DSAP: Enhancing Generalization in Goal-Conditioned Reinforcement Learning

Yiming Wang[1], Kaiyan Zhao[2], Ming Yang[3], Yan Li[4], Furui Liu[5], Jiayu Chen[6,7], Leong Hou U[1]*

[1]State Key Laboratory of Internet of Things for Smart City, University of Macau, Macao SAR, China
[2]School of Computer Science, Wuhan University, Wuhan, China
[3]Department of Management and Marketing, Hong Kong Polytechnic University, Hong Kong SAR, China
[4]Undergraduate School of Artificial Intelligence, Shenzhen Polytechnic University, Shenzhen, China
[5]Zhejiang Lab, Hangzhou, China
[6]Department of Data and Systems Engineering, University of Hong Kong, Hong Kong SAR, China
[7]INFIFORCE Intelligent Technology Co., Ltd, Hangzhou, China
{wang.yiming,yb57411}@connect.um.edu.mo, zhao.kaiyan@whu.edu.cn, miyang@polyu.edu.hk, liufurui@zhejianglab.com, jiayuc@hku.hk, ryanlhu@um.edu.mo

Abstract

Goal-conditioned Reinforcement Learning (RL) is a promising direction for training agents capable of tackling a variety of tasks. However, generalizing to new goals in different environments remains a central challenge for goal-conditioned RL agents. Existing methods often rely on state abstraction, which involves learning abstracted state representations by excluding irrelevant features, to improve generalization. Despite their success in simplified settings, these methods often fail to generalize effectively to realistic environments with varied goals. In this work, we propose to enhance generalization through state abstraction from the perspective of causal inference. We hypothesize that the generalization gap arises in part due to unobserved confounders: latent variables that simultaneously influence both the global and goal states. To address this, we introduce Deconfounded State Abstraction for Policy learning (DSAP), a novel framework that mitigates backdoor confounding by employing a learned causal graph as a *proxy* for the hidden confounders. We provide theoretical analysis demonstrating that DSAP improves both the learning process and the generalization capability of goal-conditioned policies. Extensive experiments across different settings of multiple benchmarks show that our method significantly outperforms existing methods.

1 Introduction

Reinforcement learning (RL) has achieved impressive success across a wide range of domains. However, real-world deployment (Yang et al. 2023) requires agents to generalize to factors outside the training distribution, known as the out-of-distribution (OOD) generalization problem (Krueger et al. 2021; Shen et al. 2021). This challenge is further compounded by the fact that the test-time distribution is typically unknown and can differ significantly from the training environment. Recent efforts have introduced causal reasoning into RL to improve generalization by capturing structured knowledge through directed graphical models (Wang et al. 2022a). For instance, some methods represent environment dynamics using learnable causal graphs (Ding et al. 2022), while others explain spurious correlations caused by unobserved confounders (Ding et al. 2024). Despite these advances, existing causal RL methods often struggle in *hard OOD scenarios* where the goal object or task semantics vary considerably (Yang et al. 2024). In this work, we propose to model the generalization process in goal-conditioned RL as a causal graph, providing a principled approach for addressing goal-dependent distribution shifts.

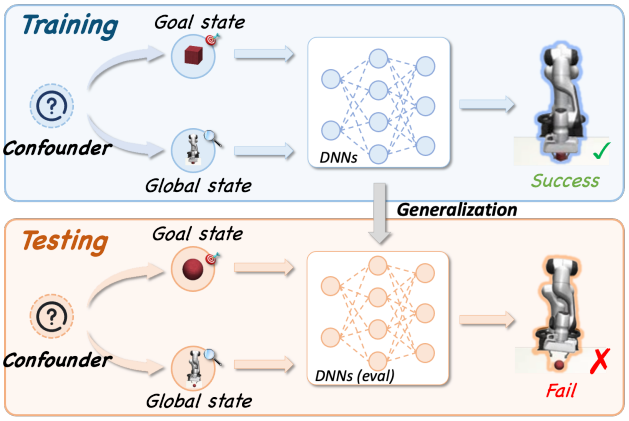


Figure 1: **Generalization Dilemma** in Goal-Conditioned RL. Unobserved confounders in the environment can simultaneously influence both the global state and the goal state, introducing confounding bias during policy learning. In the training phase (top), the agent learns to lift a cube-shaped object, with deep neural networks (DNNs) encoding both the state abstraction and the policy. However, during testing (bottom), a change in the confounder (the object's shape) modifies both the global and goal states. This distributional shift leads to a mismatch between training and testing conditions, ultimately causing policy failure.

Recent advances in deep reinforcement learning (RL)

*Corresponding author.

have leveraged state abstraction (Shanahan and Mitchell 2022; Wang et al. 2022b, 2024b) to improve transfer and generalization. Learning compact state representations that filter out task-irrelevant information is crucial for enabling agents to generalize across diverse downstream tasks in goal-conditioned RL. However, in practice, generalization often fails due to the presence of unobserved confounders: latent factors that simultaneously influence both the observed global state and the goal specification. These confounders induce confounding bias (Pearl, Glymour, and Jewell 2016), which compromises the agent's ability to generalize to novel environments. As illustrated in Figure 1, such confounding bias arises when the properties of unobserved confounders shift between training and testing. During training, deep neural networks (DNNs) learn both state abstractions and policies based on consistent environmental factors. However, at test time, changes in the confounder (e.g., object shape) alter both global and goal states, introducing a distribution mismatch that leads to policy failure.

To address the aforementioned issue, we model policy learning with state abstraction through a novel perspective of causal inference[1]. As shown in Figure 2(a), we formulate the process as a causal graph where unobserved confounders z influence both the global state s and the goal state g, inducing a backdoor path $s \leftarrow z \rightarrow g \rightarrow a$ that leads to confounding bias (Pearl, Glymour, and Jewell 2016). This bias causes learned policies to fail when generalizing to different testing goals introduced by the confounders. A direct solution is to estimate the interventional distribution $P(a \mid do(s))$ by cutting off the backdoor path (Pearl 1995), but this is infeasible due to the unobservability of z and the random sampling of goals g during testing. To overcome this, we introduce a proxy variable $\mathcal{G}$, learnable causal graph that depends solely on g, to estimate the abstracted state s^{ab} through the path $g \rightarrow \mathcal{G} \rightarrow s^{ab}$. As illustrated in Figure 2(b), the new path decomposes the confounding bias into two parts: $s \leftarrow z \rightarrow g \rightarrow \mathcal{G} \rightarrow s^{ab}$ and $s^{ab} \leftarrow \mathcal{G} \leftarrow g \rightarrow a$. In this way, the proxy confounder $\mathcal{G}$ enables backdoor adjustment by facilitating the estimation of $P(s^{ab} \mid do(s))$ and $P(a \mid do(s^{ab}))$. Built upon this mechanism, our DSAP framework integrates causal discovery into state abstraction and policy learning, enabling robust generalization across unseen goals. We further prove the unique identifiability of the underlying causal graph and establish theoretical guarantees on the generalization performance.

