Scalable Causal Imitation Learning

Eylam Tagor[1], **Mingxuan Li**[1], **Elias Bareinboim**[1]
eylam.tagor@columbia.edu, ml@cs.columbia.edu, eb@cs.columbia.edu

[1]**Department of Computer Science, Columbia University**

Abstract

Imitation learning enables learning a policy in an unknown environment with a latent reward signal using expert demonstrations, but it struggles when the imitator's and expert's observations are mismatched and unobserved confounders are present in expert demonstrations. By identifying appropriate adjustment sets via the sequential π-backdoor criterion, causal imitation learning (CIL) provides a framework for approximating the expert's policy from confounded data. However, existing CIL methods, Causal Behavioral Cloning (Causal BC) and Causal Generative Adversarial Imitation Learning (Causal GAIL), are designed for short-horizon, low-dimensional settings. When applied to continuous control tasks with long horizons and high-dimensional state-action spaces, these methods exhibit poor performance: Causal BC suffers from compounding errors, Causal GAIL is unstable and sample-inefficient, and sequential π-backdoor adjustment becomes impractical. We introduce Causal Soft Q Imitation Learning (SQIL) and Causal Inverse soft-Q Learning (IQ-Learn), two off-policy causal imitation learning algorithms that combine the causal adjustment framework with state-of-the-art inverse reinforcement learning objectives. Both algorithms operate on causally-adjusted state representations produced by an efficient approximation of the sequential π-backdoor criterion, exploiting the causal structure of continuous control environments to reduce the full-horizon adjustment to a fixed-size sliding window. We evaluate all methods in a suite of confounded environments and find that Causal SQIL and Causal IQ-Learn substantially outperform prior CIL algorithms on long-horizon tasks, sometimes surpassing the expert, whereas all causally unaware imitation methods fail to learn meaningful behavior.

Repository: https://github.com/scil-paper/Scalable-Causal-Imitation-Learning

1 Introduction

Imitation learning (IL) has become a central paradigm in robotics and control tasks as an alternative to reinforcement learning (RL) in domains where reward signals are unavailable, sparse, or difficult to engineer (Osa et al., 2018; Zare et al., 2024). Rather than optimizing a task-specific reward, IL seeks to learn from demonstrations collected as state-action trajectories from an expert deployed in the environment. This framework underlies a large body of work in behavioral cloning, dataset aggregation, and inverse RL (Ross et al., 2011; Ziebart et al., 2008; Ho & Ermon, 2016; Fu et al., 2018; Chi et al., 2023; Zhao et al., 2023), and is widely used in offline RL and robotics applications (Levine et al., 2020; Fu et al., 2021; Figueiredo Prudencio et al., 2024; Brohan et al., 2023).

Traditionally, IL methods assume that the expert and imitator operate with the same sensory capabilities, meaning every variable that the expert can observe is also observable to the imitator. Under this No Unobserved Confounders (NUC) assumption, the expert policy is identifiable from observational data and standard IL can recover it given sufficient demonstrations. However, realistic

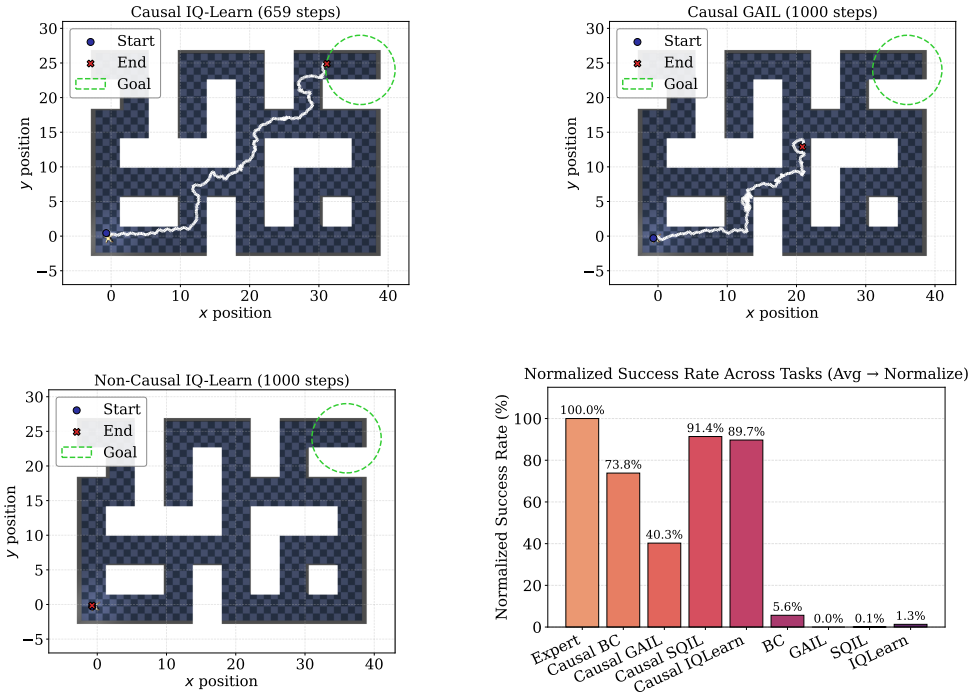

Figure 1: Performance of various imitation learning algorithms on the Confounded AntMaze Large task. (a) One of our proposed algorithms, Causal IQ-Learn. (b) Existing CIL methods such as Causal GAIL struggle to scale to high-dimensional long-horizon tasks. (c) Non-causal methods fail by overfitting to spurious correlations corrupted by unobserved confounding. (d) Success rates for all algorithms tested aggregated across all tasks evaluated; Causal SQIL and Causal IQ-Learn perform best through leveraging causal knowledge and scalable policy learning.

decision-making systems rarely satisfy NUC. In practice, experts often have access to additional sensors or context unavailable to the imitator, and latent conditions such as wind, friction, and payload may unpredictably change their distribution. These phenomena introduce unobserved confounding to the imitation task that may jointly affect the state transition, expert's action and rewards.

By falsely assuming NUC or ignoring unobserved confounding, standard IL methods overfit to spurious correlations during training, thus failing to generalize when these correlations shift at runtime (de Haan et al., 2019; Lu et al., 2023). Figure 1 illustrates this on a confounded maze navigation task: the observation contains quantities that correlate with expert actions but, due to influence from unobserved confounders, suffer distribution shift at runtime. Causally unaware methods mistake these spurious correlations for causal signals during training and incorporate them into their decision-making, ultimately failing to exhibit coherent movement during runtime and achieve near-0% success (Figure 1c,d). In safety-critical tasks in which IL is often applied due to its independence from reward engineering, such as autonomous driving (Chen et al., 2024; Codevilla et al., 2019) and robotic manipulation (Chi et al., 2023; Zheng et al., 2024), such failures are only revealed during deployment and thus pose an unaffordable risk which severely limits the utility of IL.

Causal imitation learning (CIL) addresses this by leveraging structural causal knowledge to identify which observed variables are safe to condition on (Zhang et al., 2020; Kumor et al., 2021; Ruan et al., 2023; 2024). However, existing CIL methods have remained restricted to low-dimensional, short-horizon tasks: in continuous control benchmarks (Todorov et al., 2012; Park et al., 2025) where observation spaces are high-dimensional and episodes span thousands of steps, they suffer from compounding error and training instability (Figure 1b). On the same large maze task, these

methods recover between $13 - 71\%$ of expert performance despite recovering $84 - 89\%$ on the medium maze (Table 1). A review of the CIL and IL literature is provided in Appendix A.

Collectively, we identify the gap in the current state of imitation learning literature: scalable methods can imitate long-term expert policies, but are fragile when NUC is violated; causal methods are robust under confounding, but struggle to scale. Our contributions address this gap by introducing scalable CIL through soft Q-learning methods augmented by an efficient approximation to the sequential π-backdoor adjustment sets that leverages structural properties of the environment. For evaluation, we introduce a suite of confounded control tasks based on OGBench (Park et al., 2025) where confounding biases are designed to mirror what is naturally found in real-world scenarios. We find that in these environments, existing CIL algorithms struggle to recover a consistent policy and non-causal IL algorithms fail altogether, whereas our proposed algorithms are able to achieve 90% of the expert's success rate on average (Figure 1d) and even surpass it on some tasks.

Notations. We will consistently use capital letters (X) to represent variables either in the observation or in the causal diagram, and lowercase (x) for their values. We bold capital letters ($\mathbf{X}$) for sets of variables. We denote the parents of the variable X in a causal diagram $\mathrm{pa}(X)$ and its children $\mathrm{ch}(X)$. X_t and X_H denote the instance of $\mathbf{X}$ at timestep t and at the last step, respectively. We use $P(X)$ as the probability distribution over X, $\pi(X \mid \mathbf{Z})$ as the behavioral policy distribution conditioning on $\mathbf{Z}$, and $do(x)$ as the intervention fixing X to take values x.

2 The Challenge of Imitation under Unobserved Confounding

We model the joint expert–environment system as a structural causal model (SCM).

Definition 1 (Structural Causal Model (Pearl, 2009; Bareinboim et al., 2022))**.** A structural causal model (SCM) is a tuple $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathscr{F}, P(\mathbf{u}) \rangle$, where:

- $\mathbf{U}$ is a set of exogenous variables determined by factors outside the model;
- $\mathbf{V} = \{V_1, \ldots, V_n\}$ is a set of endogenous variables determined by variables in $\mathbf{U} \cup \mathbf{V}$;
- $\mathscr{F} = \{f_V : V \in \mathbf{V}\}$ is a set of structural functions such that each $V \in \mathbf{V}$ is assigned via $V \leftarrow f_V(\mathrm{pa}(V), U_V)$, where $\mathrm{pa}(V) \subseteq \mathbf{V} \setminus \{V\}$ are endogenous parents of V and $U_V \subseteq \mathbf{U}$;
- $P(\mathbf{u})$ is a joint probability distribution over the exogenous variables $\mathbf{U}$.

Each SCM $\mathcal{M}$ induces a causal diagram $\mathcal{G}$ with one node for each $V \in \mathbf{V}$, directed edges $\mathrm{pa}(V) \to V$, and bidirected edges between variables that share an unobserved parent in $\mathbf{U}$.

In our setting, $\mathbf{V}$ includes states, actions, and latent environment variables, $\mathbf{X} \subseteq \mathbf{V}$ is the action set, and $Y \in \mathbf{U}$ is the latent reward. To differentiate between variables that are unobserved, observed, and observed by the expert only, we partition the endogenous variables into

$$\mathbf{V}^O \subseteq \mathbf{V} \quad \text{(observed to the imitator)}, \qquad \mathbf{V}^L = \mathbf{V} \setminus \mathbf{V}^O \quad \text{(latent to the imitator)}.$$

The expert demonstrations reflect the joint observational distribution $P(\mathbf{V}^O)$, whereas the imitator, operating under its own policy, induces the interventional distribution $P(\mathbf{V} \mid do(\pi))$. In the presence of latent variables $\mathbf{V}^L$, these two distributions may differ substantially: correlations between observed variables and expert actions may be driven by unobserved confounders rather than causal paths. The goal of CIL is therefore to determine, for each time step t, the sufficient subset of observed variables $\mathbf{Z}_t \subseteq \mathbf{V}^O$ for constructing an unbiased approximation of the expert's decision mechanism, $\pi_t(x_t \mid \mathbf{Z}_t) \approx P(x_t \mid \mathbf{Z}_t)$, in a way that is stable to the removal of latent confounding.

Example 1 (Confounded AntMaze)**.** Consider an ant robot navigating a maze toward a goal region, receiving a terminal reward Y upon success (see Figure 1 for visualization). The expert observes the full state $\mathbf{V}$, including its torso orientation $\mathbf{O}$, and selects joint torques $\mathbf{X}$ accordingly to move itself; Figure 2a shows this sequential structure. The imitator, however, does not observe $\mathbf{O}$ (i.e. $\mathbf{O} \in \mathbf{V}^L$). A compass sensor $\mathbf{W}$ serves as a noisy surrogate for the ant's bearing. The environment is also

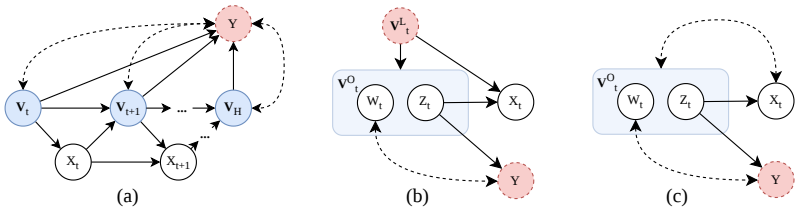

Figure 2: (a) Generic causal diagram for sequential imitation learning with unobserved confounders, as perceived by the expert. (b) Expanded single-timestep slice of a diagram with a spurious variable W_t, a useful state variable Z_t, and latent state $\mathbf{V}_t^L$. (c) Imitator perception of diagram (b), demonstrating how latent observation variables become additional unobserved confounders.

subject to a latent wind $\mathbf{U}$ (shown implicitly through bidirectional edges) that applies an external force to the dynamics, affecting the compass reading and the difficulty of reaching the goal.

Figure 2b shows a single-timestep slice of this structure. The compass W_t receives incoming edges from both the hidden orientation $O_t \in \mathbf{V}_t^L$ and the latent wind U_t, but has no outgoing edges: it is a collider that plays no causal role in determining future states, actions, or rewards. The useful state variables (position, joint angles, velocities) are represented by Z_t, which does causally influence the expert's action due to its edge into X_t. From the imitator's perspective (Figure 2c), the hidden $\mathbf{V}_t^L$ becomes an additional unobserved confounder, introducing a bidirected edge between $\mathbf{V}_t^O$ and X_t.

