# DCIL: DropConnect-based Causal Imitation Learning Under Environment Heterogeneity

**Xinyu Zhang**[1] , **Shihua Li**[1] , **Rongjie Liu**[2]
[1]Southeast University, Nanjing, Jiangsu, China
[2]University of Georgia, Athens, GA, USA
{zhang_xinyu, lsh}@seu.edu.cn, rjliu@uga.edu

## Abstract

In many real-world scenarios, agents operate across heterogeneous environments where the underlying dynamics and data-generating processes vary. Standard reinforcement learning and imitation learning methods often fail in such settings, as they typically assume stationarity and learn policies that overfit to environment-specific correlations. A key challenge is the presence of spurious correlations as observed states often contain both causal and non-causal features, with the latter introducing environment-specific biases that undermine generalization. To address this problem, we propose DropConnect-based Causal Imitation Learning (DCIL), a novel offline imitation learning framework designed to identify and exploit stable causal mechanisms across diverse environments. DCIL introduces a gradient alignment constraint that encourages the policy to align with stable causal features shared across training environments. To further mitigate overfitting to spurious correlations, DCIL involves DropConnect-based regularization, injecting stochastic perturbations into network weights to simulate parameter uncertainty and reduce reliance on unstable features. We evaluate DCIL on synthetic benchmarks derived from OpenAI Gym control tasks, where non-causal features exhibiting spurious correlations are explicitly injected to simulate environmental heterogeneity. Experimental results show that DCIL outperforms state-of-the-art imitation learning baselines, achieving superior generalization to unseen environments. These findings highlight the importance of incorporating causal invariance principles into policy learning for robust performance under environment shift.



## 1 Introduction

Reinforcement Learning (RL) has achieved impressive success in a wide range of decision-making tasks, including game playing, robotics, and resource management [Arulkumaran *et al.*, 2017; Mnih *et al.*, 2015]. However, many classical RL algorithms [Lillicrap *et al.*, 2015; Hasselt *et al.*, 2016] assume access to a well-specified and fully observable reward function, which is often impractical to define in real-world applications [Hadfield-Menell *et al.*, 2017]. Imitation Learning (IL) offers a promising alternative by learning policies directly from expert demonstrations, thereby bypassing the need for explicit reward specification. Compared to exploration-based RL, IL is often more sample-efficient and safer, especially in safety-critical domains such as autonomous driving, healthcare, and robot manipulation [Jang *et al.*, 2022; Codevilla *et al.*, 2019; Hu *et al.*, 2022].

Despite these advantages, most IL methods assume that expert demonstrations are drawn from a single stationary environment. In practice, however, expert trajectories often originate from multiple heterogeneous environments, each with different dynamics, feature distributions, or contextual factors. As a result, observed states typically contain both causal features that genuinely influence expert decisions and non-causal features that are spuriously correlated with actions due to environment-specific artifacts [Codevilla *et al.*, 2019]. Policies trained on such data may overfit to environment-dependent correlations, failing to capture the underlying decision-making mechanisms and generalizing poorly to unseen environments. This fundamental mismatch between environment-specific correlations and environment-invariant decision mechanisms poses a central challenge for offline imitation learning under distribution shift.

Such issues are widespread in real-world settings. For example, in autonomous driving, an agent trained on data collected during rainy weather may observe a strong correlation between windshield wiper usage and reduced speed. However, wiper activity does not causally influence driving policy; it simply correlates with weather conditions, which are the actual causal factor. A policy that mistakes this correlation for causation may make unsafe decisions under different conditions.

These challenges highlight the need for imitation learning algorithms that generalize across environments by identifying and leveraging stable causal mechanisms underlying expert behavior. To this end, we propose **DCIL** (*DropConnect-based Causal Imitation Learning*), a novel framework for offline imitation learning from demonstrations collected in heterogeneous environments[1]. DCIL is designed to learn

[1]Code is available at https://github.com/PapriKy/DCIL.

policies that remain robust under distribution shifts without requiring additional online interaction. DCIL incorporates two key components. First, it introduces a gradient alignment constraint that encourages the policy to focus on stable causal features shared across environments. Second, to further reduce sensitivity to spurious features, it applies a DropConnect-based regularization strategy that injects stochastic perturbations into the policy network weights during training. This improves robustness by modeling parameter uncertainty and discouraging reliance on unstable patterns. We validate DCIL on modified OpenAI Gym control tasks where non-causal features are explicitly injected to simulate environmental heterogeneity. Experimental results show that DCIL consistently outperforms state-of-the-art imitation learning methods in generalization to unseen environments. Our contributions are summarized as follows:

- Address the problem of offline imitation learning from expert trajectories collected across heterogeneous environments, focusing on generalization without additional interaction.
- Propose a novel framework that integrates causal gradient alignment constraints to promote learning of stable, transferable policies.
- Introduce a simple yet effective DropConnect-based regularization strategy to alleviate overfitting induced by conventional causal constraints, thereby improving robustness across heterogeneous environments.

## 2 Related Work

Causal inference [Pearl, 2009; Peters *et al.*, 2016] has emerged as a promising direction for improving generalization in decision-making systems, including imitation learning and reinforcement learning. By identifying and leveraging variables that are causally related to decisions, these approaches aim to produce models that remain robust under distributional shifts. De Haan et al. [de Haan *et al.*, 2019] highlight the issue of causal misidentification, where policies inadvertently rely on spurious correlations in expert data rather than true causal variables. This issue becomes more severe when distribution shifts occur during deployment, as non-causal correlations observed during training may no longer hold. To address this problem, their approach uses causal graph-structured policies and targeted interventions, such as expert queries and environment interactions, to infer the correct causal structure. However, such strategies are not applicable in fully offline settings where no additional data collection is permitted. To overcome the need for interventions, Sonar et al. [Sonar *et al.*, 2021] incorporate the principle of Invariant Risk Minimization (IRM) [Arjovsky *et al.*, 2019] into policy learning. This framework encourages policies that perform consistently across multiple environments by aligning invariant components. While effective in certain settings, IRM-based methods can still suffer from spurious correlations even when the regularization penalty approaches zero, potentially leading to overfitting [Lin *et al.*, 2022]. Invariant Causal Imitation Learning (ICIL) [Bica *et al.*, 2021] separates causal and non-causal representations and enforces invariance via adversarial training. However, the combination of mutual information maximization and adversarial objectives often leads to unstable optimization, motivating simpler regularization-based alternatives such as ours. In contrast to these approaches, DCIL avoids explicit adversarial training or causal graph specification, and instead enforces causal invariance through gradient-based constraints combined with stochastic regularization.