The main contributions of the paper are as follows:

- We introduce a novel causal perspective of the generalization caused by confounders for goal-conditioned RL, revealing how unobserved confounders induce bias through backdoor paths between global and goal states.
- We propose DSAP, a framework that leverages learnable causal graphs as proxy confounders for backdoor adjustment, and prove the unique identifiability of the true causal graph in goal-conditioned settings.
- We provide theoretical guarantees and extensive empirical results demonstrating that DSAP significantly improves generalization across diverse environments.

[1]Project details: https://github.com/YimingWangMingle/DSAP

2 Preliminaries

Goal-conditioned RL. We formulate the environments as the Goal-conditioned Markov Decision Process (MDP) setting with full observation, which is defined by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, G, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$ defines the transition dynamics, $G \subset \mathcal{S}$ is the goal state space which is a set of assignment of values to states, $r(s, g) = 1(s = g) : \mathcal{S} \times G \rightarrow \mathbb{R}$ is the one-step immediate reward function conditioned on the goal g, and γ is the discount factor. In this paper, we focus on the goal-conditioned generalization problem, where the goal for training and testing stages will be sampled from different distributions $p_{\text{train}}(g)$ and $p_{\text{test}}(g)$. Our objective is to learn a goal-conditioned policy $\pi(s, g) : \mathcal{S} \times G \rightarrow \mathcal{A}$ that maximizes the expected sum of rewards $\sum_{k=t}^{\infty} \gamma^{k-t} r(s_k, g)$. To facilitate the implementation of causal discovery, we make an assumption similar to previous work (Wang et al. 2022a; Seitzer, Schölkopf, and Martius 2021): the state space $\mathcal{S}$ can be factorized to disjoint components $\{\mathcal{S}^i\}_{i=1}^{d_{\mathcal{S}}}$, where the factorization can be achieved by abstraction methods.

Causal Graphical Models. Causal structures are commonly represented as directed acyclic graphs (DAGs) (Peters, Janzing, and Schölkopf 2017) over variables. We can model the causal relation using a Causal Graphical Model (CGM) (Peters, Janzing, and Schölkopf 2017). Consider random variables $\mathcal{X} = \{X^1 \cdots X^d\}$ with index set $\mathcal{V} = \{1, \ldots, d\}$. A directed acyclic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consists of nodes $\mathcal{V}$ and edges $\mathcal{E} \subset \mathcal{V}^2$ with (i, j) for any $i, j \in \mathcal{V}$. A node i is called a parent of j if $e^{ij} \in \mathcal{E}$ and $e^{ji} \notin \mathcal{E}$. The set of parents of j is denoted by Pa^j. We formally define causality as follows:

Definition 1 (Structural Causal Models (Peters, Janzing, and Schölkopf 2017))**.** *A structural causal model (SCM) $\mathfrak{C} := (\mathcal{X}, \mathcal{U}, \mathcal{F}, P)$. Here $\mathcal{X}$ is the set of endogenous variables, $\mathcal{U}$ is the set of endogenous (noise) variables, $\mathcal{F}$ is the collection of d structural functions $X^j := f^j(Pa^j_{\mathcal{G}}, U^j), j \in [d]$, where $Pa^j \subset \{X^1, \ldots, X^d\} \backslash \{X^j\}$ are called parents of X^j and $U^j \in \mathcal{U}$. P is the distribution of all the exogenous variables.*

Definition 2 (Backdoor Path)**.** *A backdoor path from a variable X to a variable Y in a DAG is any path that connects X to Y with an arrow pointing to Y.*

Definition 3 (Confounder (Pearl 1995))**.** *Two variables X and Y are confounded by Z if they are both caused by Z.*

Definition 4 (Do-operator (Pearl 1995))**.** *An intervention on a set of endogenous variables $X \in \mathcal{X}$ assigns a value x to X regardless of the other exogenous and endogenous variables as well as the structural functions. We denote by $do(X = x)$ the intervention on X in our work.*

Following (Ding et al. 2022), the do operation can be understood as the actions taken by an agent in RL environments. Through the do operation, the agent can intervene on state variables within the SCM, thereby influencing the state s of the environment. We denote the marginal distribution of the state variable s as $p_{do_\pi(s)}$, which can be regarded as the interventional data distribution collected by the policy π.

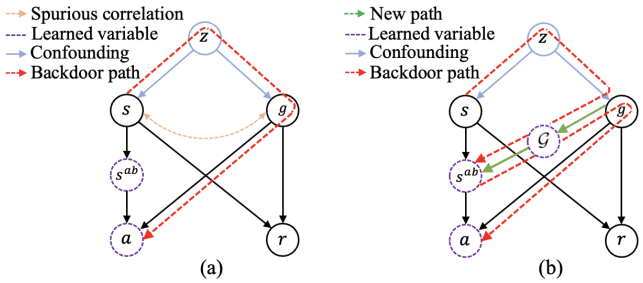


Figure 2: (a) The causal structure of state abstraction in generalizing goal-conditioned RL, where z is the unobserved confounder, r is the default reward function, s^{ab} is the learned abstracted state, a is output action of learned policy and g is the goal variable across different (training and testing) environments. There exists spurious correlation between state s and goal g when it comes to OOD generalization caused by z. (b) The causal structure of DSAP, where we introduce a new path from g to s^{ab} with the causal graph $\mathcal{G}$ of the transition dynamic as a proxy variable for achieving the backdoor adjustment.

3 Methodology

This section begins with a review of backdoor adjustment for mitigating confounding bias. We then introduce DSAP, a framework that incorporates a proxy confounder as a learnable causal graph into state abstraction and policy learning to improve generalization. Finally, we provide theoretical guarantees for its effectiveness in out-of-distribution settings.