A causally unaware imitator that conditions on all observed variables, including W_t, inadvertently opens the spurious path $Y \leftarrow U_t \rightarrow W_t \leftarrow O_t \rightarrow X_t$. During training, the compass correlates with the expert's turning behavior because both are influenced by the wind. The imitator mistakes this for a causal signal: it learns, for instance, that when W_t points east the expert turns right, not realizing both facts are driven by an eastward gust. Once deployed under a different $P(U)$, these associations become actively harmful: the imitator turns into walls whenever the wind changes direction. ■

2.1 Sequential π-Backdoor Criterion

Formally, the sequential π-backdoor criterion graphically determines what each $\mathbf{Z}_t$ must contain so that conditioning on $\mathbf{Z}_t$ blocks all noncausal paths from X_t to the final outcome Y.

Definition 2 (Sequential π-Backdoor Criterion (Kumor et al., 2021))**.** Let $\mathcal{G}$ be the causal diagram induced by the SCM. For each action X_t, define a manipulated graph $\mathcal{G}_t'$ obtained by: (i) removing all incoming edges into future actions $\mathbf{X}_{t+1:H}$, and (ii) replacing each future action X_j $(j > t)$ by a node whose parents are restricted to $\mathbf{Z}_j$. A family of sets $\{\mathbf{Z}_t\}_{t=0}^H$ satisfies the sequential π-backdoor for $(\mathcal{G}, \mathbf{X}, Y)$ if, for every t, either $(X_t \perp\!\!\!\perp Y \mid \mathbf{Z}_t)_{(\mathcal{G}_t')_{X_t}}$ or $X_t \notin \mathrm{An}_{\mathcal{G}_t'}(Y)$. Here $(\mathcal{G}_t')_{X_t}$ denotes the graph obtained from $\mathcal{G}_t'$ by deleting outgoing edges from X_t.

When $\{\mathbf{Z}_t\}$ satisfies Definition 2, conditioning on $\mathbf{Z}_t$ removes all confounding and noncausal dependencies between X_t and Y that arise from shared latent parents in $\mathbf{V}^L$, proxy variables, or unobserved factors. Crucially, $\mathbf{Z}_t$ is restricted to the observable set $\mathbf{V}^O$. Returning to Example 1 and Figure 2, applying the sequential π-backdoor to the diagram yields an adjustment set $\mathbf{Z}_t$ that contains the useful state variables (position, joint angles, velocities) but excludes the compass W_t. By conditioning only on $\mathbf{Z}_t$, the imitator's policy $\pi(x_t \mid \mathbf{z}_t)$ is indifferent to the wind-driven distributional shift in W_t and instead relies exclusively on variables that causally determine the expert's actions. Even when the orientation O_t is unobserved and the full imitability condition is broken, learning a policy over $\mathbf{Z}_t \subseteq \mathbf{V}^O$ can still approximate the expert behavior because the remaining observed variables carry sufficient causal signal for navigation.

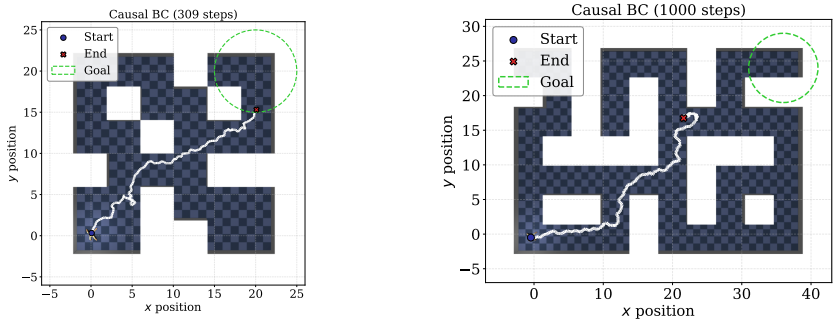

Figure 3: Causal BC on Confounded AntMaze. On Medium (left), the agent reaches the goal. On Large (right), the agent follows the expert path initially but drifts off course partway through; once outside the demonstration support, it cannot recover and fails to reach the goal.

3 Scalable Causal Imitation Learning

The sequential π-backdoor criterion identifies the correct conditioning set, but correct adjustment sets alone are not sufficient for imitation at scale. Prior CIL work has instantiated the criterion within two algorithmic paradigms: Causal Behavioral Cloning (Causal BC), or supervised cloning of the expert's conditional policy on the adjustment sets $\hat{\pi}_t(x_t \mid \mathbf{z}_t) = P(X_t \mid \mathbf{Z}_t)$ (Kumor et al., 2021), and Causal Generative Adversarial Imitation Learning (Causal GAIL), which matches occupancy measures over the adjustment sets via adversarial training (Ruan et al., 2023). Both exhibit fundamental scaling limitations, which we illustrate on the Confounded AntMaze task from Example 1.

Failure of Causal BC in long-horizon tasks. Even when supplied with correct adjustment sets, Causal BC remains a simple supervised learner. In long-horizon tasks, small prediction errors compound and the policy drifts into states not covered by the expert demonstrations, where its predictions are unreliable. While Causal BC achieves a respectable 88.9% normalized to the expert in AntMaze-Medium ($H \approx 300$ effective steps for successful solves), it drops to 71.2% in AntMaze-Large ($H \approx 700$ for successful episodes) and to 30.8% in HumanoidMaze-Medium ($H > 1000$ for successful episodes). This degradation is shown in Figure 3, where Causal BC easily navigates a medium maze; on a large maze, however, the agent initially follows the expert path but gradually drifts off course due to compounding error, entering unseen states from which it cannot recover.

Failure of Causal GAIL in high-dimensional domains. While Causal GAIL addresses compounding error by matching occupancy measures via adversarial training, it introduces its own scaling difficulties through its reliance on on-policy rollouts (typically via PPO) and a discriminator to distinguish expert from imitator trajectories. In long-horizon tasks, the agent must discover the complete path to the goal through exploration before the discriminator can provide a useful learning signal for later portions of the trajectory. In AntMaze-Medium, this exploration is still somewhat feasible and Causal GAIL achieves 84.6% success, but performance drops to 13.1% in AntMaze-Large and collapses to 0% by HumanoidMaze-Medium. Figure 1b shows that even in AntMaze-Large, the agent learns only coherent navigation for roughly the first two-thirds of the maze. This indicates a credit-assignment failure, as on-policy training collects too few complete traversals to propagate reward signal to later stages of navigation and prevents scaling to long-horizon tasks.

These failure modes point to a clear algorithmic requirement: off-policy methods that treat expert demonstrations as a static buffer, learn from self-collected transitions, and propagate reward signal across long horizons via temporal-difference learning. SQIL (Reddy et al., 2020) and IQ-Learn (Garg et al., 2021), two recent off-policy imitation learning methods built on soft Q-learning, satisfy all three properties but assume unconfounded environments. We now describe how to combine them with the causal adjustment framework.

3.1 Causal SQIL and Causal IQ-Learn

Given a π-backdoor admissible scope $\mathcal{S} = \{\langle X_t, \mathbf{Z}_t \rangle\}_{t=0}^{H-1}$, a straightforward causally-adjusted state representation would be $\mathbf{z}_t$ that concatenates the values of the variables in $\mathbf{Z}_t$ at each timestep. Both Causal SQIL and Causal IQ-Learn then operate on $(\mathbf{z}_t, x_t)$ pairs in place of the standard (s_t, a_t) pairs, applying their respective objectives to the adjusted representation.

Causal SQIL. SQIL (Reddy et al., 2020) assigns a fixed reward of $r = 1$ to expert transitions and $r = 0$ to policy transitions, then trains an SAC agent on the combined replay buffer. We apply this to causally-adjusted inputs $(\mathbf{z}_t, x_t)$: the critic minimizes the soft Bellman residual

$$y = r + \gamma\Big(\min_{j=1,2} Q_{\bar{\theta}_j}(\mathbf{z}', a') - \alpha \log \pi_\phi(a' \mid \mathbf{z}')\Big), \qquad \mathcal{L}_Q = \mathbb{E}\Big[\big(Q_\theta(\mathbf{z}, x) - y\big)^2\Big], \tag{1}$$

with $a' \sim \pi_\phi(\cdot \mid \mathbf{z}')$, and the actor maximizes the entropy-regularized objective

$$\mathcal{L}_\pi = \mathbb{E}_{\mathbf{z}}\Big[\alpha \log \pi_\phi(x \mid \mathbf{z}) - \min_{j=1,2} Q_{\theta_j}(\mathbf{z}, x)\Big], \quad x \sim \pi_\phi(\cdot \mid \mathbf{z}). \tag{2}$$

Because the adjustment set $\mathbf{Z}_t$ satisfies the π-backdoor criterion, the Q-function learns value estimates that are not confounded by latent variables, while temporal-difference learning propagates the expert signal across the full horizon. Full pseudocode is given in Algorithm 2 (Appendix C).

Causal IQ-Learn. IQ-Learn (Garg et al., 2021) learns a Q-function whose implicit reward is consistent with expert behavior. The critic enforces the soft Bellman equation on expert data,

$$\mathcal{L}_{\text{expert}} = \mathbb{E}_{(\mathbf{z}, x, \mathbf{z}') \sim \mathcal{D}_{\text{exp}}}\Big[\big(Q_\theta(\mathbf{z}, x) - \gamma\, V_{\bar{\theta}}(\mathbf{z}')\big)^2\Big], \tag{3}$$

where $V(\mathbf{z}) = \log \mathbb{E}_{x \sim \pi}[\exp Q(\mathbf{z}, x)]$, and applies a policy-consistency regularizer on policy data,

$$\mathcal{L}_{\text{reg}} = \mathbb{E}_{(\mathbf{z}, x) \sim \pi}\Big[\big(\log \pi_\phi(x \mid \mathbf{z}) - Q_\theta(\mathbf{z}, x) + V_\theta(\mathbf{z})\big)^2\Big]. \tag{4}$$

The combined loss $\mathcal{L}_Q = \mathcal{L}_{\text{expert}} + \lambda\, \mathcal{L}_{\text{reg}}$ trains the critic, grounding the implicit reward $r(\mathbf{z}, x) = Q(\mathbf{z}, x) - \gamma V(\mathbf{z}')$ in deconfounded state-action associations rather than spurious correlations. The actor uses the same entropy-regularized objective as Causal SQIL. Full pseudocode is given in Algorithm 3 (Appendix C).

The causal adjustment layer is algorithm agnostic. The downstream RL algorithm is unmodified; rather, it is the input representation that is changed. This property means that any future IL algorithm built on the soft Q learning approach can be made causal by the same procedure.

3.2 Scalable Causal Adjustment

Although the sequential π-backdoor yields a principled solution to the causal imitation problem, applying it swiftly becomes intractable as horizon and dimensionality increase. In a sequential decision-making problem with a terminal reward Y and a horizon $H \geq 1000$, the FINDOX algorithm of Kumor et al. (2021) (which returns the maximal admissible set $\mathbf{V}_X^Q$ for sequential π-backdoor adjustment; reproduced in Appendix C) operates over thousands of nodes and results in $|\mathbf{Z}_t|$ that linearly increases with t. In an SCM where the full histories of $\mathbf{V}$ and $\mathbf{X}$ have at least a potential causal effect on X_t, adjustment is inefficient and bloats observational dimensionality.

To make adjustment feasible, we propose an approximation of the true adjustment sets by generalizing structural properties of confounded control environments. We assume the following:

Assumption 1 (k-Bounded Time-Homogeneous Confounding)**.** Let $\mathcal{G}_H$ be the causal diagram induced by the SCM unrolled over horizon H, with endogenous variables $\mathbf{V}_t$ at each timestep t:

Algorithm 1 Windowed Sequential π-Backdoor Adjustment

Require: Window size k, full horizon H.

1: Construct proxy environment $\mathcal{E}_k$ with horizon $h = k + 1$ and extract its causal graph $\mathcal{G}_k$.
2: Run FINDOX$(\mathcal{G}_k, \mathbf{X}, Y) \to \mathbf{O}^X$. **If** $\mathbf{X} \not\subseteq \mathbf{O}^X$, **return** not imitable.
3: Compute ancestral graph $\mathcal{G}_k^Y$. Compute Markov boundary $\mathrm{MB} \leftarrow \mathrm{Pa}^+(C(\mathrm{Ch}^+(\mathbf{O}^X))) \setminus \mathbf{O}^X$ and boundary actions $\mathrm{BA} \leftarrow \{X_i \in \mathbf{X} \cap \mathbf{O}^X \mid \mathrm{ch}^+(X_i) \not\subseteq \mathbf{O}^X\}$ in $\mathcal{G}_k^Y$.
4: **for** each action X_t in $\mathcal{G}_k$ **do**
5: **if** $\mathbf{X}_t \in \mathrm{BA}$ **then**
6: $\mathbf{Z}_t^k \leftarrow (\mathrm{MB} \cup \mathrm{BA}) \cap \mathrm{before}(X_t)$
7: **else**
8: $\mathbf{Z}_t^k \leftarrow \varnothing$ *// $X_t \notin \mathrm{An}(Y)$ in $\mathcal{G}'_t$; satisfies condition (2) of Def. 2*
9: **end if**
10: **end for**
11: **for** $t = 0, \ldots, H-1$ **do**
12: $\mathbf{Z}_t^H \leftarrow \{(v, \tau) \in \mathbf{Z}_t^k \mid \tau \geq t - k\}$ *// Clip to window of width k*
13: **end for**
14: Build sliding window specification S: for each observed variable type V appearing in $\bigcup_t \mathbf{Z}_t^k$, enumerate lags $\{-1, \ldots, -k\}$ with dimension d_V.
15: **return** $\{\mathbf{Z}_t^H\}_{t=0}^{H-1}$, S.