## 3 Methodology

To overcome the limitations of conventional imitation learning, we propose a novel approach that introduces stable causal constraints to disentangle true causal features from environment-specific noise. Our method constrains the gradients of input-layer weights and incorporates DropConnect-based regularizations, jointly capturing stable causal influences across heterogeneous environments while mitigating overfitting to spurious correlations. This mechanism not only reduces reliance on unstable features but also enhances robustness, particularly in settings with limited environment diversity. Unlike prior work, our approach jointly applies a causal constraint that regularizes policy gradients toward stable causal features across environments, and a stochastic, weight-level regularization to mitigate overfitting, resulting in improved generalization in offline imitation learning. An overview of our method is presented in Figure 1.

### 3.1 Preliminaries

In this section, we formalize the problem of imitation learning across multiple heterogeneous environments using the framework of multi-environment Markov Decision Processes (MDPs). Without loss of generality, we assume that we observe data consisting solely of expert demonstrations collected from a set of $K$ heterogeneous environments. For the environment $e \in \mathcal{E}$, where $\mathcal{E} = \{1, \dots, K\}$ and $|\mathcal{E}| = K$, we define the MDP framework as $\mathcal{M}_e = \{\mathcal{S}^e, \mathcal{A}, \mathcal{P}^e, p_0^e, r^e, \gamma\}$, where $\mathcal{S}^e$ denotes the state space specific to the environment $e$, $\mathcal{A}$ represents the action space shared across environments, $\mathcal{P}^e(\boldsymbol{s}'|\boldsymbol{s}, a)$ denotes the transition dynamics characterizing the probability of transitioning to state $\boldsymbol{s}' \in \mathcal{S}^e$ given state $\boldsymbol{s} \in \mathcal{S}^e$ and action $a \in \mathcal{A}$, $p_0^e$ is the initial state distribution, $r^e(\boldsymbol{s}, a) : \mathcal{S}^e \times \mathcal{A} \to \mathbb{R}$ is the reward function, and $\gamma \in (0, 1)$ is the discount factor. To further formulate the imitation learning, we also denote $\rho^e$ as the corresponding occupancy measure of the expert policy, $\pi^e$, in the environment $e$. Specifically, we define $\rho_\pi^e(\boldsymbol{s}) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t P(s_t = \boldsymbol{s} \mid \pi^e)$ and $\rho_\pi^e(\boldsymbol{s}, a) = \pi(a \mid \boldsymbol{s})\rho_\pi^e(\boldsymbol{s})$, which represent the state and state-action visitation distributions under $\pi^e$. In particular, there exists a bijective correspondence between the policy $\pi^e$ and its occupancy measures $\rho_\pi^e(\boldsymbol{s})$ and $\rho_\pi^e(\boldsymbol{s}, a)$.

### 3.2 Offline Imitation Learning

In offline imitation learning, the agent has no access to environment interaction and the reward function $r^e$ is unknown. Instead, it is provided with a dataset of expert demonstrations $\mathcal{D}^e = \{(\boldsymbol{s}_t^e, a_t)\}_{t \geq 0}$ collected by an expert policy $\pi_{\text{exp}}^e$ in environment $e$.

We adopt the Maximum Entropy Inverse Reinforcement Learning (MaxEnt IRL) framework [Ho and Ermon, 2016],

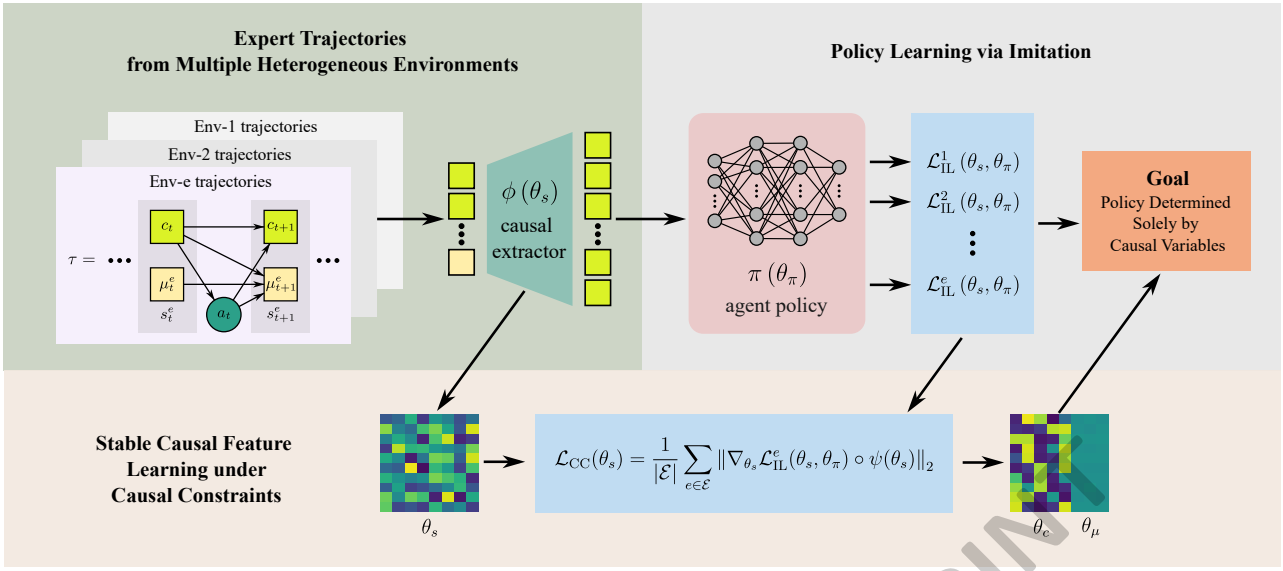



Figure 1: Block diagram of DCIL. Our DCIL performs imitation learning from expert demonstrations collected from a set of $K$ heterogeneous environments, which consists of two key components: (1) Stable causal feature learning under causal constraints to identify environment-invariant features; and (2) DropConnect-based regularization mechanism to avoid overfitting of these constraints to spurious correlations.

which formulates imitation learning as a state-action distribution matching problem between the expert and the learned policy. Following IQ-Learn [Garg *et al.*, 2021], the MaxEnt IRL objective can be reformulated into a non-adversarial optimization problem directly in the $Q$-function space by leveraging the inverse soft Bellman operator.

