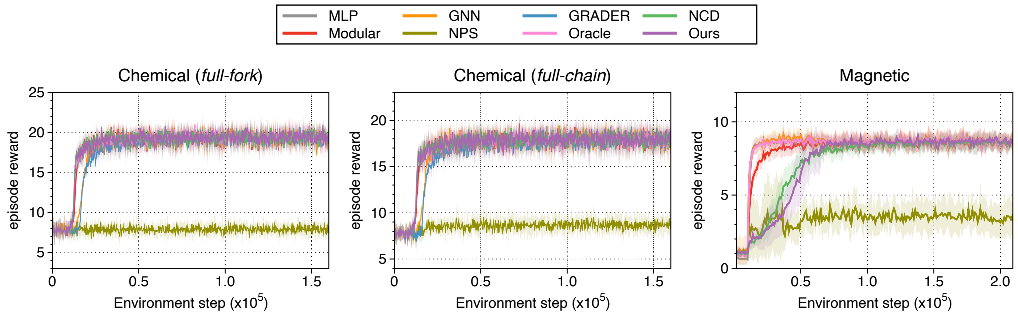


Figure 19. Learning curves during training as measured by the episode reward.

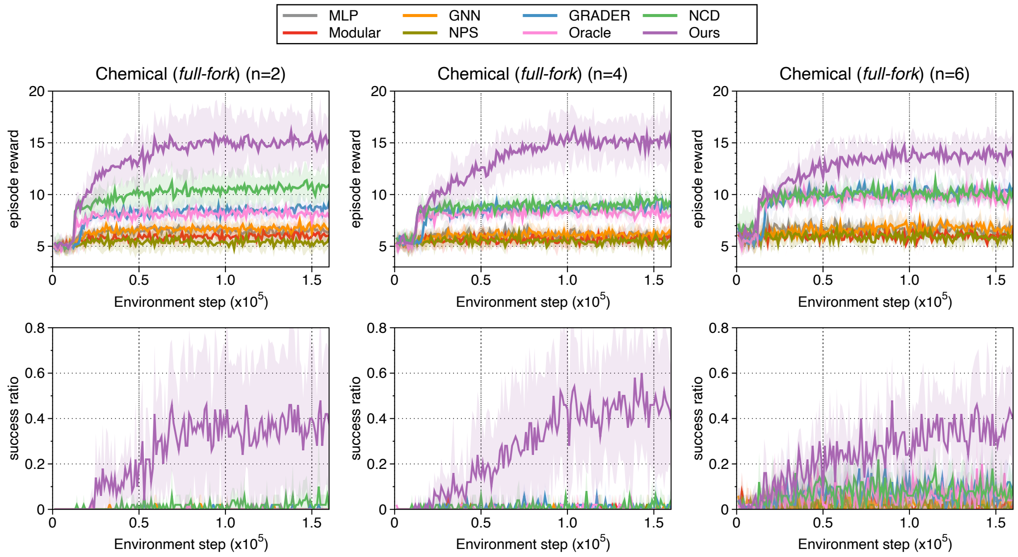


Figure 20. Learning curves on downstream tasks in Chemical (*full-fork*) as measured on the episode reward (**top**) and success rate (**bottom**).

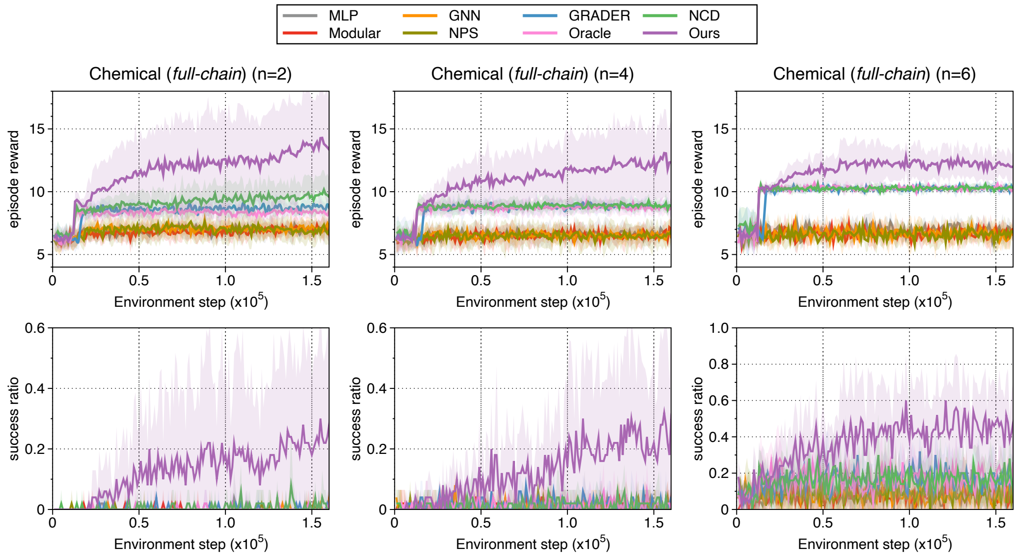


Figure 21. Learning curves on downstream tasks in Chemical (*full-chain*) as measured on the episode reward (**top**) and success rate (**bottom**).