3.1 Backdoor Adjustment

Causal Structure of Confounded Generalization. In the generalization setting of goal-conditioned RL, there exists unobserved confounders as the common cause factors between the global state and the goal state, which leads to the confounding bias when generalizing to other environments. As depicted in Figure 2(a), s and g are confounded by z, resulting in a backdoor path $s \leftarrow z \rightarrow g \rightarrow a$, which can lead to the spurious correlation between s and g during generalization when agent takes action a. Backdoor adjustment (Pearl 1995; Zhang et al. 2020b) is a widely used approach to address the confounding bias, whereby causal interventions are applied based on specified confounders to block the backdoor paths. To implement backdoor adjustment, we need to satisfy the following backdoor criterion (Peters, Janzing, and Schölkopf 2017):

Assumption 1 (Backdoor Criterion)**.** *In the SCM defined in Definition 1 and its induced directed acyclic graph (DAG), there exists an observed set $\mathcal{H}$ that satisfies the backdoor criterion, that is, the elements of $\mathcal{H}$ are not the descendants of s, and the elements of $\mathcal{H}$ d-separate every path between s and a that has an incoming arrow into s.*

Proposition 1 (Backdoor Adjustment)**.** *Under Assumption 1, it holds for all $h \in \mathcal{H}$ that:*

$$P(a \mid do(s), g) = \sum_{h} P(a \mid s, g, h)P(h) \tag{1}$$

Proof in Appendix C. Based on Proposition 1, we can achieve deconfounded policy learning for the generalization process using the backdoor adjustment with the following equation:

$$P(a \mid do(s)) = \begin{cases} \sum_z P(a \mid s, z)P(z) \\ \sum_g P(a \mid s, g)P(g) \end{cases} \tag{2}$$

However, in complex and uncontrollable environments, directly sampling the distributions of the unobserved confounder z and the goal g during training is often infeasible. This violates Assumption 1, as z and g are not contained in $\mathcal{H}$. To address this, we introduce a proxy variable as follows.

Causal Graph as Proxy Variable. To achieve the backdoor adjustment, we introduce the causal graph $\mathcal{G}$ as the proxy variable by establishing a new path $g \rightarrow \mathcal{G} \rightarrow s^{ab}$ (see Figure 2(b)). Here, $\mathcal{G}$ is the causal graph of the goal-conditioned environment transition dynamics and depends on g. ***This dependency exists because g varies across different (training and testing) environments due to changes in the properties of the goal object (confounder). Since $\mathcal{G}$ encapsulates the goal-conditioned environmental dynamics, it inherently relies on g throughout the generalization process.*** By connecting $\mathcal{G}$ and s^{ab}, we can estimate the abstracted state s^{ab} with the causal structure of the environment by removing unnecessary dependencies between actions and state variables which further benefits generalization to different tasks. As Figure 2(b) shows, the newly introduced path $g \rightarrow \mathcal{G} \rightarrow s^{ab}$ helps decompose the confounding bias on generalization process into two parts: one originates from the backdoor paths $s \leftarrow z \rightarrow g \rightarrow \mathcal{G} \rightarrow s^{ab}$, and the other arises from $s^{ab} \leftarrow \mathcal{G} \leftarrow g \rightarrow a$. By utilizing the causal graph $\mathcal{G}$ as a proxy confounder, we can cut off these two backdoor paths and achieve deconfounded policy learning. In other words, the learned causal graph $\mathcal{G}$ is observable during the training process, adhering to Assumption 1. Specifically, the aforementioned backdoor paths can be cut off by calculating $P(s^{ab} \mid do(s))$ and $P(a \mid do(s^{ab}))$ using backdoor adjustment with the following equations:

$$P(s^{ab} \mid do(s)) = \sum_{\mathcal{G}} P(s^{ab} \mid s, \mathcal{G})P(\mathcal{G}) \tag{3}$$

$$P(a \mid do(s^{ab})) = \sum_{\mathcal{G}} P(a \mid s^{ab}, \mathcal{G})P(\mathcal{G}) \tag{4}$$

The implementation details with theoretical analysis of the backdoor adjustment are discussed in the next section.

3.2 Implementation Details

Causal Graph Learning. We define the goal-conditioned causal graph of transition dynamics as a bipartite directed acyclic graph (DAG) $\mathcal{G}$, where the vertices are divided into two disjoint sets $\mathcal{T}_t = \{\mathcal{A}_t, \mathcal{S}_t\}$ and $\mathcal{T}_{t+1} = \{\mathcal{S}_{t+1}\}$. Here, $\mathcal{A}_t = \{a_t\}$ represents the action node at step t, $\mathcal{S}_t = \{s_t^1, \ldots, s_t^{d_\mathcal{S}}\}$ represents the state nodes at step t, and $d_\mathcal{S}$ represents the dimension of factored states. All edges in the graph originate from the set $\mathcal{T}_t$ and end in the set $\mathcal{T}_{t+1}$. In practice, we utilize the gradient-based causal discovery algorithm that commonly relies on least-squares (LS)

loss (Lachapelle et al. 2020) and employs neural network modeling to capture more general (non-linear) causal relationships compared with previous methods (Wang et al. 2022a; Ding et al. 2022) using Conditional Independent Test (CIT). Specifically, we seek to learn $\mathcal{G}$ using a score-based approach (Zheng et al. 2018), which defines and optimizes score functions of causal graphs to identify the underlying causal structure. Given n samples of transitions $\{s_t^1, \ldots, s_t^{d_\mathcal{S}}, a_t, s_{t+1}^1, \ldots, s_{t+1}^{d_\mathcal{S}}\}$, we denote the input feature matrix as $\mathbf{X} \in \mathbb{R}^{n \times (2d_\mathcal{S}+1)}$, and the weighted adjacency matrix of causal graph parameterized by ϕ as $\mathbf{W}(\phi)$. This optimization problem can be reformed as a SCM (see Definition 1) and uses the LS loss:

$$L(\phi) = \frac{1}{2n}\|\mathbf{X} - \mathbf{X}\mathbf{W}(\phi)\|_2 + \lambda\|\mathbf{W}(\phi)\|_1 + \alpha h(\mathbf{W}(\phi)) \tag{5}$$

where $\|\cdot\|_p$ is defined as the ℓ_p-norm on matrices, penalty term λ is used to encourage the sparsity of the matrix $\mathbf{W}(\phi)$, $h(\mathbf{W}(\phi)) = \mathrm{Tr}\left(e^{\mathbf{W}(\phi)\circ\mathbf{W}(\phi)}\right) - 2d_\mathcal{S} - 1$ is the trace exponential regularizer to enforce the "*DAG-ness*" of the causal graph and α is the scale parameter. Finally, the learned causal graph $\mathcal{G}$ is represented by a weighted adjacency matrix $\mathbf{W} = \texttt{Reduce}(\mathbf{W}(\phi)) \in \mathbb{R}^{d_\mathcal{S} \times (d_\mathcal{S}+1)}$. For more details please refer to Appendix A.