(i) **k-bounded influence.** There exists a constant $k \geq 1$ such that no edge in $\mathcal{G}_H$ spans more than k timesteps: for every edge $\mathbf{V}_s \to \mathbf{V}_t$ or $\mathbf{V}_s \leftrightarrow \mathbf{V}_t$ ($\mathbf{X} \in \mathbf{V}$ in this case) in $\mathcal{G}_H$, $t - s \leq k$. Equivalently, $\mathrm{pa}(\mathbf{V}_t)$ is at most k steps away from $\mathbf{V}_t$.

(ii) **Endogenous time-homogeneity.** The structural functions $\mathscr{F}$ are identical at every timestep: for all t, t' and for any fixed $j \leq k$, the local causal structure among $(\mathbf{V}_{t-j}, \ldots, \mathbf{V}_t, X_t)$ is isomorphic to that among $(\mathbf{V}_{t'-j}, \ldots, \mathbf{V}_{t'}, x_{t'})$.

Condition (i) ensures that all causal influence on X_t, whether from observed state variables or unobserved confounders, is fully captured within a window of length k timesteps. Condition (ii) ensures that this local causal structure applies to any timestep in the unrolled graph, such that computing adjustment sets for one k-length window yields sufficient information of the relevant causal relationships to apply for all timesteps. Both conditions of Assumption 1 are naturally satisfied in physics-based continuous control environments such as MuJoCo, where dynamics depend solely on the immediate state and external forces from confounders have temporally localized effects.

Algorithm 1 implements this approximation in two stages. The first stage (Lines 1–10) solves the exact sequential π-backdoor on a short-horizon proxy graph $\mathcal{G}_k$ with horizon $k+1$. FINDOX (Kumor et al., 2021) identifies the maximal admissible set $\mathbf{O}^X$; if $\mathbf{X} \not\subseteq \mathbf{O}^X$, the problem is not imitable. The algorithm then constructs per-action adjustment sets from the Markov boundary MB of $\mathbf{O}^X$ in the ancestral graph $\mathcal{G}_k^Y$ (the minimal observed set blocking $\mathbf{O}^X$ from all other variables) and the boundary actions BA (actions whose causal effect on Y persists regardless of future actions) (Kumor et al., 2021, Definition 3.1, Lemma 3.2): boundary actions condition on $(\mathrm{MB} \cup \mathrm{BA}) \cap \mathrm{before}(X_t)$, while non-boundary actions require no conditioning. Intuitively, in Example 1 and Figure 2, the Markov boundary selects recent instances of $\mathbf{Z}$ and $\mathbf{X}$ within the k-step window while instances of $\mathbf{W}$, a collider with no outgoing edges, is excluded from $\mathbf{O}^X$ and thus from any adjustment set.

The second stage (Lines 11-15) transfers the proxy-graph adjustment sets to the full horizon. By time-homogeneity, the sets $\{\mathbf{Z}_t^k\}$ exhibit a repeating pattern of relative lags; clipping each to a window of k steps reduces dimensionality from $O(H)$ to $O(k)$. The output is a fixed-dimensional sliding window S that specifies adjustment set lags and dimensions for each timestep, used to encode inputs for the imitator. Its correctness is guaranteed by the following theorem (Proof in Appendix B).

Theorem 1 (Correctness of Windowed Adjustment)**.** Let $\mathcal{G}_H$ be the causal diagram of an SCM unrolled over horizon H, and let Assumption 1 hold with window size k. Let $\{\mathbf{Z}_t^k\}$ be adjustment sets that satisfy the sequential π-backdoor criterion (Definition 2) for the proxy graph $\mathcal{G}_k$ with horizon

Table 1: Evaluation results on confounded tasks. Normalized $\mathbb{E}[Y]$ linearly shifts the worst-performing algorithm to 0 due to the purely negative reward. Best non-expert result per task in **bold**. Raw data and standard errors can be found in Appendix F.1.

		Expert	C-BC	C-GAIL	C-SQIL	C-IQ-Learn	BC	GAIL	SQIL	IQ-Learn
AntMaze-Medium	Norm. $\mathbb{E}[Y]$	271.4	250.1	239.5	**275.8**	257.1	0.0	10.4	4.7	2.8
	Success rate (%)	87.6	77.9	74.1	**90.7**	84.3	0.0	0.0	0.0	0.0
AntMaze-Large	Norm. $\mathbb{E}[Y]$	229.0	199.3	155.6	204.3	**225.6**	14.0	85.8	0.0	16.1
	Success rate (%)	55.9	39.8	7.3	45.0	**58.9**	0.0	0.0	0.0	0.0
HumanoidMaze-Medium	Norm. $\mathbb{E}[Y]$	224.8	92.0	0.5	**192.2**	158.9	69.4	0.0	41.5	50.1
	Success rate (%)	33.8	10.4	0.0	**24.7**	19.1	5.4	0.0	0.1	2.4
HumanoidMaze-Large	Norm. $\mathbb{E}[Y]$	136.4	125.1	0.8	**139.3**	90.0	93.0	0.0	80.7	70.8
	Success rate (%)	7.0	**8.0**	0.0	**8.0**	3.0	5.0	0.0	0.0	0.0

$k+1$. Then the transferred sets $\{\mathbf{Z}_t^H\}_{t=0}^{H-1}$ returned by Algorithm 1 satisfy the sequential π-backdoor criterion for $(\mathcal{G}_H, \mathbf{X}, Y)$.

With windowed adjustment sets $\{\mathbf{Z}_t^H\}_{t=0}^{H-1}$ and specification S in hand, the causal encoding $\mathbf{z}_t$ is constructed at each timestep: for each observed variable type V appearing in the adjustment sets, we concatenate its values at lags $\{-1, \ldots, -k\}$ relative to t. This representation comprises the state observation in both Causal SQIL and Causal IQ-Learn, making the full procedure from causal graph analysis to policy optimization accessible and feasible for arbitrarily long horizons.

4 Experiments

We evaluate the proposed algorithms on confounded continuous-control locomotion environments derived from OGBench (Park et al., 2025). Each environment is defined as an SCM to support the modeling of unobserved confounders. In Confounded AntMaze (H=1000), an 8-DoF ant navigates a maze under latent wind; the imitator observes a wind-affected compass $\mathbf{W}$ in place of orientation $\mathbf{O}$ (see Example 1). In Confounded HumanoidMaze (H=2000), a 21-DoF humanoid navigates under latent seismic tremors; the imitator observes a tremor-affected vibration sensor $\mathbf{W}$ in place of the hidden center-of-mass velocity $\mathbf{C}$. In both environments, causal methods exclude $\mathbf{W}$ from the adjustment set, while causally unaware methods condition on it and thereby fail under distributional shift. Full environment details, causal graphs, and visualizations are provided in Appendix D.

We compare eight algorithms total, four causal (Causal BC (Kumor et al., 2021), Causal GAIL (Ruan et al., 2023), Causal SQIL (ours), Causal IQ-Learn (ours)) and four causally unaware (BC (Ross et al., 2011), GAIL (Ho & Ermon, 2016), SQIL (Reddy et al., 2020), IQ-Learn (Garg et al., 2021)), to isolate the contributions of causal adjustment and algorithmic choice. Expert policies are constructed via offline-to-online RL (BC and TD3 fine-tuning); full implementation details are in Appendix E.

4.1 Results

Causal variants use Algorithm 1 to compute their per-timestep observation using windowed causal adjustment. Non-causal variants condition on the full $\mathbf{V}^O$ at each timestep. All hyperparameters remain the same between causal and non-causal variants of the same algorithm.

All non-causal methods fail catastrophically. As seen in Table 1, non-causal algorithms consistently fail across all environments: standard BC, GAIL, SQIL, and IQ-Learn achieve 0% or near-0% success rates with low $\mathbb{E}[Y]$ across all tasks. This confirms that the fundamental failure is not a consequence of any particular learning paradigm but of the decision to condition on all observed variables, and no amount of temporal-difference learning or adversarial training can overcome a fundamentally misspecified conditioning set. The improvement of non-causal algorithms in HumanoidMaze tasks from AntMaze tasks, despite the former being more difficult, can be attributed to the less disruptive confounding effect of seismic tremors than wind fields on the state dynamics.

Causal adjustment is necessary but not sufficient for scaling. Although causal adjustment alone leads to significant improvements in most algorithms (Table 1), Causal BC and especially Causal

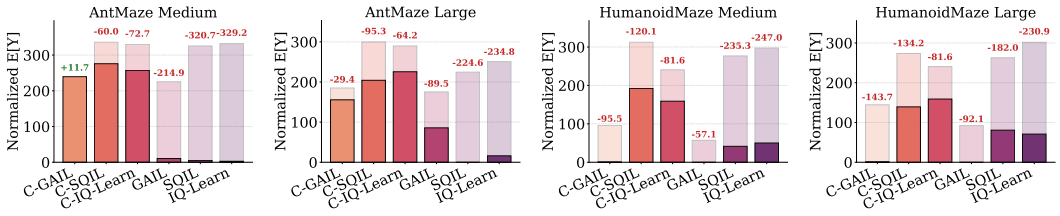

Figure 4: Evaluation return during training and runtime for GAIL, SQIL, and IQ-Learn (causal and non-causal variants). During training, both variants of each algorithm achieve comparable returns. The gap emerges only at runtime (Table 1) due to unobserved confounders causing distributional changes from where the expert demonstrations were collected, which affects only non-causal methods. Hyperparameters and other training metrics are provided in detail in Appendix E.

GAIL see substantial performance decreases as the task horizon and dimensionality increases, with Causal GAIL collapsing to about the level of non-causal GAIL by the HumanoidMaze-Medium task. Causal SQIL and Causal IQ-Learn scale more gracefully despite increasing dimensionality and horizon, and at times surpassing the expert's performance.

Confounding cannot be revealed by in-distribution evaluation. Figure 4 reveals that during training, when the imitator operates under the same confounder distribution $P(\mathbf{U})$ as the expert, non-causal variants of each algorithm achieve nearly identical evaluation returns to causal variants, and in some cases appear better since conditioning on the spurious proxy $\mathbf{W}$ provides additional predictive signal that is useful when under the training $P(\mathbf{U})$. At runtime, when $P(\mathbf{U})$ shifts, every non-causal method collapses to near-0% success while the causal methods retain significantly more of the performance. Thus, confounding cannot be diagnosed during training.

Windowed approximation is necessary. Figure 5 demonstrates that a moderate $k \in [1, 10]$ is ideal for imitation, whereas $k = 0$ (pure Markov) and $k = 100$ see significant drops in most methods. This implies that while the environments require sequential decision-making capabilities, the relevant causal effects on any X_t can be captured within a few timesteps; meanwhile, large k bloats the representation. Interestingly, at $k = 100$, Q-learning methods collapse due to bootstrapping instability in high-dimensional state spaces while BC remains robust due to its supervised learning approach.

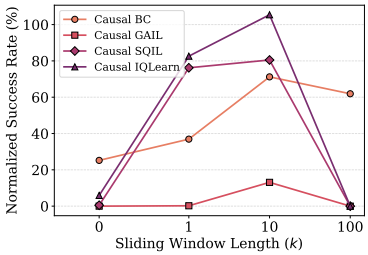

Figure 5: Sensitivity of causal algorithms to k in Algorithm 1 on Confounded AntMaze-Large.

The causality gap dominates the algorithmic gap. In every task, the causal variant of each algorithm outperforms its non-causal counterpart, with the exception of Causal GAIL and non-causal GAIL both achieving 0% success rate in HumanoidMaze tasks (Table 1). Thus, causal reasoning remains the primary determinant of success in confounded environments. The algorithm choice becomes the secondary yet still substantial factor that determines how well a causal method performs.

5 Conclusion

We introduce Causal SQIL and Causal IQ-Learn, off-policy CIL algorithms that combine sequential π-backdoor adjustment with soft Q-learning objectives, and windowed causal adjustment that is tractable for long-horizon control. Our experiments demonstrate that causal adjustment is necessary for robust policy learning and that off-policy Q-learning is necessary for scaling. While our approach relies on knowledge of the causal diagram (see Appendix G for a full discussion on limitations), it is algorithm-agnostic. Thus, looking forward, it can be composed with expressive policy classes (e.g. diffusion policies (Chi et al., 2023) or action-chunking transformers (Zhao et al., 2023)), high-

dimensional sensory inputs where confounding manifests through pixel-level corruptions (Li et al., 2025; 2026; Juliani et al., 2026), and any setting in which the expert and imitator operate under different sensor configurations, such as tele-operation, sim-to-real transfer, and multi-agent imitation. As unobserved confounding is the norm in real-world deployment, integrating causal reasoning with scalable policy learning is essential for trustworthy imitation in safety-critical domains.

References

Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, pp. 1, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581138385. DOI: 10.1145/1015330.1015430. URL https://doi.org/10.1145/1015330.1015430.

Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. *On Pearl's Hierarchy and the Foundations of Causal Inference*, pp. 507–556. Association for Computing Machinery, New York, NY, USA, 1 edition, 2022. ISBN 9781450395861. URL https://doi.org/10.1145/3501714.3501743.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alex Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *arXiv preprint arXiv:2307.15818*, 2023.

Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12):10164–10183, December 2024. ISSN 0162-8828. DOI: 10.1109/TPAMI.2024.3435937. URL https://doi.org/10.1109/TPAMI.2024.3435937.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

Felipe Codevilla, Eder Santana, Antonio Lopez, and Adrien Gaidon. Exploring the limitations of behavior cloning for autonomous driving. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9328–9337, 2019. DOI: 10.1109/ICCV.2019.00942.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/947018640bf36a2bb609d3557a285329-Paper.pdf.

Rafael Figueiredo Prudencio, Marcos R. O. A. Maximo, and Esther Luna Colombini. A survey on offline reinforcement learning: Taxonomy, review, and open problems. *IEEE Transactions on Neural Networks and Learning Systems*, 35(8):10237–10257, August 2024. ISSN 2162-2388. DOI: 10.1109/tnnls.2023.3250269. URL http://dx.doi.org/10.1109/TNNLS.2023.3250269.

Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adverserial inverse reinforcement learning. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rkHywl-A-.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4{rl}: Datasets for deep data-driven reinforcement learning, 2021. URL https://openreview.net/forum?id=px0-N3_KjA.

Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning*, pp. 1587–1596, 2018. URL https://arxiv.org/abs/1802.09477.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 4028–4039. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/210f760a89db30aa72ca258a3483cc7f-Paper.pdf.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pp. 1861–1870, 2018. URL https://arxiv.org/abs/1801.01290.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/cc7e2b878868cbae992d1fb743995d8f-Paper.pdf.

Mateo Juliani, Mingxuan Li, and Elias Bareinboim. Confounding robust continuous control via automatic reward shaping. In *The 25th International Conference on Autonomous Agents and Multi-Agent Systems*, 2026. URL https://openreview.net/forum?id=ZFtjCJqEQf.

Daniel Kumor, Junzhe Zhang, and Elias Bareinboim. Sequential causal imitation learning with unobserved confounders. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 14669–14680. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/7b670d553471ad0fd7491c75bad587ff-Paper.pdf.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020. URL https://arxiv.org/abs/2005.01643.

Mingxuan Li, Junzhe Zhang, and Elias Bareinboim. Confounding robust deep reinforcement learning: A causal approach. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=9fUr5iFU9j.

Mingxuan Li, Junzhe Zhang, and Elias Bareinboim. Causal flow q-learning for robust offline reinforcement learning, 2026. URL https://arxiv.org/abs/2602.02847.

Yiren Lu, Justin Fu, George Tucker, Xinlei Pan, Eli Bronstein, Rebecca Roelofs, Benjamin Sapp, Brandyn White, Aleksandra Faust, Shimon Whiteson, Dragomir Anguelov, and Sergey Levine. Imitation is not enough: Robustifying imitation with reinforcement learning for challenging driving scenarios. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7553–7560, 2023. DOI: 10.1109/IROS55552.2023.10342038.

Pedro A Ortega, Markus Kunesch, Grégoire Delétang, Tim Genewein, Jordi Grau-Moya, Joel Veness, Jonas Buchli, Jonas Degrave, Bilal Piot, Julien Perolat, et al. Shaking the foundations: delusions in sequence models for interaction and control. *arXiv preprint arXiv:2110.10819*, 2021. URL https://arxiv.org/abs/2110.10819.

Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J. Andrew Bagnell, Pieter Abbeel, and Jan Peters. An algorithmic perspective on imitation learning. *Foundations and Trends® in Robotics*, 7(1–2): 1–179, March 2018. ISSN 1935-8261. DOI: 10.1561/2300000053. URL http://dx.doi.org/10.1561/2300000053.

Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. OGBench: Benchmarking offline goal-conditioned RL. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=M992mjgKzI.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

Samuel Pfrommer, Yatong Bai, Hyunin Lee, and Somayeh Sojoudi. Initial state interventions for deconfounded imitation learning. In *CDC*, pp. 2312–2319, 2023. URL https://doi.org/10.1109/CDC49753.2023.10383252.

Dean A. Pomerleau. Alvinn: An autonomous land vehicle in a neural network. In D. Touretzky (ed.), *Advances in Neural Information Processing Systems*, volume 1. Morgan-Kaufmann, 1988. URL https://proceedings.neurips.cc/paper_files/paper/1988/file/812b4ba287f5ee0bc9d43bbf5bbe87fb-Paper.pdf.

Siddharth Reddy, Anca D. Dragan, and Sergey Levine. SQIL: imitation learning via reinforcement learning with sparse rewards. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=S1xKd24twB.

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 627–635, 2011. URL https://arxiv.org/abs/1011.0686.

Kangrui Ruan, Junzhe Zhang, Xuan Di, and Elias Bareinboim. Causal imitation learning via inverse reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=B-z41MBL_tH.

Kangrui Ruan, Junzhe Zhang, Xuan Di, and Elias Bareinboim. Causal imitation for markov decision processes: a partial identification approach. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 87592–87620. Curran Associates, Inc., 2024. DOI: 10.52202/079017-2781. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/9f7f2f57d8eaf44b2f09020f64ff6d96-Paper-Conference.pdf.

Daqian Shao, Thomas Kleine Buening, and Marta Kwiatkowska. A unifying framework for causal imitation learning with hidden confounders. In *Workshop on Spurious Correlation and Shortcut Learning: Foundations and Solutions*, 2025. URL https://openreview.net/forum?id=arlxXpjWGZ.

Jonathan C. Spencer, Sanjiban Choudhury, Arun Venkatraman, Brian D. Ziebart, and J. Andrew Bagnell. Feedback in imitation learning: The three regimes of covariate shift. *CoRR*, abs/2102.02872, 2021. URL https://arxiv.org/abs/2102.02872.

Gokul Swamy, Sanjiban Choudhury, J. Bagnell, and Steven Z. Wu. Sequence model imitation learning with unobserved contexts. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17665–17676. Curran Associates, Inc., 2022a. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/708e58b0b99e3e62d42022b4564bad7a-Paper-Conference.pdf.

Gokul Swamy, Sanjiban Choudhury, J. Andrew Bagnell, and Zhiwei Steven Wu. Causal imitation learning under temporally correlated noise. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20877–20890, 2022b. URL https://proceedings.mlr.press/v162/swamy22a.html.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012. DOI: 10.1109/IROS.2012.6386109.

Risto Vuorio, Pim De Haan, Johann Brehmer, Hanno Ackermann, Daniel Dijkman, and Taco Cohen. Deconfounding imitation learning with variational inference. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=3FsVtsISHW. Expert Certification.

Chuan Wen, Jierui Lin, Trevor Darrell, Dinesh Jayaraman, and Yang Gao. Fighting copycat agents in behavioral cloning from observation histories. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 2564–2575. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1b113258af3968aaf3969ca67e744ff8-Paper.pdf.

Maryam Zare, Parham M. Kebria, Abbas Khosravi, and Saeid Nahavandi. A survey of imitation learning: Algorithms, recent developments, and challenges. *IEEE Transactions on Cybernetics*, 54(12):7173–7186, 2024. DOI: 10.1109/TCYB.2024.3395626.

Yan Zeng, Shenglan Nie, Feng Xie, Libo Huang, Peng Wu, and Zhi Geng. Confounded causal imitation learning with instrumental variables. *CoRR*, abs/2507.17309, July 2025. URL https://doi.org/10.48550/arXiv.2507.17309.

Junzhe Zhang, Daniel Kumor, and Elias Bareinboim. Causal imitation learning with unobserved confounders. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12263–12274. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/8fdd149fcaa7058caccc9c4ad5b0d89a-Paper.pdf.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems*, 2023. URL https://arxiv.org/abs/2304.13705.

Boyuan Zheng, Sunny Verma, Jianlong Zhou, Ivor W. Tsang, and Fang Chen. Imitation learning: Progress, taxonomies and challenges. *IEEE Transactions on Neural Networks and Learning Systems*, 35(5):6322–6337, 2024. DOI: 10.1109/TNNLS.2022.3213246.

Brian D. Ziebart, Andrew L. Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, pp. 1433–1438, 2008. URL https://cdn.aaai.org/AAAI/2008/AAAI08-227.pdf.

Supplementary Materials

The following content was not necessarily subject to peer review.

A Related Work

Imitation learning. Imitation learning (IL) trains policies from expert demonstrations without access to a reward signal. Behavioral cloning (BC) reduces IL to supervised learning (Pomerleau, 1988), but suffers from compounding errors due to covariate shift (Ross et al., 2011). Interactive methods such as DAgger (Ross et al., 2011) mitigate this by querying the expert on-policy, though at the cost of requiring an interactive demonstrator. Inverse reinforcement learning (IRL) recovers a reward function rationalizing expert behavior (Abbeel & Ng, 2004; Ziebart et al., 2008); adversarial formulations such as GAIL (Ho & Ermon, 2016) and AIRL (Fu et al., 2018) cast IRL as occupancy-measure matching, avoiding explicit reward modeling but relying on on-policy rollouts and adversarial training that scale poorly to long horizons. More recently, off-policy value-based methods have achieved strong results on continuous-control benchmarks: SQIL (Reddy et al., 2020) reformulates IL as soft Q-learning with binary rewards, and IQ-Learn (Garg et al., 2021) learns a Q-function whose implicit reward is consistent with expert data. Both build on the SAC framework (Haarnoja et al., 2018) and propagate the expert signal across trajectories via temporal-difference learning, making them substantially more effective than BC or GAIL on long-horizon tasks. Modern IL has also scaled to real-world robotics through expressive policy classes such as diffusion policies (Chi et al., 2023) and action-chunking transformers (Zhao et al., 2023). All of these methods, however, assume that the expert and imitator share the same observation space, i.e., No Unobserved Confounders (NUC). The present work retains the scalability advantages of off-policy soft Q-learning while relaxing NUC via causal adjustment.

Formal objectives of prior CIL and IL methods. For reference, we provide the formal objectives of the methods discussed in the main text. Causal BC (Kumor et al., 2021) directly clones the expert's conditional policy over the admissible adjustment variables, $\hat{\pi}_t(x_t \mid \mathbf{z}_t) = P(X_t \mid \mathbf{Z}_t)$, via supervised learning on expert demonstrations. Causal GAIL (Ruan et al., 2023) extends the framework to inverse reinforcement learning by matching expert and imitator occupancy measures over the adjusted variables:

$$\min_{\pi \sim \mathcal{S}} \max_{D} \; \mathbb{E}[\log D(\mathbf{z}_t, x_t)] + \mathbb{E}_{\pi}[\log(1 - D(\mathbf{z}_t, x_t))] \,. \tag{5}$$

In non-causal settings, SQIL (Reddy et al., 2020) assigns a fixed reward of 1 to expert transitions and 0 to policy transitions, and then applies SAC on the combined replay buffer. The SAC critic minimizes the soft Bellman residual

$$y = r + \gamma \Big(\min_{j=1,2} Q_{\bar{\theta}_j}(s', a') - \alpha \log \pi_\phi(a' \mid s') \Big), \qquad \mathcal{L}_Q = \mathbb{E}\Big[\big(Q_\theta(s,a) - y \big)^2 \Big], \tag{6}$$

with $a' \sim \pi_\phi(\cdot \mid s')$. The actor maximizes the entropy-regularized objective

$$\mathcal{L}_\pi = \mathbb{E}_s \Big[\alpha \log \pi_\phi(a \mid s) - \min_{j=1,2} Q_{\theta_j}(s,a) \Big], \quad a \sim \pi_\phi(\cdot \mid s). \tag{7}$$

IQ-Learn (Garg et al., 2021) learns a Q-function whose induced reward is consistent with expert behavior. The critic enforces the soft Bellman equation on expert data,

$$\mathcal{L}_{\text{expert}} = \mathbb{E}_{(s,a,s') \sim \mathcal{D}_{\text{exp}}} \Big[\big(Q_\theta(s,a) - \gamma \, V_{\bar{\theta}}(s') \big)^2 \Big], \tag{8}$$

where $V(s) = \log \mathbb{E}_{a \sim \pi}[\exp Q(s,a)]$, and applies a policy-consistency regularizer on policy data,

$$\mathcal{L}_{\text{reg}} = \mathbb{E}_{(s,a) \sim \pi} \Big[\big(\log \pi_\phi(a \mid s) - Q_\theta(s,a) + V_\theta(s) \big)^2 \Big]. \tag{9}$$

The combined loss $\mathcal{L}_Q = \mathcal{L}_{\text{expert}} + \lambda \, \mathcal{L}_{\text{reg}}$ trains the critic, while the actor uses the same objective as above. Despite their scalability advantages, these methods assume fully observed (i.e., unconfounded) environments; using them in the presence of latent confounders leads to biased policies.