Specifically, the resulting offline imitation learning objective for environment $e$ is given by

$$\begin{aligned}\mathcal{L}^e_{\mathrm{IL}}(\boldsymbol{\theta_\pi}) = (1-\gamma)\,\mathbb{E}_{p^e_0}\left[V^Q(\boldsymbol{s}^e_0;\boldsymbol{\theta_\pi})\right] \\ - \mathbb{E}_{\rho^e_{\exp}}\left[\mathcal{T}^\pi[Q](\boldsymbol{s}^e,a;\boldsymbol{\theta_\pi})\right],\end{aligned} \tag{1}$$

where $V^Q(\boldsymbol{s}) = \log\sum_{a'}\exp(Q(\boldsymbol{s},a'))$ denotes the soft value function and $\mathcal{T}^\pi$ is the inverse soft Bellman operator defined under policy $\pi$.

The objective in Eq. (1) encourages the learned policy to match expert behavior using only offline data. However, when expert demonstrations are collected from heterogeneous environments, directly minimizing this objective may lead the policy to exploit spurious correlations between states and actions, resulting in poor generalization. In the following subsection, we introduce causal constraints to mitigate this issue and promote learning of environment-invariant decision mechanisms.

### 3.3 Stable Causal Feature Learning Under Causal Constraints

In each environment $e$, the observed state $\boldsymbol{s}^e_t$ admits a latent representation that can be decomposed into environment-invariant causal factors and environment-specific nuisance factors. Specifically, we assume that $\boldsymbol{s}^e_t$ is mapped into a latent space $\mathcal{Z}$ via a (possibly implicit) representation function, yielding

$$\boldsymbol{z}^e_t = (\boldsymbol{c}_t, \boldsymbol{\mu}^e_t), \tag{2}$$

where $\boldsymbol{c}_t$ denotes the latent causal variables that are shared across environments, and $\boldsymbol{\mu}^e_t$ captures the remaining environment-dependent factors. Motivated by the observation that expert actions often depend only on a subset of task-relevant features [Peters *et al.*, 2016; Yin *et al.*, 2024], we assume that actions are generated solely from the causal representation, i.e.,

$$\mathrm{Pa}(a_t) = \boldsymbol{c}_t. \tag{3}$$

However, since the mapping from the observed state $\boldsymbol{s}^e_t$ to its latent decomposition $(\boldsymbol{c}_t, \boldsymbol{\mu}^e_t)$ is unknown, the underlying causal features governing action generation remain unobserved.

To model these latent causal features and learn a policy that generalizes across environments, we introduce the following assumptions.

**Assumption 1** (Heterogeneity)**.** *The observed state $\boldsymbol{s}^e_t$ can be decomposed into causal variables $\boldsymbol{c}_t$ and non-causal variables $\boldsymbol{\mu}^e_t$. The non-causal variables are generated via a structural equation (illustrated in Figure 2): $\boldsymbol{\mu}^e_t \leftarrow f_\mu(\boldsymbol{c}_{t-1}, \boldsymbol{\mu}^e_{t-1}, a_{t-1}, \xi^e)$ and $\mathrm{Pa}(\boldsymbol{\mu}^e_t) = \{\boldsymbol{c}_{t-1}, \boldsymbol{\mu}^e_{t-1}, a_{t-1}, \xi^e\}$, where $\boldsymbol{\xi}^e$ denotes an environment-specific exogenous variable that represents a soft or hard intervention on the non-causal variables.*

**Assumption 2** (Invariance)**.** *The causal variables $\boldsymbol{c}_t$ are independent of the environment-specific exogenous variable $\xi^e$, i.e., $\boldsymbol{c}_t \perp \xi^e$. The causal variables are generated via a structural equation (illustrated in Figure 2): $\boldsymbol{c}_{t+1} \leftarrow f_c(\boldsymbol{c}_t, a_t)$ with $\mathrm{Pa}(\boldsymbol{c}_{t+1}) = \{\boldsymbol{c}_t, a_t\}$, and the corresponding transition dynamics across environments satisfy the invariance condition:*

$$\mathcal{P}^e(\boldsymbol{c}_{t+1} \mid \boldsymbol{c}_t, a_t) = \mathcal{P}^{e'}(\boldsymbol{c}_{t+1} \mid \boldsymbol{c}_t, a_t), \quad \forall e, e' \in \mathcal{E}.$$

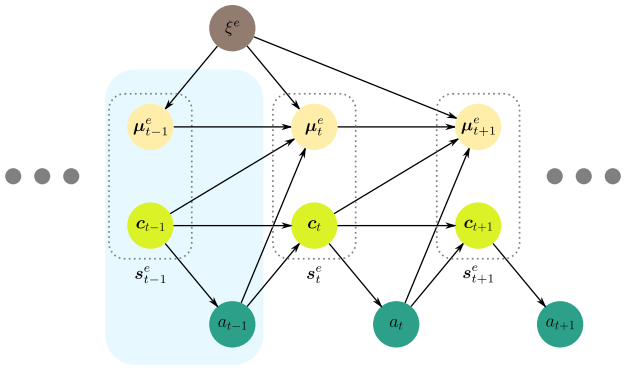



Figure 2: Illustration of the structural causal model in Assumptions 1 and 2.