State Abstraction. In order to accomplish Equation (3), we utilize the learned causal graph $\mathcal{G}$ as input to perform state abstraction, which also allows us to filter out non-causal states that are unrelated to the action and other state variables. For instance, as depicted in Figure 3(d), the agent cannot interact with the black block. Consequently, the state of unmovable black block is causally independent of the agent's actions and other state variables which makes it an interference term during training and hinders generalization. Based on Equation (3), we implement the corresponding $P(s^{ab} \mid do(s))$ with abstraction network F: $s^{ab} = F(s_t, \mathcal{G})$. Specifically, the abstraction network F has an encoder-decoder structure with the causal graph $\mathcal{G}$ as an intermediate linear transformation. The encoder processes the input state and outputs features $f_e \in \mathbb{R}^{(d_\mathcal{S}+1)\times d_f}$, where d_f represents the feature dimension and $d_\mathcal{S}+1$ accounts for the combined dimensions of the state and action, which are then integrated with the causal graph. Then, the causal graph is multiplied to generate the feature for the decoder $f_d = \mathcal{G} * f_e \in \mathbb{R}^{d_\mathcal{S} \times d_f}$. The detailed architecture of the state abstraction can be found in Appendix B. Since we cannot iterate all the $\mathcal{G}$, we approximate the backdoor adjustment in Equation (3) via Monte Carlo sampling:

$$P(s^{ab} \mid do(s)) \approx \frac{1}{K}\sum_{k=1}^{K} P(s^{ab} \mid s, \mathcal{G}_k) \tag{6}$$

where K represents the number of sampling times (ablation study in Appendix D), and $\mathcal{G}_k$ refers to the sampled graph at times k. Thus the abstracted state is computed as:

$$s^{ab} = \frac{1}{K}\sum_{k=1}^{K} F(s, \mathcal{G}_k) \tag{7}$$

based on which the backdoor adjustment $P(s^{ab} \mid do(s))$ is completed.

Deconfounded Policy Learning. Based on Equation (4), we additionally utilize the causal graph $\mathcal{G}$ as the input fed into the policy network $\pi(\theta)$:

$$a = \pi(s^{ab}, \mathcal{G}, g) \tag{8}$$

This incorporation equips the agent with the ability of causal reasoning, as the agent itself has knowledge of the causal graph (Nair et al. 2019), which significantly benefits generalization. Similarly, combined with Equation (7), we complete the backdoor adjustment in Equation (4) via Monte Carlo sampling:

$$a = \frac{1}{M}\sum_{m=1}^{M} \pi(s_m^{ab}, \mathcal{G}_m, g_m) \tag{9}$$

where M represents the number of sampling times. Thus we achieve the deconfounded policy learning with state abstraction by calculating $P(s^{ab} \mid do(s))$ and $P(a \mid do(s^{ab}))$ with Equation (7) and Equation (9) respectively. The pseudo-code is demonstrated in Algorithm 1 in Appendix E.

3.3 Theoretical Analysis

To analyze the performance of the optimization of different blocks, we provide a theoretical analysis on the learning process. All the detailed proofs can be found in Appendix C.

Theorem 1 (Identifiability)**.** *Suppose that the joint state variables $\{s^i\}_{i=1}^{d_\mathcal{S}}$, and action a are observable at all times, the properties of Markov assumption and causal faithfulness assumption are maintained, then the causal graph is uniquely identifiable.*

Under Theorem 1, we can identify a unique $\mathcal{G}$ under optimality during the causal graph learning process. However, sampling data for the causal graph learning from a goal-conditioned MDP is not a trivial problem. Usually, the random policy is not enough to satisfy the oracle assumption because some nodes cannot be fully explored when the task is complicated and has a long horizon. To make this assumption empirically possible, it is necessary to simultaneously optimize the policy to access more samples close to finishing the task, which is further analyzed in Theorem 3.

Theorem 2 (Task Generalization)**.** *Given the mapping $F : \mathcal{S} \mapsto \mathcal{S}^{ab}$ that maps states to the abstracted state defined in Equation (7), $\mathcal{S}^{ab}$ encodes all the state information about the causal ancestors of the reward $Ancestor(R)$ after convergence of the optimal causal graph.*

With the converged optimal causal graph under Theorem 1 and drawing upon Theorem 2, we can conclude that the learned converged abstracted state demonstrates the ability to generalize to unseen reward functions, as long as the new reward function shares a subset of the same causal ancestors. Additionally, the learned representation will be robust to spurious correlations, or changes in state factors that are not in causal ancestors of the reward $Ancestor(R)$.

Theorem 3 (Distribution Alignment From Policy Learning)**.** *During the state abstraction process, let T represent the maximum step count per episode. Assuming that the Wasserstein distance between the marginal state distribution $p_{do_\pi(s)}$ in collected interventional data buffer and the abstracted state distribution $p_{s^{ab}}$ is bounded by ϵ^{ab}: $\mathcal{W}(p_{do_\pi(s)}, p_{s^{ab}}) \leq \epsilon^{ab}$. Then the Wasserstein distance between the collected interventional data distribution $p_{do_\pi(s)}$ and the goal distribution p_g is upper-bounded by the value function:*

$$\mathcal{W}(p_{do_\pi(s)}, p_g) \leq \frac{1 + \epsilon^{ab} - (1-\gamma)V^\pi(s)}{1-\gamma^T} \tag{10}$$

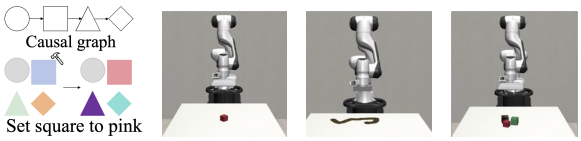
Based on Theorem 3, we conclude that during state abstraction and policy learning process, a better policy characterized by a higher value $V^\pi(s)$ aligns the distribution of interventional data collected under policy π more closely with the target goal distribution, which underscores the benefit of our policy learning for agents tackling goal-oriented tasks. See experimental results in Appendix D.

Remark. According to Theorem1, the optimal causal graph of the transition dynamics is identifiable and the causal graph learning error bound is further discussed in Lemma 3 in Appendix C. With the optimal causal graph, we can ensure the abstracted state encodes all relevant information about the causal ancestors of the reward based on Theorem 2, which demonstrate the robustness and generalization ability of learned representation. Furthermore, drawing upon Theorem 3, policy learning effectively steers the collected interventional data distribution towards the goal distribution, which in turn enhances the node exploration for causal graph learning, leading to a virtuous cycle.