Causal imitation learning. Causal imitation learning (CIL) leverages structural causal knowledge to approximate the expert policy in the presence of unobserved confounding. Zhang et al. (2020) introduces the π-backdoor criterion, a complete graphical condition for determining policy imitability from observational data when the expert and imitator have different sensory inputs. Kumor et al. (2021) extend this to sequential decision-making with the sequential π-backdoor criterion and the FINDOX algorithm for constructing per-timestep adjustment sets. Ruan et al. (2023) develop CIL via inverse reinforcement learning, and Ruan et al. (2024) introduce a partial-identification approach that can enable the imitator to surpass expert performance. In practice, these methods have been instantiated as Causal BC and Causal GAIL, both of which assume access to a π-backdoor admissible scope. Despite their theoretical appeal, existing CIL algorithms have remained restricted to low-dimensional, short-horizon domains: Causal BC inherits the covariate-shift limitations of behavioral cloning, Causal GAIL's reliance on on-policy rollouts and adversarial training leads to sample and stability challenges at scale, and the adjustment sets produced by FINDOX grow linearly with the horizon, making exact computation infeasible in long-horizon settings. Our work addresses all three limitations by combining the causal adjustment framework with off-policy soft Q-learning objectives and introducing a windowed approximation that reduces the adjustment set computation to a fixed-size sliding window.

Confounding, causal confusion, and spurious correlations in IL. A related line of work studies the problem of spurious correlations in IL without assuming access to a causal graph. de Haan et al. (2019) identify "causal confusion," showing that conditioning on non-causal features can degrade imitation performance, and propose targeted interventions to select the correct causal model. Ortega et al. (2021) formalize "delusions" in sequence models, demonstrating that treating past actions as observations rather than interventions leads to incorrect inference. Variations of this problem have been studied under different names: as feedback-driven covariate shift (Spencer et al., 2021), as the copycat problem in BC from observation histories (Wen et al., 2020), as confounding in driving settings (Codevilla et al., 2019), and as deconfounding via initial-state interventions (Pfrommer et al., 2023). These works diagnose the problem and propose heuristic or environment-specific solutions. In contrast, the CIL framework this paper builds on provides a complete, non-parametric graphical criterion for determining imitability and identifying the correct adjustment set.

Several works address confounded IL through mechanisms other than graphical backdoor adjustment. Swamy et al. (2022b) apply instrumental variable (IV) regression to handle temporally correlated noise in expert actions, proposing DoubIL and ResiduIL. Swamy et al. (2022a) prove that on-policy sequence models can asymptotically recover expert behavior when the expert observes privileged information, while showing that off-policy methods "latch" onto confounded features—a failure mode that directly motivates the causal adjustment layer in our off-policy algorithms. Vuorio et al. (2024) train a variational inference model to infer the expert's latent information and learn a latent-conditional policy. Shao et al. (2025) propose DML-IL, a unifying framework that reformulates causal IL as a conditional moment restriction problem solved via IV regression, handling both expert-observable and expert-unobservable confounders. Zeng et al. (2025) similarly leverage IVs for confounded sequential IL. These IV and latent-inference approaches typically assume additive or temporally bounded confounding for instrument validity, or require ergodic dynamics for latent inference; they do not require a known causal graph but trade this for stronger distributional assumptions. Our approach instead assumes a known causal graph and applies non-parametric graphical adjustment, which is complementary: it provides exact deconfounding guarantees when the graph is available, and our windowed approximation makes it tractable at scale.

Causal reinforcement learning. Causal reasoning has been increasingly integrated into RL to handle confounded offline data and distribution shift. Li et al. (2025) develop confounding-robust deep RL via causal adjustment in the presence of unobserved confounders; Li et al. (2026) extend this to offline settings via causal flow Q-learning; and Juliani et al. (2026) address confounders in continuous control through automatic reward shaping. These works operate in the RL setting where a reward signal is available, whereas the present paper addresses the harder IL setting where the

reward is entirely latent. Nevertheless, the structural causal formalization and the challenge of off-policy learning under confounders are shared, and our causal adjustment layer is compatible with advances in causal RL.

B Proof

Theorem 2 (Restatement of Theorem 1)**.** Let $\mathcal{G}_H$ be the causal diagram of an SCM unrolled over horizon H, and let Assumption 1 hold with window size k. Let $\{\mathbf{Z}_t^k\}$ be adjustment sets that satisfy the sequential π-backdoor criterion (Definition 2) for the proxy graph $\mathcal{G}_k$ with horizon $k{+}1$. Then the transferred sets $\{\mathbf{Z}_t^H\}_{t=0}^{H-1}$ returned by Algorithm 1 satisfy the sequential π-backdoor criterion for $(\mathcal{G}_H, \mathbf{X}, Y)$.

Proof. We show that for every $t \in \{0, \dots, H{-}1\}$, one of the two conditions of Definition 2 holds for $\mathbf{Z}_t^H$ in the full-horizon manipulated graph $(\mathcal{G}_H)'_t$. We handle $t \geq k$ first, then $t < k$.

Setup and notation. Let $(\mathcal{G}_H)'_t$ denote the manipulated graph at timestep t: all incoming edges into future actions X_j ($j > t$) are removed and each X_j is given parents $\mathbf{Z}_j^H$. Let $((\mathcal{G}_H)'_t)_{X_t}$ further delete all outgoing edges from X_t. Define $\mathrm{Win}_t := \{t{-}k, \dots, t\}$. We use V_r (non-bold) for an individual variable at timestep r.

Step 1: Structure of the manipulated graph. By k-bounded influence (Assumption 1(i)), every edge in $\mathcal{G}_H$ spans at most k timesteps. In $(\mathcal{G}_H)'_t$, the original incoming edges to future actions are replaced by edges from $\mathbf{Z}_j^H$, which by Algorithm 1 contain only variables within k steps of j. Therefore the k-bounded influence property is preserved in $(\mathcal{G}_H)'_t$: every edge spans at most k steps. It follows that all parents and confounders of X_t lie in Win_t, and no edge connects any variable at timestep $\leq t - k - 1$ to any variable in Win_t.

We decompose $(\mathcal{G}_H)'_t$ into two parts:

- The local subgraph $\mathcal{L}_t$: the restriction of $(\mathcal{G}_H)'_t$ to variables at timesteps in Win_t.
- The tail subgraph $\mathcal{T}_t$: the restriction of $(\mathcal{G}_H)'_t$ to variables at timesteps in $\{t{+}1, \dots, H\}$, including Y.

By k-bounded influence, the only edges between $\mathcal{L}_t$ and $\mathcal{T}_t$ are edges from variables in $\{t{-}k{+}1, \dots, t\} \subset \mathrm{Win}_t$ to variables in $\{t{+}1, \dots, t{+}k\} \subset \mathcal{T}_t$. We call the variables in $\mathcal{T}_t$ at timesteps $\{t{+}1, \dots, t{+}k\}$ that have at least one parent in $\mathcal{L}_t$ the *interface variables*, and denote them $\mathbf{I}_t$. No other edges cross between $\mathcal{L}_t$ and $\mathcal{T}_t$. Additionally, since all incoming edges to actions $X_j \in \mathcal{T}_t$ have been replaced by edges from $\mathbf{Z}_j^H$, no bidirected edges cross the boundary either (by Assumption 1(i) bidirected edges in the original graph span at most k steps and those entering future actions are removed by the manipulated graph construction).

Step 2: Isomorphism between the local subgraph and the proxy graph. Consider the proxy graph $\mathcal{G}_k$ on timesteps $\{0, \dots, k\}$ with terminal reward Y_k. Define the time-shift $\phi_t(\tau) = \tau{+}(t{-}k)$. By time-homogeneity (Assumption 1(ii)), ϕ_t induces a graph isomorphism between $\mathcal{G}_k$ restricted to $\{0, \dots, k\}$ and $\mathcal{L}_t$, preserving all directed and bidirected edges and the $\mathbf{V}^O/\mathbf{V}^L$ partition.

This extends to manipulated graphs. In the proxy manipulated graph $(\mathcal{G}_k)'_k$, timestep k is the final step so there are no future actions to modify. In $\mathcal{L}_t$, the action X_t is likewise the last action (all actions at $j > t$ belong to $\mathcal{T}_t$). The adjustment sets satisfy $\mathbf{Z}^H_{\phi_t(\tau)} = \phi_t(\mathbf{Z}^k_\tau)$ by the identical relative-lag construction of Algorithm 1. Therefore $\mathcal{L}_t$ with the manipulated-graph modifications is isomorphic to $(\mathcal{G}_k)'_k$ under ϕ_t.

Step 3: Relating Y_k and Y via the interface. In $\mathcal{G}_k$, the terminal reward Y_k receives edges from variables at timesteps $\{k{-}k, \dots, k\} = \{0, \dots, k\}$ (by k-bounded influence). Since Y is a terminal sink node with only incoming edges in both $\mathcal{G}_k$ and $\mathcal{G}_H$, and its parent structure follows the same

time-homogeneous template, the structural role of Y_k relative to $\{0, \dots, k\}$ in $\mathcal{G}_k$ is identical to that of Y relative to $\{H-k, \dots, H\}$ in $\mathcal{G}_H$. Under the isomorphism, these correspond to variables in Win_t in $\mathcal{G}_H$.

In $(\mathcal{G}_H)'_t$, the terminal reward Y at timestep H is not solely adjacent to Win_t; it is connected to Win_t only through paths that traverse $\mathcal{T}_t$. However, by Step 1, *all* connections from $\mathcal{L}_t$ into $\mathcal{T}_t$ pass through the interface $\mathbf{I}_t$ at timesteps $\{t+1, \dots, t+k\}$. This means that $\mathbf{I}_t$ d-separates Win_t from Y in $\mathcal{T}_t$, in the following sense: every path from any variable in $\mathcal{L}_t$ to Y in $(\mathcal{G}_H)'_t$ must pass through $\mathbf{I}_t$.

The parents of Y_k in $\mathcal{G}_k$ (variables at timesteps $\{0, \dots, k\}$) are mapped by ϕ_t to variables in Win_t that are exactly the parents of the interface variables $\mathbf{I}_t$ within $\mathcal{L}_t$. Thus Y_k in $\mathcal{G}_k$ and Y in $\mathcal{G}_H$ are connected to $\mathcal{L}_t$ through the same set of nodes (up to isomorphism): Y_k directly, and Y through $\mathbf{I}_t$ and the tail.

Step 4: Transferring the sequential π-backdoor conditions. Since $\{\mathbf{Z}_t^k\}$ satisfies the sequential π-backdoor on $\mathcal{G}_k$, at timestep k either:

Case A: $X_k \notin \mathrm{An}_{(\mathcal{G}_k)'_k}(Y_k)$.

This means there is no directed path from X_k to Y_k in $(\mathcal{G}_k)'_k$. By the isomorphism of Step 2, there is no directed path from X_t to any interface variable in $\mathbf{I}_t$ within $\mathcal{L}_t$. Since every directed path from X_t to Y in $(\mathcal{G}_H)'_t$ must pass through $\mathbf{I}_t$ (Step 3), and no directed path from X_t reaches $\mathbf{I}_t$, we conclude $X_t \notin \mathrm{An}_{(\mathcal{G}_H)'_t}(Y)$.

Case B: $(X_k \perp\!\!\!\perp Y_k \mid \mathbf{Z}_k^k)_{((\mathcal{G}_k)'_k)_{X_k}}$.

We show $(X_t \perp\!\!\!\perp Y \mid \mathbf{Z}_t^H)_{((\mathcal{G}_H)'_t)_{X_t}}$. Let p be any path from X_t to Y in $((\mathcal{G}_H)'_t)_{X_t}$.

Since outgoing edges from X_t are deleted, p must leave X_t via an incoming edge from a parent or confounder, all of which lie in Win_t (Step 1). In the backward direction, p cannot reach timesteps $< t-k$ because no edges connect Win_t to earlier timesteps (Step 1). Therefore, the portion of p that is "behind" X_t (in the direction away from Y) is entirely contained in $\mathcal{L}_t$.

Since p connects X_t to Y and $Y \notin \mathcal{L}_t$, the path p must at some point cross from $\mathcal{L}_t$ into $\mathcal{T}_t$. By Step 1, this crossing must go through an interface variable $I \in \mathbf{I}_t$.

Now consider the path p as having two segments: the segment p_L from X_t to the first interface variable I (contained in $\mathcal{L}_t \cup \{I\}$), and the segment p_T from I onward to Y (contained in $\mathcal{T}_t$). We show p is blocked by $\mathbf{Z}_t^H$ by showing p_L is blocked.

Construct the corresponding path p'_L in $((\mathcal{G}_k)'_k)_{X_k}$: apply ϕ_t^{-1} to map p_L from $\mathcal{L}_t$ to $\mathcal{G}_k$. The interface variable I at timestep $t+j$ (for some $1 \le j \le k$) maps to a variable at timestep $k+j$ in $\mathcal{G}_k$. However, $\mathcal{G}_k$ only contains timesteps $\{0, \dots, k\}$, so if $j \ge 1$, the variable I maps to a timestep beyond $\mathcal{G}_k$.

We handle this as follows. The interface variable I has parents in $\mathcal{L}_t$ (by definition of $\mathbf{I}_t$). In the proxy graph, Y_k also has parents in $\{0, \dots, k\}$ corresponding (under ϕ_t) to the same variables in Win_t that are parents of $\mathbf{I}_t$. Therefore, any path from X_k that reaches a parent of Y_k in $((\mathcal{G}_k)'_k)_{X_k}$ corresponds to a path from X_t that reaches a parent of $\mathbf{I}_t$ in $((\mathcal{G}_H)'_t)_{X_t}$. In particular, the path p_L from X_t to I passes through a parent of I in $\mathcal{L}_t$, say V_q. The sub-path from X_t to V_q lies entirely in $\mathcal{L}_t$ and maps under ϕ_t^{-1} to a path from X_k to $\phi_t^{-1}(V_q)$ in $((\mathcal{G}_k)'_k)_{X_k}$. Since V_q is a parent of an interface variable, $\phi_t^{-1}(V_q)$ is a parent of Y_k in $\mathcal{G}_k$ (by the structural correspondence established in Step 3). Hence the path from X_k to $\phi_t^{-1}(V_q)$ can be extended to Y_k via the edge $\phi_t^{-1}(V_q) \to Y_k$, forming a path from X_k to Y_k in $((\mathcal{G}_k)'_k)_{X_k}$.