*Furthermore, the expert policy conditioned on causal states is also invariant across environments:*

$$\pi^e_{exp}(a_t \mid \boldsymbol{c}_t) = \pi^{e'}_{exp}(a_t \mid \boldsymbol{c}_t), \quad \forall e, e' \in \mathcal{E}.$$

Assumption 1 facilitates the modeling of spurious correlations by distinguishing invariant causal variables from environment-specific confounders, thereby enabling us to account for heterogeneity across environments. Assumption 2 is crucial for learning a policy that generalizes well. It ensures that both the transition dynamics and expert behavior with respect to the causal variables remain stable across environments, allowing us to apply invariance principles to identify and exploit the shared causal mechanisms. Both assumptions are commonly adopted in invariance-based causal inference [Krueger *et al.*, 2021; Yin *et al.*, 2024].

Based on Assumptions 1 and 2, we consider a representation of the causal variables, i.e., $\boldsymbol{c}_t = \phi(\boldsymbol{s}^e_t; \boldsymbol{\theta_s})$, to encode the causal features of the action $a_t$, where $\boldsymbol{\theta_s} \in \mathbb{R}^{m\times|\mathcal{S}|}$ denotes the input-layer parameters of the policy network with $m$ the number of hidden units and $|\mathcal{S}|$ the input dimension. Then, the objective function in Eq. (1) can be adjusted as

$$\begin{aligned}\mathcal{L}^e_{\text{IL}}(\boldsymbol{\theta_s}, \boldsymbol{\theta_\pi}) = (1-\gamma)\mathbb{E}_{p^e_0}[V^Q(\phi(\boldsymbol{s}^e_0; \boldsymbol{\theta_s}); \boldsymbol{\theta_\pi})] \\ -\mathbb{E}_{\rho^e_{\text{exp}}}[\mathcal{T}^\pi[Q](\phi(\boldsymbol{s}^e; \boldsymbol{\theta_s}), a; \boldsymbol{\theta_\pi})].\end{aligned} \tag{4}$$

Next, we introduce causal constraints for optimization regularization, i.e.,

$$\mathcal{L}^e_{\text{CC}}(\boldsymbol{\theta_s}) = \|\nabla_{\boldsymbol{\theta_s}} \mathcal{L}^e_{\text{IL}}(\boldsymbol{\theta_s}, \boldsymbol{\theta_\pi}) \circ \psi(\boldsymbol{\theta_s})\|_2\,, \tag{5}$$

where $\circ$ denotes the Hadamard product and $\psi(\boldsymbol{\theta_s}) \in \mathbb{R}^{m\times|\mathcal{S}|}$ denotes the element-wise weighting matrix. Here we use the same weight for all $m$ elements in each column of $\boldsymbol{\theta_s}$. Specifically, we define the elements in the $j$-th column of $\psi(\boldsymbol{\theta_s})$ as $\tilde{w}_j \cdot \|\boldsymbol{\theta_s}[:, j]\|_2$, where $\boldsymbol{\theta_s}[:, j] \in \mathbb{R}^m$ denotes the $j$-th column of $\boldsymbol{\theta_s}$ and the scalar $\tilde{w}_j = \frac{1}{\|\boldsymbol{\theta_s}[:,j]\|_2+\varepsilon}$ acts as a column-wise normalization factor with $\varepsilon > 0$ a small constant added for numerical stability. Consequently, under Assumptions 1 and 2, the optimality condition of invariant causal features requires the gradients of the imitation learning objective with respect to their associated parameters to vanish across environments. In contrast, parameters corresponding to non-causal features induce environment-dependent gradients and are therefore suppressed through weight shrinkage. By penalizing the weighted gradient norm in Eq. (5), the proposed constraint drives gradients associated with causal features toward zero, while forcing the weights of non-causal features to diminish, thereby enforcing causal invariance at the representation level. By integrating this into Eq. (4), we arrive at the following loss function:

$$\mathcal{L}(\boldsymbol{\theta_s}, \boldsymbol{\theta_\pi}) = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}\left(\mathcal{L}^e_{\text{IL}}(\boldsymbol{\theta_s}, \boldsymbol{\theta_\pi}) + \mathcal{L}^e_{\text{CC}}(\boldsymbol{\theta_s})\right). \tag{6}$$

### 3.4 DropConnect-based Regularization Mechanism

Although causal constraints can encourage invariance across environments, recent studies have shown that such constraints may still lead to spurious solutions and overfitting, even when the corresponding penalty term vanishes [Lin *et al.*, 2022; Zhou *et al.*, 2022]. In practice, enforcing strong causal regularization often introduces additional optimization bias, making the learned policy overly sensitive to specific parameter configurations or training environments. Several existing approaches attempt to mitigate this issue through Bayesian formulations [Lin *et al.*, 2022] or sparsity-inducing regularization [Zhou *et al.*, 2022], which explicitly model uncertainty or restrict model capacity. While effective, these methods typically incur additional modeling complexity or require careful prior specification. In contrast, we adopt a simpler and more practical strategy based on DropConnect. By introducing stochastic perturbations directly at the weight level during training, DropConnect implicitly regularizes the causal constraint, discourages reliance on fragile feature–parameter associations, and effectively alleviates overfitting without altering the underlying optimization objective.

During each forward pass, we introduce a binary mask matrix $\boldsymbol{M} \in \{0,1\}^{m\times|\mathcal{S}|}$, whose elements are independently sampled from a Bernoulli distribution, i.e., $\boldsymbol{M}_{ij} \sim \text{Bernoulli}(1-p)$, where $p$ is the drop rate hyperparameter. The effective weight matrix used for forward computation is regularized as $\tilde{\boldsymbol{\theta}}_{\boldsymbol{s}} = \frac{1}{1-p} \cdot (\boldsymbol{\theta_s} \circ \boldsymbol{M})$, which are then used to compute the layer output in the standard manner. Accordingly, the total loss in Eq. (6) can be reformulated as $\mathcal{L}\left(\tilde{\boldsymbol{\theta}}_{\boldsymbol{s}}, \tilde{\boldsymbol{\theta}}_{\boldsymbol{\pi}}\right)$:

$$\frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}\left(\mathcal{L}^e_{\text{Drop-IL}}\left(\tilde{\boldsymbol{\theta}}_{\boldsymbol{s}}, \tilde{\boldsymbol{\theta}}_{\boldsymbol{\pi}}\right) + \mathcal{L}^e_{\text{Drop-CC}}\left(\tilde{\boldsymbol{\theta}}_{\boldsymbol{s}}\right)\right). \tag{7}$$

This stochastic masking mechanism introduces random variations in the model structure during training, encouraging the learned representations to be less dependent on individual connections and more robust to spurious correlations.