4 Experiments

Our experiments evaluate the following hypotheses:
H1. DSAP achieves better performance under both in-distribution and OOD generalization settings comparing to other competitive methods.
H2. DSAP identifies the ground-truth causal graph and achieves better accuracy than Conditional Independent Test (CIT) based methods.
H3. State abstractions derived by DSAP improve sample efficiency when generalizing to downstream tasks compared to full state space and other baselines.
H4. DSAP exhibits greater scalability and robustness when compared to other goal-conditioned RL methods.

4.1 Environments and Baselines



(a) Chemistry (b) Lift (c) Wipe (d) Stack

Figure 3: The evaluation environments.

We examine the generalization on modified commonly used RL benchmarks following previous work (Wang et al. 2022a; Ke et al. 2021). We design two settings: In-Distribution (ID) and Out-Of-Distribution (OOD) corresponding to different goal distributions for generalization. We use $p_{\text{train}}(g)$ and $p_{\text{test}}(g)$ to represent the goal distribution during training and testing, respectively. ID uses the same $p_{\text{train}}(g)$ and $p_{\text{test}}(g)$. OOD introduces the difference between $p_{\text{train}}(g)$ and $p_{\text{test}}(g)$ caused by confounder. The settings are briefly summarized in the following:

We evaluate generalization on modified versions of standard RL benchmarks, following prior work (Wang et al. 2022a; Ke et al. 2021). Two settings are considered: In-Distribution (ID) and Out-of-Distribution (OOD), corresponding to different goal distributions. Let $p_{\text{train}}(g)$ and $p_{\text{test}}(g)$ denote the goal distributions during training and testing, respectively. In the ID setting, these distributions are identical. In the OOD setting, a shift is introduced through changes induced by confounders.

(1) In the Chemistry environment adapted from (Ke et al. 2021), two fixed-color nodes are added as distractors, and an underlying causal graph controls the color-changing mechanism of remaining five nodes. In one step, the agent can change the color of one node. The goal is to match given colors of the nodes. In the OOD setting, we change the color of different nodes to introduce the spurious relationship.

(2) We use the robosuite manipulation suite (Zhu et al. 2020) to assess DSAP under realistic dynamics. We evaluate three tasks: lift, wipe, and stack. In the OOD variant of the lift task, the block's shape is changed to a sphere. For the wipe task, we increase table friction to make wiping more difficult. In the stack task, an immovable black block is added as a distractor, and we change the shape and color of the two goal blocks to construct a composite OOD setting. Full implementation details are provided in Appendix E.

Baselines. To demonstrate the effectiveness of DSAP, we compare it against following competitive baseline methods. **SAC:** (Haarnoja et al. 2018) Soft Actor-Critic is a famous RL baseline which uses entropy to increase the diversity of action. **ICIN:** (Nair et al. 2019) It endows the agent with the capability of causal reasoning for completing goal-directed tasks where it assumes that the true causal graph is accessed. **ICIL:** (Bica, Jarrett, and van der Schaar 2021) The method proposes an invariant feature learning structure that captures the implicit causality of multiple tasks to benefit generalizing to different tasks. **DBC:** (Zhang et al. 2021) The method proposes learning the abstracted state using the bisimulation metric to group similar states that considers effective downstream control. **GRADER:** (Ding et al. 2022) It formulates the goal-conditioned RL problem into variational likelihood maximization with causal graphs as latent variables to benefit generalization. **CDL:** (Wang et al. 2022a) The model-based method learns the causal dynamics model that strictly removes dependencies between state variables and the action to help improve generalization. **CBM:** (Wang et al. 2024b) The method learns the causal relationships in the dynamics and reward functions for each task to derive a minimal and task-specific abstraction.

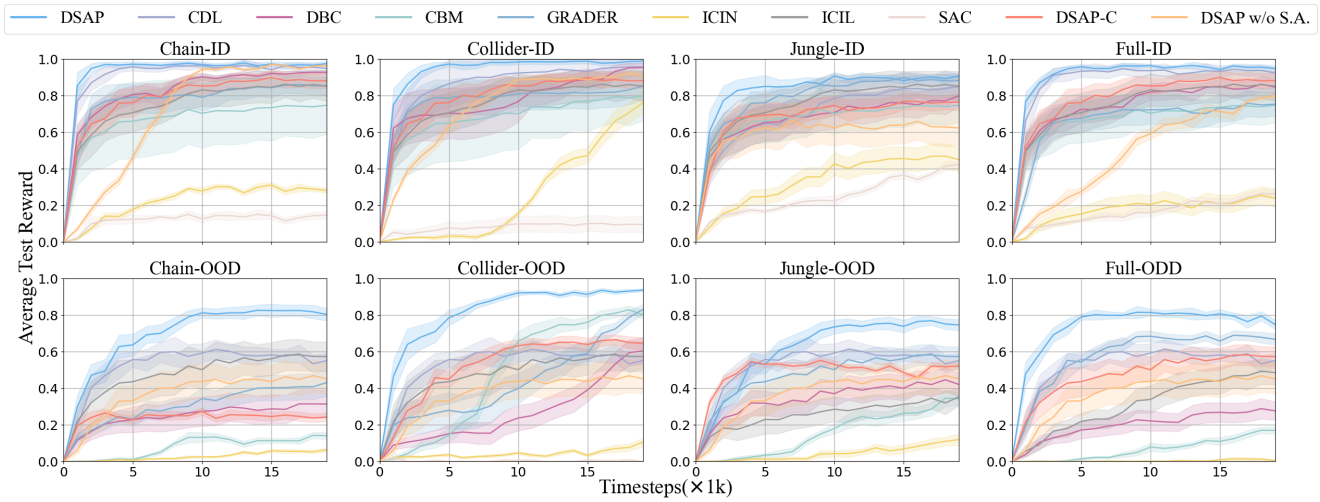


Figure 4: Comparisons of DSAP and baselines (all results are the average test reward with mean and standard deviation over 10 random seeds) in three types of causal graphs [Chain, Collider, Jungle, Full] under ID and ODD setting of Chemistry.