By the d-separation hypothesis, this path is blocked by $\mathbf{Z}_k^k$. The blocking must occur on the sub-path from X_k to $\phi_t^{-1}(V_q)$ (since Y_k is not in $\mathbf{Z}_k^k$ and the final edge $\phi_t^{-1}(V_q) \to Y_k$ cannot introduce a blocking collider). Mapping back via ϕ_t, the same blocking occurs on p_L at the corresponding variable in $\mathbf{Z}_t^H = \phi_t(\mathbf{Z}_k^k)$. Therefore p is blocked by $\mathbf{Z}_t^H$.

Boundary timesteps ($t < k$). For $t < k$, the window is $\{0, \dots, t\}$. The $\mathrm{before}(X_t)$ constraint in Algorithm 1 (Line 6) ensures $\mathbf{Z}_t^k$ contains only variables at timesteps $\leq t-1$. Since $t < k$, all such variables satisfy $\tau \geq 0 > t - k$, so the clipping in Lines 7–9 is vacuous: $\mathbf{Z}_t^H = \mathbf{Z}_t^k$. The subgraph of $\mathcal{G}_H$ restricted to $\{0, \dots, t\}$ is identical to the subgraph of $\mathcal{G}_k$ on the same timesteps (a direct subgraph inclusion requiring no time-shift). The manipulated graph $(\mathcal{G}_H)'_t$ restricted to $\{0, \dots, t\}$ matches $(\mathcal{G}_k)'_t$ restricted to $\{0, \dots, t\}$, since the modifications at timesteps $> t$ (replacement of future action parents) do not alter the subgraph at $\{0, \dots, t\}$. Therefore the d-separation or non-ancestry condition holding for X_t in $(\mathcal{G}_k)'_t$ also holds in $(\mathcal{G}_H)'_t$.

Since for every $t \in \{0, \dots, H-1\}$ one of the two conditions of Definition 2 holds, $\{\mathbf{Z}_t^H\}_{t=0}^{H-1}$ satisfies the sequential π-backdoor criterion for $(\mathcal{G}_H, \mathbf{X}, Y)$. □

C Algorithm Pseudocode

Algorithms 2 and 3 give the full training procedures for Causal SQIL and Causal IQ-Learn, respectively. Both algorithms share the same causal encoding layer (Algorithm 1) and SAC-style actor update, differing only in how the critic is trained. In Causal SQIL the critic minimizes a standard soft Bellman residual on binary-labeled transitions, whereas in Causal IQ-Learn it minimizes a chi-squared divergence objective on implicit rewards. All other components (entropy tuning, soft target updates, replay buffer management, and the causal adjustment algorithm) are identical.

Algorithm 2 Causal SQIL

Require: Expert demonstrations $\mathcal{D}_{\text{exp}}$, causal graph $\mathcal{G}$, window size k, horizon H, discount γ, soft update rate τ, batch size B.

Compute windowed adjustment sets $\{\mathbf{Z}_t^H\}_{t=0}^{H-1}$ and window specification S via Algorithm 1.
Build encoder $\textsc{Encode}(\mathbf{s}, t)$ from S: concatenates values of $V \in \mathbf{V}^O \cap \mathbf{Z}_t^H$ at lags $\{0, -1, \ldots, -k\}$ relative to t, zero-padding when $t < k$.
Initialize actor π_ϕ, twin Q-networks $Q_{\theta_1}, Q_{\theta_2}$, target networks $Q_{\bar{\theta}_1} \leftarrow Q_{\theta_1}$, $Q_{\bar{\theta}_2} \leftarrow Q_{\theta_2}$.
Initialize entropy coefficient $\log \alpha \leftarrow 0$, target entropy $\bar{\mathcal{H}} \leftarrow -|\mathbf{A}|$.
$\mathcal{B}_{\text{exp}} \leftarrow \varnothing$, $\mathcal{B}_\pi \leftarrow \varnothing$.
for each transition $(\mathbf{s}_t, \mathbf{x}_t, \mathbf{s}_{t+1}, d_t)$ in $\mathcal{D}_{\text{exp}}$ **do**
  $\mathcal{B}_{\text{exp}} \leftarrow \mathcal{B}_{\text{exp}} \cup \{(\textsc{Encode}(\mathbf{s}_t, t),\ \mathbf{x}_t,\ r{=}1,\ \textsc{Encode}(\mathbf{s}_{t+1}, t{+}1),\ d_t)\}$
end for
for each episode $e = 1, 2, \ldots$ **do**
  Reset environment, observe $\mathbf{s}_0$.
  for $t = 0, \ldots, H-1$ **do**
    $\mathbf{z}_t \leftarrow \textsc{Encode}(\mathbf{s}_t, t)$
    Sample $\mathbf{x}_t \sim \pi_\phi(\cdot \mid \mathbf{z}_t)$, execute, observe $\mathbf{s}_{t+1}$, d_t.
    $\mathcal{B}_\pi \leftarrow \mathcal{B}_\pi \cup \{(\mathbf{z}_t,\ \mathbf{x}_t,\ r{=}0,\ \textsc{Encode}(\mathbf{s}_{t+1}, t{+}1),\ d_t)\}$
  end for
  for each gradient step **do**
    Sample $\{(\mathbf{z}, \mathbf{x}, r, \mathbf{z}', d)\}_{i=1}^{B}$ with $B/2$ from $\mathcal{B}_{\text{exp}}$ and $B/2$ from $\mathcal{B}_\pi$.
    // Critic update
    $\mathbf{x}' \sim \pi_\phi(\cdot \mid \mathbf{z}')$
    $y \leftarrow r + \gamma(1-d)\big(\min_j Q_{\bar{\theta}_j}(\mathbf{z}', \mathbf{x}') - \alpha \log \pi_\phi(\mathbf{x}' \mid \mathbf{z}')\big)$
    $\mathcal{L}_Q \leftarrow \frac{1}{2}\sum_{j=1}^{2} \big\| Q_{\theta_j}(\mathbf{z}, \mathbf{x}) - y \big\|^2$
    Update θ_1, θ_2 by $\nabla_\theta \mathcal{L}_Q$.
    // Actor update
    $\tilde{\mathbf{x}} \sim \pi_\phi(\cdot \mid \mathbf{z})$ (reparameterized)
    $\mathcal{L}_\pi \leftarrow \mathbb{E}\big[\alpha \log \pi_\phi(\tilde{\mathbf{x}} \mid \mathbf{z}) - \min_j Q_{\theta_j}(\mathbf{z}, \tilde{\mathbf{x}})\big]$
    Update ϕ by $\nabla_\phi \mathcal{L}_\pi$.
    // Entropy tuning
    $\mathcal{L}_\alpha \leftarrow -\log \alpha\ \mathbb{E}\big[\log \pi_\phi(\tilde{\mathbf{x}} \mid \mathbf{z}) + \bar{\mathcal{H}}\big]$
    Update $\log \alpha$ by $\nabla \mathcal{L}_\alpha$, $\alpha \leftarrow \exp(\log \alpha)$.
    // Soft target update
    $\bar{\theta}_j \leftarrow \tau\, \theta_j + (1-\tau)\, \bar{\theta}_j$ for $j = 1, 2$.
  end for
end for
return π_ϕ.

Algorithm 3 Causal IQ-Learn

Require: Expert demonstrations $\mathcal{D}_{\text{exp}}$, causal graph $\mathcal{G}$, window size k, horizon H, discount γ, soft update rate τ, batch size B, number of value samples K.

Compute adjustment sets and build encoder ENCODE as in Algorithm 2, steps 1–2.
Initialize actor π_ϕ, twin Q-networks $Q_{\theta_1}, Q_{\theta_2}$, target networks $Q_{\bar{\theta}_1} \leftarrow Q_{\theta_1}$, $Q_{\bar{\theta}_2} \leftarrow Q_{\theta_2}$.
Initialize entropy coefficient $\log \alpha \leftarrow 0$, target entropy $\bar{\mathcal{H}} \leftarrow -|\mathbf{A}|$.
$\mathcal{B}_{\text{exp}} \leftarrow \varnothing, \quad \mathcal{B}_\pi \leftarrow \varnothing$.
Encode expert demonstrations into $\mathcal{B}_{\text{exp}}$ as $\{(\mathbf{z}_t, \mathbf{x}_t, \mathbf{z}_{t+1}, d_t)\}$ (no reward labels).
for each episode $e = 1, 2, \ldots$ **do**
  Roll out π_ϕ in the environment with causal encoding (as in Algorithm 2, steps 9–13). Store transitions in $\mathcal{B}_\pi$.
  for each gradient step **do**
    Sample $\{(\mathbf{z}^e, \mathbf{x}^e, \mathbf{z}'^e, d^e)\}_{i=1}^{B/2}$ from $\mathcal{B}_{\text{exp}}$ and $\{(\mathbf{z}^p, \mathbf{x}^p, \mathbf{z}'^p, d^p)\}_{i=1}^{B/2}$ from $\mathcal{B}_\pi$.
    Let $(\mathbf{z}, \mathbf{x}, \mathbf{z}', d)$ denote the concatenation of expert and policy batches.
    // Compute soft state value via Monte Carlo
    $\{\tilde{\mathbf{x}}_m\}_{m=1}^K \sim \pi_\phi(\cdot \mid \mathbf{z}')$
    $\bar{V}(\mathbf{z}') \leftarrow \frac{1}{K} \sum_{m=1}^K \left[\min_j Q_{\bar{\theta}_j}(\mathbf{z}', \tilde{\mathbf{x}}_m) - \alpha \log \pi_\phi(\tilde{\mathbf{x}}_m \mid \mathbf{z}') \right]$
    // Implicit reward
    $\hat{r}_j(\mathbf{z}, \mathbf{x}) \leftarrow Q_{\theta_j}(\mathbf{z}, \mathbf{x}) - \gamma(1 - d)\, \bar{V}(\mathbf{z}') \quad$ for $j = 1, 2$
    // Critic update (chi-squared divergence)
    $\mathcal{L}_{Q_j} \leftarrow -\mathbb{E}_{\text{exp}}\left[\hat{r}_j(\mathbf{z}^e, \mathbf{x}^e)\right] + \frac{1}{2}\, \mathbb{E}_{\text{all}}\left[\hat{r}_j(\mathbf{z}, \mathbf{x})^2\right] \quad$ for $j = 1, 2$
    Update θ_1, θ_2 by $\nabla_\theta \mathcal{L}_Q$.
    // Actor update (identical to Algorithm 2, steps 21–24)
    $\tilde{\mathbf{x}} \sim \pi_\phi(\cdot \mid \mathbf{z}), \quad \mathcal{L}_\pi \leftarrow \mathbb{E}\left[\alpha \log \pi_\phi(\tilde{\mathbf{x}} \mid \mathbf{z}) - \min_j Q_{\theta_j}(\mathbf{z}, \tilde{\mathbf{x}})\right]$
    Update ϕ by $\nabla_\phi \mathcal{L}_\pi$.
    // Entropy tuning and soft target update (identical to Algorithm 2, steps 26–29)
    Update $\log \alpha$, α, and $\bar{\theta}_1, \bar{\theta}_2$.
  end for
end for
return π_ϕ.

For completeness, we reproduce the FINDOX algorithm of Kumor et al. (2021) (Algorithm 1 in that paper), which returns the maximal set $\mathbf{O}_X \subseteq \mathbf{V}^O$ from which sequential π-backdoor admissible sets can be constructed. Given the causal diagram $\mathcal{G}$, action set $\mathbf{X}$, and target Y, FINDOX iteratively grows $\mathbf{O}_X$ by checking, for each observed node, whether a valid backdoor adjustment exists that would make that node a non-ancestor of Y in the manipulated graph. A sequential π-backdoor exists for $(\mathcal{G}, \mathbf{X}, Y)$ if and only if $\mathbf{X} \subseteq \mathbf{O}_X$ (Kumor et al., 2021, Theorem 3.1). Once $\mathbf{O}_X$ is obtained, the per-action adjustment sets $\mathbf{Z}_t$ are constructed from the Markov boundary of $\mathbf{O}_X$ in $\mathcal{G}^Y_{\mathbf{X}'}$ (where $\mathbf{X}' = \mathbf{O}_X \cap \mathbf{X}$), intersected with $\mathrm{before}(\mathbf{X}_t)$. Our windowed adjustment procedure (Algorithm 1) invokes FINDOX on a short-horizon proxy graph $\mathcal{G}_k$ rather than the full-horizon graph $\mathcal{G}_H$.