## 4 Experiments

### 4.1 Experimental Concerns

Our experiments are designed to answer the following questions: **Q1:** How does **DCIL** perform in unseen environments? **Q2:** What is the impact of the number of expert trajectories per environment on the algorithm? **Q3:** What is the

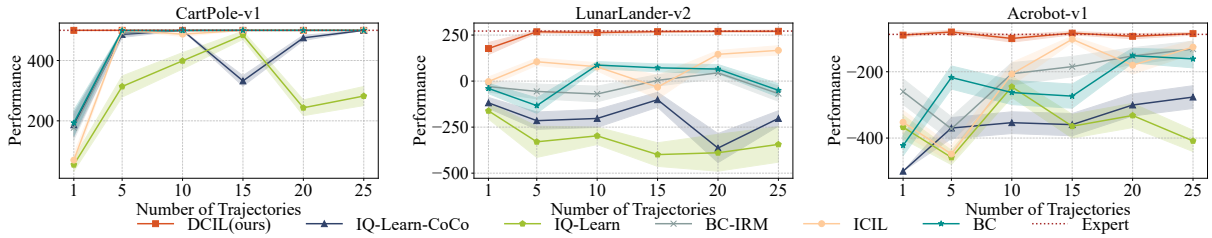



Figure 3: Learning performance evaluation with respect to the number of expert trajectories.

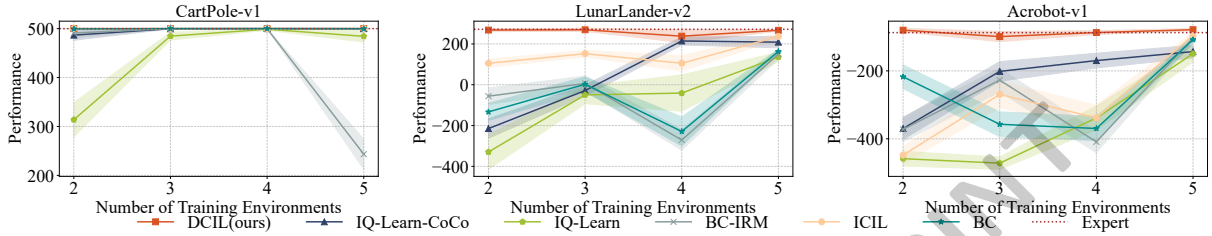



Figure 4: Learning performance evaluation with respect to the number of training environments.

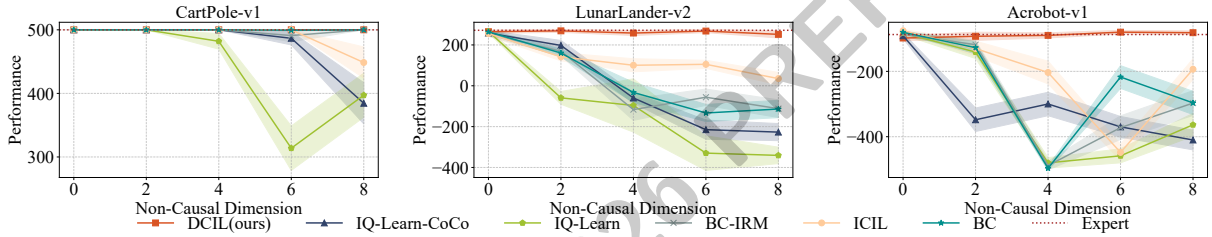



Figure 5: Learning performance evaluation with respect to the dimension of non-causal variables.



impact of the number of heterogeneous environments on the algorithm? **Q4:** What is the impact of the dimensionality of non-causal variables on the algorithm? **Q5:** How effective are the DropConnect-based causal constraints in mitigating the influence of spurious correlations from non-causal variables? **Q6:** What is the contribution of causal constraints and DropConnect-based regularization to the performance of DCIL? The experiments are organized into 5 parts: OpenAI Gym control tasks, the synthetic GMM dataset [Yin *et al.*, 2024], the visual dataset ColoredMNIST [Arjovsky *et al.*, 2019], the real-world dataset [Beery *et al.*, 2019], and an ablation study.

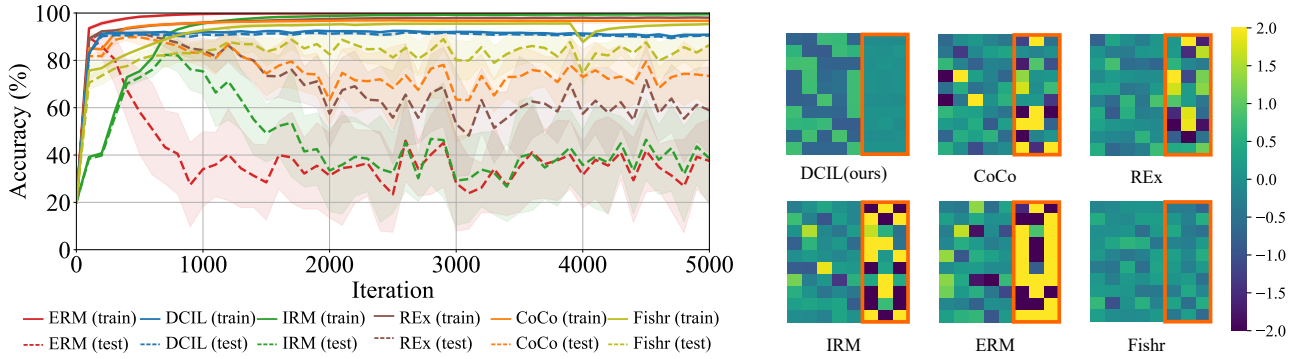
## 4.2 Control Tasks From OpenAI Gym

### Benchmark Methods

We compare DCIL with representative offline imitation learning baselines designed for multi-environment or invariant learning settings: (1) **BC**, which learns policies via supervised imitation; (2) **BC-IRM**, which combines BC with the IRMv1 objective [Arjovsky *et al.*, 2019] to encourage environment-invariant representations; (3) **IQ-Learn** [Garg *et al.*, 2021], a non-adversarial distribution-matching approach; (4) **IQ-Learn-CoCo**, which incorporates the CoCo regularization [Yin *et al.*, 2024] into IQ-Learn to promote invariant representations; and (5) **ICIL** [Bica *et al.*, 2021], which uses adversarial optimization to identify causal invariance across environments.