4.2 Main Results

Overall Performance (H1). As shown in Figure 4 and Table 2, DSAP dominates all baselines in the chemistry environment (ID and OOD setting) and achieves the best performance in all the OOD setting of the manipulation tasks. Note that the gap between our method and baselines in OOD setting of the chemistry environment is more significant than in the ID setting, showing that our method still works well in the non-trivial generalization task. SAC struggles in all tasks of chemistry and manipulation environments since they have very sparse rewards for goal-conditioned tasks. Without learning the causal structure of the environment, DBC fails in most OOD generalization tasks. The ability to uncover causality behind the tasks provides ICIN, ICIL, and GRADER with an edge over SAC, CBM and DBC, leading to superior performance. CDL's effective state abstraction leads to strong performance, but vulnerability to spurious correlations introduced by confounders in OOD settings hinders its performance compared to DSAP.

Environment	DSAP	GRADER	CDL
Chain	**100.0** $\pm$ **0.0**	99.5 $\pm$ 0.2	**100.0** $\pm$ **0.0**
Collider	**100.0** $\pm$ **0.0**	99.7 $\pm$ 0.1	**100.0** $\pm$ **0.0**
Jungle	**99.5** $\pm$ **0.2**	98.1 $\pm$ 0.1	99.1 $\pm$ 0.2
Full	**99.3** $\pm$ **0.3**	97.7 $\pm$ 0.5	99.1 $\pm$ 0.1
Manipulation	**93.7** $\pm$ **0.1**	82.3 $\pm$ 0.4	89.9 $\pm$ 0.2

Table 1: Causal graph accuracy (%) comparison

Causal Discovery Analysis (H2). To verify the effectiveness of the causal graph learning in DSAP, we evaluate the performance of causal discovery with Conditional Independent Test (CIT) used in GRADER and CDL. We assess the performance of these methods in chemistry and manipulation environments in terms of accuracy, recall, precision, F1 score and ROC AUC. For all metrics, the higher the better. We present the results of causal graph accuracy in Table 1 (Full results in Appendix D). We can observe that our gradient-based causal graph trainer achieves improved performance, particularly in environments with complex causal graphs. This advantage stems from NNs' ability to capture non-linear relationships unseen by CIT-based methods. We visualize the learned causal graph and ground-truth graph of *Collider* as depicted in Figure 5. Our causal discovery process reveals the accurate causal graph behind the environment.

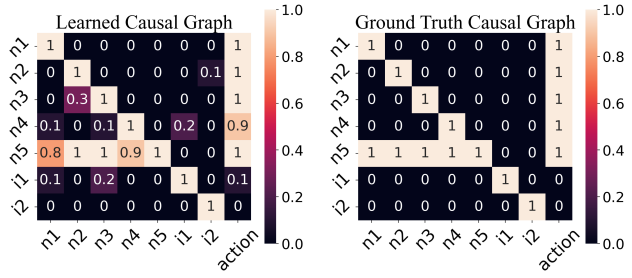


Figure 5: Learned and ground-truth causal graph.

Counterfactual analysis is widely adopted in explainable RL (Madumal et al. 2020b,a) to interpret agent behavior. To evaluate the *interpretability* of DSAP, we design a counterfactual variant, referred to as **DSAP-C**, which uses a predefined counterfactual causal graph during training (see Appendix E for details). As shown in Figure 4 and Table 2, this counterfactual setting leads to a notable drop in DSAP's performance. Intuitively, the use of a counterfactual causal graph introduces redundant and spurious correlations among state variables, injecting additional noise into the learning process and degrading generalization.

Advantage of State Abstraction (H3). To validate the efficacy of the state abstraction process, we compare DSAP with and without state abstraction (denoted as "**DSAP w/o**

Method	Lift-OOD	Wipe-OOD	Stack-OOD	Reach	Pick	Match
SAC	23.8%±11.4%	21.5%±12.3%	25.7%±12.6%	46.9%±12.5%	25.6%±10.7%	13.4%±6.7%
ICIN	90.1%±7.4%	78.4%±13.5%	61.5%±13.0%	72.9%±13.5%	68.6%±16.7%	52.7%±14.5%
ICIL	75.6%±14.7%	64.5%±12.4%	57.6%±14.5%	77.7%±15.9%	71.9%±17.6%	58.4%±17.7%
CBM	81.5%±10.8%	69.9%±10.1%	65.8%±12.3%	88.6%±12.6%	87.6%±11.8%	70.4%±11.5%
DBC	61.2%±11.3%	51.7%±11.2%	42.8%±12.5%	85.9%±11.3%	80.1%±12.4%	69.8%±13.5%
CDL	77.6%±12.9%	62.7%±10.0%	59.9%±12.4%	**94.8%±4.4%**	81.3%±7.9%	71.6%±6.8%
GRADER	88.9%±13.2%	81.6%±9.1%	86.9%±8.8%	83.6%±9.6%	65.5%±11.2%	51.6%±12.3%
DSAP	**94.6%±8.6%**	**96.9%±9.3%**	**89.9%±11.3%**	92.8%±6.7%	**91.8%±6.9%**	**85.4%±7.6%**
DSAP-counterfact	77.4%±9.5%	75.6%±11.2%	81.3%±4.6%	74.5%±8.4%	64.7%±13.4%	59.5%±11.6%
DSAP w/o S.A.	69.6%±11.3%	71.8%±11.2%	58.3%±13.6%	71.9%±12.4%	61.8%±12.1%	50.9%±10.6%

Table 2: Mean success rate (average over 10 random seeds) for the OOD testing settings of three manipulation tasks (Lift, Wipe, Stack) and downstream tasks (Reach, Pick, Match) in the manipulation environment.

S.A.") against three competitive abstraction methods: DBC, CBM and CDL. Results, as presented in Figure 4 and Table 2, consistently demonstrate DSAP's superior performance in both ID and OOD settings. DBC's lack of causality integration results in lower performance compared to DSAP and CDL. CBM struggles in OOD settings due to the existence of confounding bias during the generalization process; As for CDL, the strict removal of state variables based on the causal dynamic transition model, while effective in certain scenarios, can lead to loss of critical information for generalization and vulnerability to spurious relations in OOD settings. Also the performance of **DSAP w/o S.A.** exhibits a huge decline which highlights benefits of state abstraction.

The Scalability of DSAP (H4). To validate the DSAP's scalability compared to baselines, we evaluate on the chemistry environments with increasing entity nodes (7, 12, 17). As the node count increases, the complexity of the causal graph also increases, making it more challenging for the agent to achieve the goal. The results can be found in Appendix D. Also we conduct experiments on downstream tasks on the manipulation environment. As Table 2 shows, DSAP still achieves the best performance when handling increasingly complex environments and different downstream tasks, indicating the robustness and scalability of deconfounded policy learning with state abstraction.