Algorithm 4 FINDOX (Kumor et al., 2021): Find largest valid $\mathbf{O}_X$ in ancestral graph of Y

Require: Causal diagram $\mathcal{G}$, action set $\mathbf{X}$, target Y.

function HASVALIDADJUSTMENT($\mathcal{G}, \mathbf{O}^X, O_i, X_i$)
  $C \leftarrow$ the c-component of O_i in $\mathcal{G}^Y$
  $\mathcal{G}_C \leftarrow$ the subgraph of $\mathcal{G}^Y$ containing only $\mathrm{Pa}^+(C)$ and intermediate latent variables
  $\mathbf{O}^C \leftarrow C \setminus (\mathbf{O}^X \cup \{O_i\})$ *// Elements of c-component that might be ancestors of Y in $\mathcal{G}'_i$*
  return $(O_i \perp\!\!\!\perp \mathbf{O}^C \mid \mathbf{O}^C \cap \mathrm{before}(X_i))$ in $\mathcal{G}_C$

function FINDOX($\mathcal{G}, \mathbf{X}, Y$)
  $\mathscr{O}^X \leftarrow$ empty map from elements of $\mathbf{V}^O$ to elements of $\mathbf{X}$
  repeat
    for $O_i \in \mathbf{V}^O$ of $\mathcal{G}^Y$ (ancestral graph of Y) in reverse temporal order **do**
      if $|\mathrm{ch}^+(O_i)| > 0$ and $\mathrm{ch}^+(O_i) \subseteq \mathrm{keys}(\mathscr{O}^X)$ **then**
        $X_i \leftarrow$ earliest element of $\mathscr{O}^X[\mathrm{ch}^+(O_i)]$ in temporal order
        if HASVALIDADJUSTMENT($\mathcal{G}, \mathrm{keys}(\mathscr{O}^X), O_i, X_i$) **then**
          $\mathscr{O}^X[O_i] \leftarrow X_i$
      else if $O_i \in \mathbf{X}$ and HASVALIDADJUSTMENT($\mathcal{G}, \mathrm{keys}(\mathscr{O}^X), O_i, O_i$) **then**
        $\mathscr{O}^X[O_i] \leftarrow O_i$
  while $|\mathscr{O}^X|$ changed in most recent pass
  return $\mathrm{keys}(\mathscr{O}^X)$

D Environment Details

We describe each confounded environment in detail, including the causal structure, the confounding mechanism, and the observation partition. Table 2 summarizes the key dimensions and setup.

Confounded AntMaze. The base task is goal-conditioned navigation in a windy maze using an 8-DoF ant robot as described in Example 1. The imitator does not observe $\mathbf{O}$, or torso orientation; to compensate, a 2D compass reading $\mathbf{W}$ is added providing a noisy surrogate for heading. The latent wind field $\mathbf{U}$ follows a piecewise-constant gust process and affects both the dynamics and the reward function. The compass is prone to distributional shift between expert and imitator environments due to changing influence from $\mathbf{U}$. Causally unaware methods that condition on $\mathbf{W}$ conflate the wind's influence on the compass with the ant's true heading, learning policies that turn into walls when the wind changes direction. Causal methods exclude $\mathbf{W}$ from their adjustment set and learn to navigate the maze using position, joint angles, and velocities alone, without explicit orientation information.

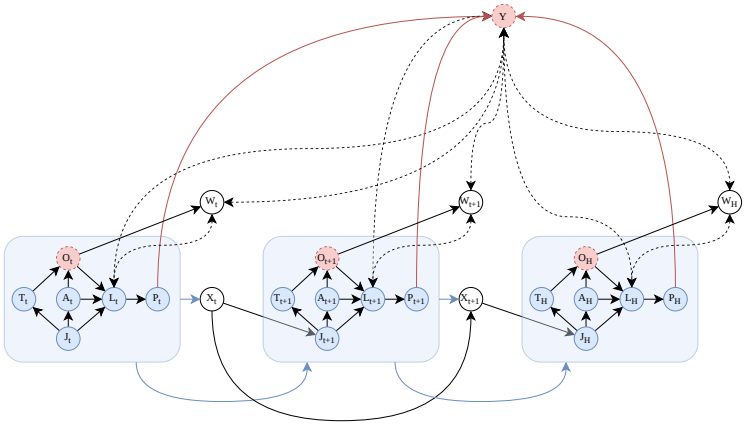

Figure 6: Confounded AntMaze. $\mathbf{P}$ is global position, $\mathbf{L}$ is torso linear velocity, $\mathbf{O}$ is torso orientation, $\mathbf{A}$ is joint angles, $\mathbf{T}$ is torso angular velocity, $\mathbf{J}$ is joint angular velocities, $\mathbf{U}$ is the latent wind field, $\mathbf{W}$ is the compass, $\mathbf{X}$ is joint torques, and $\mathbf{Y}$ is the latent terminal reward.

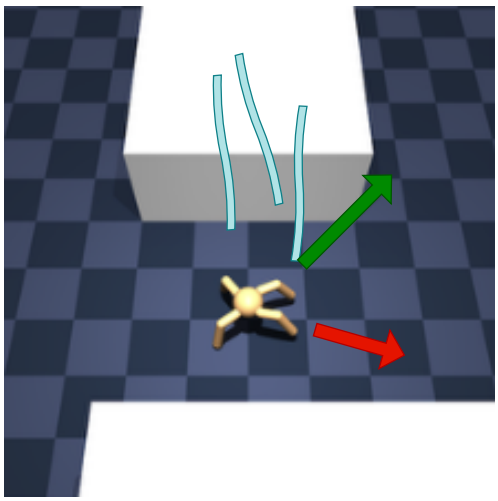

Figure 7: Visualization for Confounded AntMaze, demonstrating the effect of latent winds $\mathbf{U}$ (blue) on the heading compass $\mathbf{W}$ (red) despite the original heading $\mathbf{O}$ (green) being completely different.

Confounded HumanoidMaze. The base task is goal-conditioned navigation in a maze using a 21-DoF humanoid robot. The imitator does not observe $\mathbf{C}$, the 3D center-of-mass velocity; to compensate, a 2D ground-vibration reading $\mathbf{W}$ is added that provides a noisy surrogate for locomotion velocity. The latent seismic tremor $\mathbf{U}$ follows a piecewise-constant impulse process and affects both the dynamics (applying external force to the torso) and the reward function. The vibration sensor is prone to distributional shift between expert and imitator environments due to changing influence from $\mathbf{U}$. Causally unaware methods that condition on $\mathbf{W}$ conflate the tremor's influence on the vibration reading with the humanoid's true velocity, learning policies that stumble or over-correct when the tremor changes direction. Causal methods exclude $\mathbf{W}$ from their adjustment set and must learn to navigate the maze using position, joint angles, head height, extremity positions, torso orientation, and joint velocities alone, without explicit velocity information. The confounding effect of the seismic tremor is less disruptive than that of the wind field in AntMaze.

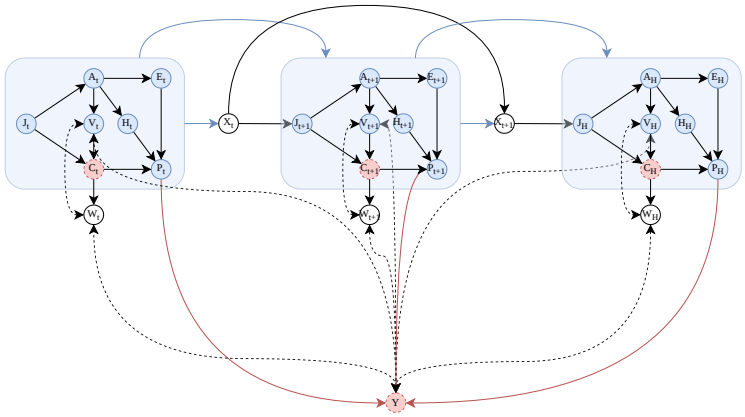

Figure 8: Confounded HumanoidMaze. $\mathbf{P}$ is global position, $\mathbf{A}$ is joint angles, $\mathbf{H}$ is head height, $\mathbf{E}$ is extremities in torso frame, $\mathbf{V}$ is torso vertical, $\mathbf{C}$ is center-of-mass gravity, $\mathbf{J}$ is joint velocities, $\mathbf{U}$ is the latent seismic tremor, $\mathbf{W}$ is the vibration sensor, $\mathbf{X}$ is joint torques, and $\mathbf{Y}$ is the latent reward.

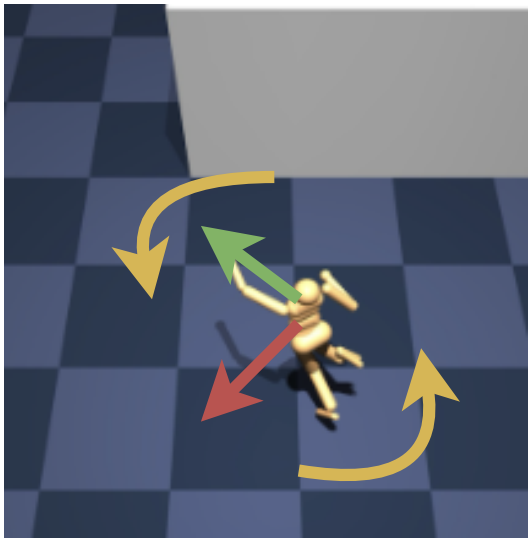

Figure 9: Visualization for Confounded HumanoidMaze, demonstrating the effect of latent seismic tremors $\mathbf{U}$ (yellow) on the perceived ground vibration sensor $\mathbf{W}$ (red) despite the true center-of-mass velocity $\mathbf{C}$ (green) pointing in a completely different direction.

Environment	H	$\lvert\mathbf{V}\rvert$	$\lvert\mathbf{X}\rvert$	Latent Confounder	Collider
AntMaze	1000	35	8	Wind field	Compass
HumanoidMaze	2000	89	21	Seismic tremors	Vibration

Table 2: Summary of confounded environments. $\lvert\mathbf{V}\rvert$ denotes the imitator's observation dimensionality (excluding hidden variables). $\lvert\mathbf{X}\rvert$ is the action dimensionality.

E Hyperparameters and Training Details

Table 3 summarizes the hyperparameters for all algorithms. Causal and non-causal variants of each algorithm use identical hyperparameters and network architectures; the only difference is the input representation (causally-adjusted $\mathbf{Z}_t$ vs. full observation $\mathbf{V}_t^O$). Where hyperparameters differ between environment families we write $\mathcal{A}$ (AntMaze) and $\mathcal{H}$ (HumanoidMaze). Note: due to limited computational resources, HumanoidMaze-Large evaluations consisted of 100 episodes rather than the other tasks' 1000 episodes, and non-causal SQIL for HumanoidMaze-Large was trained on only $1M$ timesteps instead of $2M$; this decision was informed by the fact that every algorithm in every task reached its best checkpoint before the $1M$ timestep mark during training.

Table 3: Hyperparameters for all algorithms. Causal and non-causal variants share identical settings. "—" denotes a hyperparameter not applicable to that algorithm. $\mathcal{A}$ / $\mathcal{H}$ distinguishes AntMaze and HumanoidMaze values where they differ.

Hyperparameter	BC	GAIL	SQIL	IQ-Learn
Network Architecture				
Hidden dimension	256	256	256	256
Actor residual blocks	4	3	3	3
Actor dropout	0.0	0.05	0.05	0.05
Layer normalization	✓	✓	✓	✓
Activation	SiLU	SiLU	SiLU	SiLU
Output squashing	Tanh	Tanh	Tanh	Tanh
Optimization				
Optimizer	Adam	Adam	Adam	Adam
Actor learning rate	3×10^{-4}	1×10^{-4}	3×10^{-4}	3×10^{-4}
Critic learning rate	—	3×10^{-4}	3×10^{-4}	3×10^{-4}
Batch size	2048	1024	256	256
Discount γ	—	0.99	0.99	0.99
Max gradient norm	—	0.5	1.0	1.0
BC-Specific				
Loss function	Huber	—	—	—
Training epochs	$\mathcal{A}$: 100 / $\mathcal{H}$-Med: 100 / $\mathcal{H}$-Large: 200	—	—	—
Early stopping patience	$\mathcal{A}$: 15 / $\mathcal{H}$-Med: 30 / $\mathcal{H}$-Large: 15	—	—	—
Validation fraction	0.2	—	—	—
GAIL-Specific (PPO + Discriminator)				
PPO clip ratio ϵ	—	0.2	—	—
GAE λ	—	0.95	—	—
PPO epochs per round	—	$\mathcal{A}$: 4 / $\mathcal{H}$: 2	—	—
PPO entropy coefficient	—	10^{-2}	—	—
Value loss coefficient	—	0.5	—	—
Normalize advantages	—	✓	—	—
Discriminator learning rate	—	3×10^{-4}	—	—
Discriminator dropout	—	0.2	—	—
Discriminator updates per round	—	2	—	—
Discriminator minibatch size	—	1024	—	—
Gradient penalty λ_{GP}	—	5.0	—	—
Episodes per round	—	$\mathcal{A}$: 20 / $\mathcal{H}$: 10	—	—
Total training rounds	—	$\mathcal{A}$: 500 / $\mathcal{H}$: 200	—	—
Disc. LR scheduler	—	StepLR(100, 0.5)	—	—
SQIL / IQ-Learn (SAC-Based)				
Soft update rate τ	—	—	0.005	0.005
Entropy coefficient α	—	—	auto-tuned	auto-tuned
Entropy learning rate	—	—	3×10^{-4}	3×10^{-4}
Target entropy $\mathcal{H}$	—	—	$-\lvert\mathbf{A}\rvert$	$-\lvert\mathbf{A}\rvert$
α clamp range	—	—	$[e^{-\log 1000}, e^{-\log 10}]$	$[e^{-\log 1000}, e^{-\log 10}]$
Replay buffer capacity	—	—	10^6	10^6
Expert buffer ratio	—	—	0.5	0.5
Random exploration steps	—	—	5000	5000
Update-to-data ratio	—	—	$\mathcal{A}$: 0.25 / $\mathcal{H}$: 0.5	$\mathcal{A}$: 0.25 / $\mathcal{H}$: 0.5
Total timesteps	—	—	2×10^6	2×10^6
V estimation	—	—	—	$\mathcal{A}$: MC (K=16) / $\mathcal{H}$: single-sample
Critic LR scheduler	—	—	CosineAnnealing	CosineAnnealing
Shared				
Lookback window k	$\mathcal{A}$: 10 / $\mathcal{H}$: 2	$\mathcal{A}$: 10 / $\mathcal{H}$: 2	$\mathcal{A}$: 10 / $\mathcal{H}$: 2	$\mathcal{A}$: 10 / $\mathcal{H}$: 2
Max episode steps	$\mathcal{A}$: 1000 / $\mathcal{H}$: 2000	$\mathcal{A}$: 1000 / $\mathcal{H}$: 2000	$\mathcal{A}$: 1000 / $\mathcal{H}$: 2000	$\mathcal{A}$: 1000 / $\mathcal{H}$: 2000

Expert construction. Expert policies for the maze tasks are constructed using offline-to-online RL. For each environment, we begin by training a goal-conditioned behavioral cloning (BC) policy on provided demonstrations from the base OGBench dataset. This policy provides an initialization that captures the global structure of the task, but it is not yet ready for the confounders introduced in the modified environment. To obtain an expert that reflects performance under the confounded dynamics, we then fine-tune this BC policy through a period of off-policy actor–critic training using TD3 (Fujimoto et al., 2018). Reward shaping is added optionally during fine-tuning to compensate for the sparse reward signals. The resulting expert policy is capable of operating effectively under the latent disturbances, partial observability, and altered transition dynamics of the confounded environment.