### Environments and Data Generation

We evaluate DCIL on three OpenAI Gym control tasks: Acrobot, CartPole, and LunarLander. To simulate heterogeneous environments, we augment the original state with additional environment-specific non-causal variables that introduce spurious correlations. Note that the original state may already contain non-causal features, and the augmentation allows us to explicitly control the dimensionality of spurious variables. For each environment $e \in \mathcal{E}$, we construct non-causal variables $\boldsymbol{\mu}_t^e \in \mathbb{R}^{d_z}$ from the original state features. Specifically, the non-causal variables are generated via an environment-specific linear transformation $\boldsymbol{T}^e \in \mathbb{R}^{d_c \times d_z}$, with $\boldsymbol{T}^e \sim \mathcal{N}(\xi^e, 1)$, where $\xi^e$ is an exogenous variable inducing distributional shifts across environments. We compute $\boldsymbol{\mu}_t^e = \begin{cases} \boldsymbol{c}_{\text{init}}^\top \boldsymbol{T}^e, & t = 0, \\ \boldsymbol{c}_{t-1}^\top \boldsymbol{T}^e, & t > 0, \end{cases}$ where $\boldsymbol{c}_{\text{init}} \sim \mathcal{U}(-1, 1)$. The augmented state is then given by $\boldsymbol{s}_t^e = [\boldsymbol{c}_t, \boldsymbol{\mu}_t^e] \in \mathbb{R}^{d_c + d_z}$. When $d_z = 0$, no additional non-causal variables are introduced, corresponding to the original state space, which is evaluated in Figure 5.




(a) Training and Testing Accuracy on GMM

(b) Heatmap of Causal Variable Weights on GMM

Figure 6: Performance evaluation on the synthetic GMM task. (a) Training/testing accuracy for CoCo, IRM, ERM, REx, Fishr and our DCIL. (b) Visualization of learned input-layer weights across different methods. The weights related to non-causal features $x_2$ are highlighted by an orange box.



**Learning Performance Evaluations**

To answer **Q1** and **Q2**, expert demonstrations were collected from two training environments, where the environment-specific matrices were sampled as $\boldsymbol{T}^1 \sim \mathcal{N}(1,1)^{d_c \times d_z}$ and $\boldsymbol{T}^2 \sim \mathcal{N}(-1,1)^{d_c \times d_z}$, with the non-causal dimensionality set to $d_z = 6$. **Test environments** were generated by sampling $\boldsymbol{T}^{\text{test}} \sim \mathcal{N}(4,1)^{d_c \times d_z}$. As shown in Figure 3, the x-axis denotes the number of expert trajectories per training environment, and the y-axis reports the average return of the learned policy, evaluated over **100** randomly sampled test environments. The results show that **DCIL** achieves near-expert performance using as few as 5 expert trajectories per training environment, and only one trajectory is sufficient in the CartPole task. In contrast, **ICIL** improves gradually with more demonstrations but fails to reach comparable performance, while **BC**, **BC-IRM**, and **IQ-Learn** remain limited in this setting.

To address **Q3**, we fixed the number of expert trajectories in each training environment to 5 and set $d_z = 6$. Source environments were constructed by sampling environment-specific matrices from $\mathcal{N}(1,1)$, $\mathcal{N}(-1,1)$, $\mathcal{N}(0,1)$, $\mathcal{N}(-2,1)$, and $\mathcal{N}(2,1)$, while test environments were sampled from $\mathcal{N}(4,1)^{d_c \times 6}$. Figure 4 illustrates the effect of the number of training environments on policy performance, where the x-axis indicates the number of source environments and the y-axis shows the average return evaluated over **100** randomly sampled test environments. As the number of training environments increases, the performance of **BC** and **ICIL** improves and approaches expert-level performance when trained on five environments. **IQ-Learn** also benefits from additional environments but remains consistently below the expert policy. **BC-IRM** exhibits unstable behavior in CartPole. In comparison, **DCIL** achieves near-expert performance using only two training environments.

To answer **Q4**, we collected expert demonstrations from two training environments with 5 expert trajectories each, where $\boldsymbol{T}^1 \sim \mathcal{N}(1,1)^{d_c \times d_z}$ and $\boldsymbol{T}^2 \sim \mathcal{N}(-1,1)^{d_c \times d_z}$. Test environments were sampled from $\mathcal{N}(4,1)^{d_c \times d_z}$. Figure 5 examines the impact of the non-causal dimensionality $d_z \in \{0,2,4,6,8\}$ on policy learning performance. The x-axis represents different values of $d_z$, and the y-axis shows the average return evaluated over **100** randomly sampled test environments. The results indicate that **DCIL** consistently maintains near-expert performance as $d_z$ increases, whereas the performance of baseline methods becomes increasingly unstable, suggesting that **DCIL** is robust to spurious correlations introduced by high-dimensional non-causal variables.

Next, to address **Q5**, we consider a synthetic Gaussian mixture model (GMM) classification task [Yin *et al.*, 2024], ColoredMNIST [Arjovsky *et al.*, 2019], and iWildCam 2019 dataset [Beery *et al.*, 2019]. These benchmarks involve high-dimensional and highly non-linear input spaces, and do not rely on imitation learning objectives. This setting allows us to isolate and evaluate the effectiveness of the proposed DropConnect-based causal constraint in suppressing spurious correlations under complex, non-linear representations, independent of the full DCIL framework.