5 Related Work

RL Generalization. Generalization is a core challenge in GCRL (Liu, Zhu, and Zhang 2022), which aims to train agents across diverse tasks. Common approaches include experience replay (Zhao et al. 2024), reward shaping (Wang et al. 2023, 2024a, 2025), and latent dynamics models (Nair, Savarese, and Finn 2020). Other works reformulate GCRL as variational inference (Tang and Kucukelbir 2021; Rudner et al. 2021), or apply causal reasoning to enhance generalization (Ding et al. 2022). Beyond optimization, task decomposition (Huang et al. 2019; Kipf et al. 2019), symbolic reasoning (Yang et al. 2021; Landajuela et al. 2021), and image-based data augmentation (Hansen, Su, and Wang 2021) have also been explored. GCRL generalization has further been studied under domain adaptation (Tobin et al. 2017; Mehta et al. 2020), meta-learning (Sæmundsson, Hofmann, and Deisenroth 2018; Finn, Abbeel, and Levine 2017), multi-task learning (Schaul et al. 2015), and curriculum learning (Narvekar et al. 2020).

State Abstraction. State abstraction improves sample efficiency by mapping multiple states to a compact representation that omits task-irrelevant variables. Bisimulation (Even-Dar and Mansour 2003) groups states with equivalent future reward distributions under any action sequence. Its connection to causality has been highlighted by showing that bisimulation preserves only the causal ancestors of the reward (Zhang et al. 2020a). Recent methods such as DeepMDP (Gelada et al. 2019) and DBC (Zhang et al. 2021) learn approximate bisimulation via end-to-end training with abstracted dynamics and rewards. Model-based approaches (Wang et al. 2022a, 2024b) further improve generalization by removing spurious dependencies in dynamics.

Causal RL. Causality has been integrated into reinforcement learning across both model-based and model-free settings (Lattimore, Lattimore, and Reid 2016; Madumal et al. 2020c). (Lu, Schölkopf, and Hernández-Lobato 2018) estimate structural causal models (SCMs) from observational data with confounders between actions and rewards. (Gasse et al. 2021) leverage both observational and interventional data to learn latent transition models for deconfounding. In model-free settings, (Wang, Yang, and Wang 2021) improve sample efficiency using observational data, while recent works extend deconfounding techniques to multi-agent (Li et al. 2022) and history-based RL (Gao et al. 2023).

6 Conclusion and Future Work

In this work, we propose deconfounded state abstraction for policy learning (DSAP) to use the learned causal graph of dynamics as the proxy confounder for backdoor adjustment, which removes the confounding bias between global and goal states during generalization. Further theoretical analysis guarantees the improvement on learning and generalization process and extensive experiments have demonstrated the superiority of our algorithm. The factorized state assumption is the key limitation. Integrating semantic-level state segmentation methods (e.g. VLMs), presents a promising direction for enhancing scalability in the future.

Acknowledgments

This work was supported by National Science and Technology Major Project (2023ZD0121401), the Science and Technology Development Fund Macau SAR (0003/2023/RIC, 0052/2023/RIA1, 0031/2022/A, 001/2024/SKL for SKL-IOTSC) and Shenzhen-Hong Kong-Macau Science and Technology Program Category C (SGDX20230821095159012), NSF of China 62402325 and the Research Foundation of Shenzhen Polytechnic University under Grant 6022310014K. This work was performed in part at SICC which is supported by SKL-IOTSC, University of Macau.

References

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

Ding, W.; Lin, H.; Li, B.; and Zhao, D. 2022. Generalizing goal-conditioned reinforcement learning with variational causal reasoning. *Advances in Neural Information Processing Systems*, 35: 26532–26548.

Ding, W.; Shi, L.; Chi, Y.; and Zhao, D. 2024. Seeing is not believing: Robust reinforcement learning against spurious correlation. *Advances in Neural Information Processing Systems*, 36.

Even-Dar, E.; and Mansour, Y. 2003. Approximate equivalence of Markov decision processes. In *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. Proceedings*, 581–594. Springer.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, 1126–1135. PMLR.

Gao, H.; Zhang, T.; Yang, Z.; Guo, Y.; Ren, J.; Guo, S.; and Chen, F. 2023. Fast counterfactual inference for history-based reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 7613–7623.

Gasse, M.; Grasset, D.; Gaudron, G.; and Oudeyer, P.-Y. 2021. Causal reinforcement learning using observational and interventional data. *arXiv preprint arXiv:2106.14421*.

Gelada, C.; Kumar, S.; Buckman, J.; Nachum, O.; and Bellemare, M. G. 2019. Deepmdp: Learning continuous latent space models for representation learning. In *International Conference on Machine Learning*, 2170–2179. PMLR.

Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, 1861–1870. PMLR.

Hansen, N.; Su, H.; and Wang, X. 2021. Stabilizing deep q-learning with convnets and vision transformers under data augmentation. *Advances in neural information processing systems*, 34: 3680–3693.

Huang, D.-A.; Nair, S.; Xu, D.; Zhu, Y.; Garg, A.; Fei-Fei, L.; Savarese, S.; and Niebles, J. C. 2019. Neural task graphs: Generalizing to unseen tasks from a single video demonstration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8565–8574.

Ke, N. R.; Didolkar, A.; Mittal, S.; Goyal, A.; Lajoie, G.; Bauer, S.; Rezende, D.; Bengio, Y.; Mozer, M.; and Pal, C. 2021. Systematic evaluation of causal discovery in visual model based reinforcement learning. *arXiv preprint arXiv:2107.00848*.

Kipf, T.; Li, Y.; Dai, H.; Zambaldi, V.; Sanchez-Gonzalez, A.; Grefenstette, E.; Kohli, P.; and Battaglia, P. 2019. Compile: Compositional imitation learning and execution. In *International Conference on Machine Learning*, 3418–3428. PMLR.

Krueger, D.; Caballero, E.; Jacobsen, J.-H.; Zhang, A.; Binas, J.; Zhang, D.; Le Priol, R.; and Courville, A. 2021. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, 5815–5826. PMLR.

Lachapelle, S.; Brouillard, P.; Deleu, T.; and Lacoste-Julien, S. 2020. Gradient-Based Neural DAG Learning. In *ICLR*. OpenReview.net.

Landajuela, M.; Petersen, B. K.; Kim, S.; Santiago, C. P.; Glatt, R.; Mundhenk, N.; Pettit, J. F.; and Faissol, D. 2021. Discovering symbolic policies with deep reinforcement learning. In *International Conference on Machine Learning*, 5979–5989. PMLR.