Network architecture. All algorithms share a residual MLP backbone for the actor. The actor network maps the causally-adjusted encoding $\mathbf{z}_t$ to actions through a linear projection into hidden dimension 256, followed by residual blocks, and a final linear output with tanh squashing to the action bounds. Each residual block consists of LayerNorm, SiLU, a linear layer, a second LayerNorm, SiLU, dropout, and a second linear layer, with a skip connection from input to output. For GAIL, SQIL, and IQ-Learn, the actor outputs the mean of a squashed Gaussian with a state-independent learnable log-variance; for BC, the actor outputs a deterministic action. The Q-networks (SQIL and IQ-Learn) use the same residual MLP architecture, taking concatenated $(\mathbf{z}_t, \mathbf{x}_t)$ as input and producing a scalar output. GAIL's value network (critic) uses a simpler three-layer MLP with ReLU activations, taking $\mathbf{z}_t$ as input. GAIL's discriminator similarly uses a three-layer MLP with ReLU activations, skip connections between layers, higher dropout (0.2), and binary cross-entropy loss to classify $(\mathbf{z}_t, \mathbf{x}_t)$ pairs as expert or policy-generated.

SQIL training details. Causal SQIL uses SAC with twin Q-networks and soft target updates ($\tau = 0.005$). The entropy coefficient α is automatically tuned toward a target entropy of $-|\mathbf{X}|$ and clamped to $[\exp(-\log 1000),\ \exp(-\log 10)] \approx [0.001, 0.1]$. The replay buffer is split equally: 50% capacity for expert transitions (labeled $r = 1$) and 50% for policy transitions ($r = 0$), with each training batch sampled $50/50$ from both halves. The first $5{,}000$ timesteps use random actions for exploration before policy rollouts begin. Training runs for 2×10^6 environment steps. The update-to-data (UTD) ratio is 0.25 for AntMaze tasks (one gradient step per four environment steps) and 0.5 for HumanoidMaze tasks. A cosine annealing schedule is applied to the critic learning rate.

IQ-Learn training details. Causal IQ-Learn shares the SAC actor update and twin Q-network architecture with Causal SQIL. The key difference is the critic loss: instead of minimizing a soft Bellman residual on binary-labeled transitions, IQ-Learn minimizes the chi-squared divergence between the implicit reward distribution under the expert and the combined (expert + policy) data. For AntMaze tasks, the soft state value $V(\mathbf{z}')$ is estimated via $K = 16$ Monte Carlo samples from the current policy; for HumanoidMaze tasks, a single-sample estimate is used (matching the standard SAC target computation), which avoids the high variance of multi-sample estimates in the higher-dimensional action space. The entropy coefficient α is auto-tuned and clamped to $[\exp(-\log 1000),\ \exp(-\log 10)] \approx [0.001, 0.1]$ to prevent entropy collapse or explosion. IQ-Learn uses the same UTD ratios as SQIL (0.25 for AntMaze, 0.5 for HumanoidMaze).

GAIL training details. Causal GAIL alternates between on-policy rollouts and discriminator–policy updates. Each round collects 20 episodes for AntMaze (10 for HumanoidMaze), up to H steps each, using the current policy, computes advantages via GAE ($\lambda = 0.95$, $\gamma = 0.99$), and performs PPO updates with clipping ratio $\epsilon = 0.2$, entropy regularization coefficient 10^{-2}, and value loss coefficient 0.5. The number of PPO epochs per round is 4 for AntMaze and 2 for HumanoidMaze. The discriminator is updated 2 times per round on minibatches of 1024 with gradient penalty ($\lambda_{\mathrm{GP}} = 5.0$). The discriminator learning rate follows a StepLR schedule, halving every 100 rounds. Training runs for 500 rounds on AntMaze and 200 rounds on HumanoidMaze.

BC training details. Causal BC performs supervised learning on expert demonstrations using Huber loss, optimized with Adam at learning rate 3×10^{-4}. The actor uses 4 residual blocks with no dropout, trained for up to 100 epochs (200 for HumanoidMaze-Large) with early stopping on a held-out 20% validation split. The early stopping patience is 15 for all tasks except HumanoidMaze-Medium, which uses a patience of 30. Training uses large batches of 2048 to reduce variance in the gradient estimates. BC is the only algorithm that does not interact with the environment during training; it learns entirely from the static expert dataset.

F Additional Results

F.1 Raw Evaluation Data

Table 4: Evaluation results on confounded tasks without normalization. Best non-expert result per task in **bold**.

		Expert	C-BC	C-GAIL	C-SQIL	C-IQL	BC	GAIL	SQIL	IQL
Ant-Med	$\mathbb{E}[Y]$	-104.0 ± 87.5	-125.2 ± 100.9	-135.8 ± 113.6	$\mathbf{-99.6 \pm 85.7}$	-118.2 ± 101.3	-375.3 ± 119.2	-364.9 ± 112.2	-370.7 ± 119.1	-372.5 ± 120.8
	SR (%)	87.6 ± 1.0	77.9 ± 1.3	74.1 ± 1.4	$\mathbf{90.7 \pm 0.9}$	84.3 ± 1.15	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Ant-Large	$\mathbb{E}[Y]$	-255.2 ± 122.8	-284.9 ± 121.6	-328.5 ± 122.6	-279.9 ± 125.9	$\mathbf{-258.6 \pm 127.9}$	-470.2 ± 150.2	-398.4 ± 132.9	-484.2 ± 151.5	-468.1 ± 153.5
	SR (%)	55.9 ± 1.6	39.8 ± 1.5	7.3 ± 0.8	45.0 ± 1.6	$\mathbf{58.9 \pm 1.6}$	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Hum-Med	$\mathbb{E}[Y]$	-522.9 ± 218.4	-655.7 ± 172.7	-747.2 ± 161.9	$\mathbf{-555.5 \pm 217.2}$	-588.8 ± 203.8	-678.4 ± 169.1	-747.7 ± 162.5	-706.2 ± 157.0	-697.6 ± 169.7
	SR (%)	33.8 ± 1.5	10.4 ± 1.0	0.0 ± 0.0	$\mathbf{24.7 \pm 1.4}$	19.1 ± 1.2	5.4 ± 0.7	0.0 ± 0.0	0.1 ± 0.1	2.4 ± 0.5
Hum-Large	$\mathbb{E}[Y]$	-840.6 ± 212.9	-858.6 ± 227.9	-976.3 ± 210.2	$\mathbf{-837.7 \pm 226.0}$	-887.1 ± 220.5	-884.1 ± 204.6	-977.0 ± 208.9	-896.3 ± 194.9	-906.2 ± 193.3
	SR (%)	7.0 ± 0.8	$\mathbf{8.0 \pm 2.7}$	0.0 ± 0.0	$\mathbf{8.0 \pm 2.7}$	3.0 ± 1.7	5.0 ± 2.2	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0

Remark on expert success rates. We note that the raw expert success rates in Table 4 are substantially lower than those typically reported in standard MuJoCo benchmarks. This is by design: our confounded environments are partially observable even to the expert, as the exogenous disturbances $\mathbf{U}$ (represented by bidirected arrows in the causal diagrams) are unobserved by all agents, including the expert. The unpredictable external forces generated by $\mathbf{U}$ make consistent task completion inherently difficult regardless of the agent's sensory access or learning algorithm. In this sense, the environments are designed to reflect realistic deployment conditions in which no agent has full knowledge of the environment dynamics, rather than the fully observed settings common in standard benchmarks. Nevertheless, the expert has strictly more information than the imitator (observing $\mathbf{V}^O \cup \mathbf{V}^L$ rather than $\mathbf{V}^O$ alone) and serves as a meaningful upper bound on what can be achieved given the imitator's observation set. The normalized metrics in Table 2 should therefore be interpreted relative to this upper bound: an algorithm recovering, say, 80% of the expert's success rate is operating near the frontier of what is achievable given the imitator's partial observability, not at 80% of a trivially solvable task. Cases in which a causal method exceeds 100% normalized performance (e.g., Causal SQIL on AntMaze-Medium, Causal IQ-Learn on AntMaze-Large) indicate that the imitator's learned policy is slightly more robust to the stochastic disturbances than the expert policy obtained via offline-to-online RL, likely because the off-policy Q-learning objective implicitly averages over the disturbance distribution encountered during training rather than committing to the point estimates used during expert fine-tuning.

F.2 Episode Lengths

We report the mean episode length for successfully solved episodes in Table 5. Shorter episodes indicate more efficient navigation. Only algorithms with nonzero success rates are included.

Table 5: Mean episode length (steps) for successfully solved episodes ($\pm$ std). Only algorithms with nonzero success rates are shown. Best result per task in **bold**.

Environment	Expert	C-BC	C-GAIL	C-SQIL	C-IQL	BC	GAIL	SQIL	IQL
AntMaze-Med	301 ± 57	318 ± 79	309 ± 75	$\mathbf{299 \pm 62}$	321 ± 90	—	—	—	—
AntMaze-Large	648 ± 99	683 ± 116	709 ± 134	695 ± 119	$\mathbf{636 \pm 108}$	—	—	—	—
HumanoidMaze-Med	976 ± 410	1356 ± 344	—	$\mathbf{912 \pm 425}$	1024 ± 443	1364 ± 320	—	1800 ± 0	1085 ± 396
HumanoidMaze-Large	1463 ± 366	1477 ± 421	—	$\mathbf{1251 \pm 316}$	1848 ± 84	1663 ± 237	—	—	—

G Limitations

Known causal graph. Our framework assumes that the causal diagram $\mathcal{G}$ is specified *a priori*. In practice, the graph must be elicited from domain expertise or estimated from data. Misspecification of $\mathcal{G}$, such as missing an edge from a confounder to the action, can lead to invalid adjustment sets and biased policies. Integrating causal discovery from observational data, either as a preprocessing step or jointly with policy learning, remains an important open problem.

Bounded temporal influence of confounders. The windowed approximation in Algorithm 1 assumes that the influence of any single confounder realization decays within k timesteps. This is justified by the MuJoCo physics of the environments we consider, but environments with long-range latent dependencies (e.g. persistent hidden goals, slowly drifting dynamics, or latent agent intentions in multi-agent settings) would violate this assumption. In such cases, alternative approximation strategies (e.g., hierarchical windows or attention-based aggregation over history) or exact methods on compressed representations would be required.

No exploitation of reward structure. Our algorithms treat the reward as entirely latent and make no parametric assumptions about its form. Prior work on partial identification for CIL (Ruan et al., 2024) has shown that incorporating reward priors can tighten bounds on the imitating policy and even enable the imitator to surpass expert performance. Combining Causal IQ-Learn with such reward priors is a natural extension that we leave to future work.

Sensitivity to expert optimality. While Causal SQIL and Causal IQ-Learn successfully address the compounding error and credit assignment issues seen in Causal BC and Causal GAIL, they are more sensitive than BC to the quality and optimality of expert demonstrations. Because these off-policy algorithms utilize demonstrations to ground an implicit reward signal or a value function via soft Q-learning objectives, they are more susceptible to noise in high-dimensional state-action spaces where the expert signal may be weak. This dependency is clearly demonstrated in the HumanoidMaze-Large task (Table 1), where the expert itself is sub-optimal, achieving a success rate of only 7.0%. While Causal BC maintains a success rate of 8.0%, Causal IQ-Learn collapses to 3.0% and Causal SQIL matches at 8.0% despite both algorithms outperforming BC in other tasks where the expert is more consistent. These results suggest that when the expert is sub-optimal, the off-policy grounding of value functions is more easily corrupted than the supervised cloning objective, indicating that the scalability benefits of our proposed methods are most reliably realized when provided with a high-quality expert signal.