## 4.3 Gaussian Mixture Simulation

We study multi-class classification with spurious correlations using a modified Gaussian Mixture Model. For environment $e$, data are generated via:

$$
\begin{aligned}
x_1^e &\leftarrow \sum_{k=1}^{K} \frac{1}{K} \mathcal{N}(\mu_k, \mathbf{I}), \\
y^e &\leftarrow \text{Categorical}(p_1, \ldots, p_K), \\
x_2^e &\leftarrow (1-p^e)\delta_{u_{y^e}} + p^e \delta_{u_{k_1}},
\end{aligned}
$$

where $p_k = \mathcal{N}(x_1^e; \mu_k, \mathbf{I}) / \sum_{k'} \mathcal{N}(x_1^e; \mu_{k'}, \mathbf{I})$ and $k_1 \sim \text{Multinomial}(1/K, \ldots, 1/K)$. Here $x_1^e$ is causal with invariant mapping to $y^e$, while $x_2^e$ has spurious environment-dependent correlations with $y^e$. We set $K = 5$, $\mu_k = \sqrt{1.5K} e_k \in \mathbb{R}^K$, and $u_k^e \sim \prod_{i=1}^{\lfloor k/2 \rfloor} \text{U}(0,1)$. Training uses $p^e \in \{0.01, 0.02\}$; test uses $p^f = 0$ with new $\{u_k^f\}$ to break spurious correlations. For performance comparison, we included several benchmark methods, including **CoCo** [Yin *et al.*, 2024], **IRM** [Arjovsky *et al.*, 2019], **ERM**, **REx** [Krueger *et al.*, 2021], and **Fishr** [Rame *et al.*, 2022]. The training/testing classification accuracy and the learned input-layer

| Method | 0.10 | 0.15 | 0.20 | 0.25 | 0.30 | 0.35 | 0.40 | 0.45 | 0.50 |
|---|---|---|---|---|---|---|---|---|---|
| ERM | $53.3 \pm 2.7$ | $63.4 \pm 2.6$ | $70.8 \pm 2.5$ | $75.4 \pm 2.5$ | $79.8 \pm 2.6$ | $82.1 \pm 2.7$ | $85.2 \pm 2.7$ | $87.4 \pm 2.6$ | $87.9 \pm 2.7$ |
| Drop | $61.2 \pm 5.6$ | $73.9 \pm 5.7$ | $82.2 \pm 5.5$ | $86.7 \pm 5.1$ | $89.7 \pm 4.3$ | $91.8 \pm 3.6$ | $93.5 \pm 3.2$ | $94.4 \pm 3.0$ | $95.1 \pm 2.9$ |
| IRMv1 | $25.6 \pm 4.8$ | $44.4 \pm 8.2$ | $59.7 \pm 7.9$ | $69.8 \pm 7.4$ | $77.3 \pm 6.1$ | $82.9 \pm 5.3$ | $86.9 \pm 4.5$ | $89.9 \pm 3.8$ | $91.9 \pm 3.1$ |
| CoCo | $52.5 \pm 5.0$ | $68.9 \pm 5.4$ | $79.2 \pm 5.7$ | $84.9 \pm 5.6$ | $88.6 \pm 5.2$ | $91.4 \pm 4.7$ | $93.2 \pm 4.3$ | $94.6 \pm 3.8$ | $95.5 \pm 3.2$ |
| **DCIL** | $\mathbf{62.9 \pm 9.5}$ | $\mathbf{78.5 \pm 10.2}$ | $\mathbf{85.4 \pm 9.6}$ | $\mathbf{89.0 \pm 8.8}$ | $\mathbf{91.4 \pm 8.0}$ | $\mathbf{93.0 \pm 7.1}$ | $\mathbf{94.3 \pm 6.2}$ | $\mathbf{95.2 \pm 4.8}$ | $\mathbf{95.9 \pm 3.7}$ |

Table 1: Test Accuracy (%) on CMNIST under Different Environment Probabilities ($p_e$).

weights across different methods are illustrated in Figure 6. In particular, the learned weights related to the non-causal variables $x_2$, which exhibit spurious correlations, are highlighted by an orange box in Figure 6 (b). Compared to other benchmark methods, the DropConnect-based causal constraint in DCIL successfully reduces spurious reliance by suppressing weights on $x_2$, promoting better generalization through causal stability.

### 4.4 ColoredMNIST

ColoredMNIST (CMNIST) [Arjovsky *et al.*, 2019] is a binary classification benchmark where digit colors are spuriously correlated with labels. To evaluate robustness under unknown environments, all training samples are pooled into a single mixed environment without source environment labels. The training data are generated with $p_e \in \{0.1, 0.15, \ldots, 0.5\}$ and the test environment uses $p_e = 0.9$. Additionally, 10% label noise is introduced. As shown in Table 1, DCIL consistently achieves the best performance across different $p_e$ settings, particularly under strong spurious correlations (small $p_e$), demonstrating superior robustness to environmental and label distribution shifts.

### 4.5 iWildCam 2019 Dataset

To further assess real-world applicability, we additionally conducted experiments on the iWildCam 2019 dataset [Beery *et al.*, 2019], which exhibits substantial distribution shifts across camera traps (e.g., background, illumination, and location). As shown in Table 2, DCIL achieves the best test accuracy (87.28%), outperforming all baseline methods. These results demonstrate that DCIL improves robustness under realistic distribution shifts.

### 4.6 Ablation Study

Finally, to answer **Q6**, we conducted ablation studies comparing DCIL with two ablated variants, DCIL[noDrop] and DCIL[noCC]. Specifically, **DCIL[noDrop]** disables the DropConnect-based regularization while **DCIL[noCC]** removes the causal constraint component used for identifying stable causal features. In addition, we also included the standard baseline **ERM** for comparison. Figure 7 visualizes the heatmaps of the first-layer weight matrices $\boldsymbol{\theta_s}$ learned by different methods on the GMM task under varying numbers of training environments ($|\mathcal{E}| = 2, 3, 4$). Each $10 \times 8$ matrix represents weights connecting the 8-dimensional input to a hidden layer with 10 units. The first 5 input elements correspond to causal variables, while the remaining 3 elements represent non-causal variables. It can be found that our DCIL outperforms other competing methods and optimizes the weights associated with non-causal inputs (highlighted by orange rectangles) to be closer to 0.

| Metric | ERM | Drop | V-REx | IRMv1 | Fishr | CoCo | DCIL |
|---|---|---|---|---|---|---|---|
| Train Accuracy | 100.00 | 100.00 | 91.60 | 90.71 | 89.15 | 88.21 | **84.07** |
| Test Accuracy | 57.80 | 61.46 | 79.19 | 83.24 | 83.04 | 78.23 | **87.28** |

Table 2: Training and test accuracy (%) on the iWildCam 2019 wildlife classification task.