Lattimore, F.; Lattimore, T.; and Reid, M. D. 2016. Causal bandits: Learning good interventions via causal inference. *Advances in neural information processing systems*, 29.

Li, J.; Kuang, K.; Wang, B.; Liu, F.; Chen, L.; Fan, C.; Wu, F.; and Xiao, J. 2022. Deconfounded value decomposition for multi-agent reinforcement learning. In *International Conference on Machine Learning*, 12843–12856. PMLR.

Liu, M.; Zhu, M.; and Zhang, W. 2022. Goal-conditioned reinforcement learning: Problems and solutions. *arXiv preprint arXiv:2201.08299*.

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

Madumal, P.; Miller, T.; Sonenberg, L.; and Vetere, F. 2020a. Distal explanations for explainable reinforcement learning agents. *arXiv preprint arXiv:2001.10284*.

Madumal, P.; Miller, T.; Sonenberg, L.; and Vetere, F. 2020b. Explainable reinforcement learning through a causal lens. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 2493–2500.

Madumal, P.; Miller, T.; Sonenberg, L.; and Vetere, F. 2020c. Explainable reinforcement learning through a causal lens. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 2493–2500.

Mehta, B.; Diaz, M.; Golemo, F.; Pal, C. J.; and Paull, L. 2020. Active domain randomization. In *Conference on Robot Learning*, 1162–1176. PMLR.

Nair, S.; Savarese, S.; and Finn, C. 2020. Goal-aware prediction: Learning to model what matters. In *International Conference on Machine Learning*, 7207–7219. PMLR.

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

Narvekar, S.; Peng, B.; Leonetti, M.; Sinapov, J.; Taylor, M. E.; and Stone, P. 2020. Curriculum learning for reinforcement learning domains: A framework and survey. *The Journal of Machine Learning Research*, 21(1): 7382–7431.

Pearl, J. 1995. Causal diagrams for empirical research. *Biometrika*, 82(4): 669–688.

Pearl, J.; Glymour, M.; and Jewell, N. P. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

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

Rudner, T. G.; Pong, V.; McAllister, R.; Gal, Y.; and Levine, S. 2021. Outcome-driven reinforcement learning via variational inference. *Advances in Neural Information Processing Systems*, 34: 13045–13058.

Sæmundsson, S.; Hofmann, K.; and Deisenroth, M. P. 2018. Meta reinforcement learning with latent variable gaussian processes. *arXiv preprint arXiv:1803.07551*.

Schaul, T.; Horgan, D.; Gregor, K.; and Silver, D. 2015. Universal value function approximators. In *International conference on machine learning*, 1312–1320. PMLR.

Seitzer, M.; Schölkopf, B.; and Martius, G. 2021. Causal influence detection for improving efficiency in reinforcement learning. *Advances in Neural Information Processing Systems*, 34: 22905–22918.

Shanahan, M.; and Mitchell, M. 2022. Abstraction for deep reinforcement learning. *arXiv preprint arXiv:2202.05839*.

Shen, Z.; Liu, J.; He, Y.; Zhang, X.; Xu, R.; Yu, H.; and Cui, P. 2021. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*.

Tang, Y.; and Kucukelbir, A. 2021. Hindsight expectation maximization for goal-conditioned reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, 2863–2871. PMLR.

Tobin, J.; Fong, R.; Ray, A.; Schneider, J.; Zaremba, W.; and Abbeel, P. 2017. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, 23–30. IEEE.

Wang, L.; Yang, Z.; and Wang, Z. 2021. Provably efficient causal reinforcement learning with confounded observational data. *Advances in Neural Information Processing Systems*, 34: 21164–21175.

Wang, Y.; Yang, M.; Dong, R.; Sun, B.; Liu, F.; et al. 2023. Efficient Potential-based Exploration in Reinforcement Learning using Inverse Dynamic Bisimulation Metric. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Wang, Y.; Zhao, K.; Li, Y.; and U, L. H. 2025. BILE: an effective behavior-based latent exploration scheme for deep reinforcement learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 6497–6505.

Wang, Y.; Zhao, K.; Liu, F.; et al. 2024a. Rethinking exploration in reinforcement learning with effective metric-based exploration bonus. *Advances in Neural Information Processing Systems*, 37: 57765–57792.

Wang, Z.; Wang, C.; Xiao, X.; Zhu, Y.; and Stone, P. 2024b. Building minimal and reusable causal state abstractions for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15778–15786.

Wang, Z.; Xiao, X.; Xu, Z.; Zhu, Y.; and Stone, P. 2022a. Causal Dynamics Learning for Task-Independent State Abstraction. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, 23151–23180. PMLR.

Wang, Z.; Xiao, X.; Xu, Z.; Zhu, Y.; and Stone, P. 2022b. Causal dynamics learning for task-independent state abstraction. *arXiv preprint arXiv:2206.13452*.

Yang, M.; Wang, Y.; Yu, Y.; Zhou, M.; and U, L. H. 2023. MixLight: Mixed-Agent Cooperative Reinforcement Learning for Traffic Light Control. *IEEE Transactions on Industrial Informatics*, 1–9.

Yang, M.; Zhao, K.; Wang, Y.; Dong, R.; Du, Y.; Liu, F.; Zhou, M.; and U, L. H. 2024. Team-wise effective communication in multi-agent reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 38(2): 36.

Yang, Y.; Inala, J. P.; Bastani, O.; Pu, Y.; Solar-Lezama, A.; and Rinard, M. 2021. Program synthesis guided reinforcement learning for partially observed environments. *Advances in neural information processing systems*, 34: 29669–29683.

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

Zhang, A.; McAllister, R. T.; Calandra, R.; Gal, Y.; and Levine, S. 2021. Learning Invariant Representations for Reinforcement Learning without Reconstruction. In *ICLR*. OpenReview.net.

Zhang, S.; Jiang, T.; Wang, T.; Kuang, K.; Zhao, Z.; Zhu, J.; Yu, J.; Yang, H.; and Wu, F. 2020b. Devlbert: Learning deconfounded visio-linguistic representations. In *Proceedings of the 28th ACM International Conference on Multimedia*, 4373–4382.

Zhao, K.; Wang, Y.; Chen, Y.; Li, Y.; Niu, X.; et al. 2024. Efficient Diversity-based Experience Replay for Deep Reinforcement Learning. *arXiv preprint arXiv:2410.20487*.

Zheng, X.; Aragam, B.; Ravikumar, P. K.; and Xing, E. P. 2018. Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31.

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