## 5 Conclusion and Discussion

In this paper, we proposed a novel offline imitation learning framework designed to learn robust and generalizable policies from expert demonstrations collected across heterogeneous environments. To address the challenge of spurious correlations in multi-environment data, we introduced a causal constraint that regularized policy gradients toward stable causal features shared across environments. Additionally, we applied DropConnect-based regularization to inject stochastic perturbations into network parameters during training, which helped reduce overfitting to environment-specific non-causal patterns. This work highlighted the importance of integrating causal reasoning and principled regularization to enhance the robustness and transferability of imitation learning in offline and real-world settings.

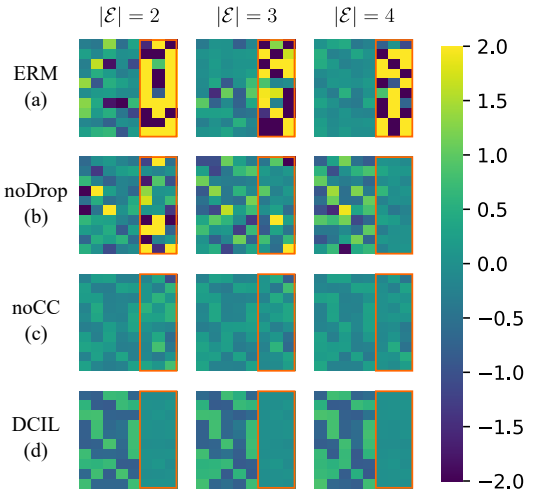



Figure 7: Heatmaps of first-layer weight matrices $\boldsymbol{\theta_s}$ from neural networks trained by ERM, DCIL[noDrop], DCIL[noCC], and DCIL on the GMM task.

## Acknowledgments

We thank the anonymous reviewers for their helpful comments on an earlier draft of the paper. Dr. Shihua Li and Xinyu Zhang were supported in this work by the Natural Science Foundation of Jiangsu Province of China [grant number 7708009046A]. Corresponding authors: Dr. Shihua Li and Dr. Rongjie Liu.

## References

[Arjovsky *et al.*, 2019] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2019.

[Arulkumaran *et al.*, 2017] Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 34(6):26–38, 2017.

[Beery *et al.*, 2019] Sara Beery, Grant van Horn, Oisin Mac Aodha, and Pietro Perona. The iwildcam 2018 challenge dataset, 2019.

[Bica *et al.*, 2021] Ioana Bica, Daniel Jarrett, and Mihaela van der Schaar. Invariant causal imitation learning for generalizable policies. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, pages 3952–3964. Curran Associates, Inc., 2021.

[Codevilla *et al.*, 2019] Felipe Codevilla, Eder Santana, Antonio M. Lopez, and Adrien Gaidon. Exploring the limitations of behavior cloning for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9328–9337, 2019.

[de Haan *et al.*, 2019] Pim de Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, pages 11693–11704. Curran Associates, Inc., 2019.

[Garg *et al.*, 2021] Divyansh Garg, Shuvam Chakraborty, Chris Cundy, Jiaming Song, and Stefano Ermon. Iq-learn: Inverse soft-q learning for imitation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, pages 4028–4039. Curran Associates, Inc., 2021.

[Hadfield-Menell *et al.*, 2017] Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J Russell, and Anca Dragan. Inverse reward design. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, pages 6765–6774. Curran Associates, Inc., 2017.

[Hasselt *et al.*, 2016] Hado van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pages 2094–2100, 2016.

[Ho and Ermon, 2016] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, pages 4565–4573. Curran Associates, Inc., 2016.

[Hu *et al.*, 2022] Anthony Hu, Gianluca Corrado, Nicolas Griffiths, Zachary Murez, Corina Gurau, Hudson Yeo, Alex Kendall, Roberto Cipolla, and Jamie Shotton. Model-based imitation learning for urban driving. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, pages 20703–20716. Curran Associates, Inc., 2022.

[Jang *et al.*, 2022] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In Aleksandra Faust, David Hsu, and Gerhard Neumann, editors, *Proceedings of the 5th Conference on Robot Learning*, pages 991–1002. PMLR, 2022.

[Krueger *et al.*, 2021] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021.

[Lillicrap *et al.*, 2015] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning, 2015.

[Lin *et al.*, 2022] Yong Lin, Hanze Dong, Hao Wang, and Tong Zhang. Bayesian invariant risk minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16021–16030, 2022.

[Mnih *et al.*, 2015] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[Pearl, 2009] J. Pearl. *Causality*. Cambridge University Press, 2009.

[Peters *et al.*, 2016] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016.

[Rame *et al.*, 2022] Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, pages 18347–18377. PMLR, 2022.

[Sonar *et al.*, 2021] Anoopkumar Sonar, Vincent Pacelli, and Anirudha Majumdar. Invariant policy optimization: Towards stronger generalization in reinforcement learning. In Ali Jadbabaie, John Lygeros, George J. Pappas, Pablo A. Parrilo, Benjamin Recht, Claire J. Tomlin, and Melanie N. Zeilinger, editors, *Proceedings of the 3rd Conference on Learning for Dynamics and Control*, pages 21–33. PMLR, 2021.

[Yin *et al.*, 2024] Mingzhang Yin, Yixin Wang, and David M. Blei. Optimization-based causal estimation from heterogeneous environments. *Journal of Machine Learning Research*, 25(168):1–44, 2024.

[Zhou *et al.*, 2022] Xiao Zhou, Yong Lin, Weizhong Zhang, and Tong Zhang. Sparse invariant risk minimization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, pages 27222–27244. PMLR, 2022.