CAUSAL IMITATION LEARNING VIA INVERSE REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

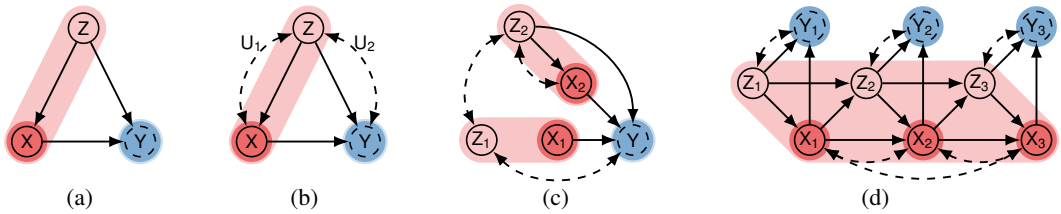
One of the most common ways children learn when unfamiliar with the environment is by mimicking adults. Imitation learning concerns an imitator learning to behave in an unknown environment from an expert's demonstration; reward signals remain latent to the imitator. This paper studies imitation learning through causal lenses and extends the analysis and tools developed for behavior cloning (Zhang, Kumor, Bareinboim, 2020) to inverse reinforcement learning. First, we propose novel graphical conditions that allow the imitator to learn a policy performing as well as the expert's behavior policy, even when the imitator and the expert's state-action space disagree, and unobserved confounders (UCs) are present. When provided with parametric knowledge about the unknown reward function, such a policy may outperform the expert's. Also, our method is easily extensible and allows one to leverage existing IRL algorithms even when UCs are present, including the multiplicative-weights algorithm (MWAL) (Syed & Schapire, 2008) and the generative adversarial imitation learning (GAIL) (Ho & Ermon, 2016). Finally, we validate our framework by simulations using real-world and synthetic data.

1 INTRODUCTION

Reinforcement Learning (RL) has been deployed and shown to perform extremely well in highly complex environments in the past decades (Sutton & Barto, 1998; Mnih et al., 2013; Silver et al., 2016; Berner et al., 2019). One of the critical assumptions behind many of the classical RL algorithms is that the reward signal is fully observed, and the reward function could be well-specified. In many real-world applications, however, it might be impractical to design a suitable reward function that evaluates each and every scenario (Randløv & Alstrøm, 1998; Ng et al., 1999). For example, in the context of human driving, it is challenging to design a precise reward function, and experimenting in the environment could be ill-advised; still, watching expert drivers operating is usually feasible.

In machine learning, the *imitation learning* paradigm investigates the problem of how an agent should behave and learn in an environment with an unknown reward function by observing demonstrations from a human expert (Argall et al., 2009; Billard et al., 2008; Hussein et al., 2017; Osa et al., 2018). There are two major learning modalities that implements IL – *behavioral cloning* (BC) (Widrow, 1964; Pomerleau, 1989; Muller et al., 2006; Mülling et al., 2013; Mahler & Goldberg, 2017) and *inverse reinforcement learning* (IRL) Ng et al. (2000); Ziebart et al. (2008); Ho & Ermon (2016); Fu et al. (2017). BC methods directly mimic the expert's behavior policy by learning a mapping from observed states to the expert's action via supervised learning. Alternatively, IRL methods first learn a potential reward function under which the expert's behavior policy is optimal. The imitator then obtains a policy by employing standard RL methods to maximize the learned reward function. Under some common assumptions, both BC and IRL are able to obtain policies that achieve the expert's performance (Kumor et al., 2021; Swamy et al., 2021). Moreover, when additional parametric knowledge about the reward function is provided, IRL may produce a policy that outperforms the expert's in the underlying environment (Syed & Schapire, 2008; Li et al., 2017; Yu et al., 2020).

For concreteness, consider a learning scenario depicted in Fig. 1a, describing trajectories of human-driven cars collected by drones flying over highways (Krajewski et al., 2018; Etesami & Geiger, 2020). Using such data, we want to learn a policy $X \leftarrow \pi(Z)$ deciding on the acceleration (action) $X \in \{0, 1\}$ of the demonstrator car based on velocities and locations Z of surrounding cars. The driving performance is measured by a latent reward signal Y. Consider an instance where $Y \leftarrow (1 - X)Z +$



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

Figure 1: Causal diagrams where X represents an action (shaded red) and Y represents a latent reward (shaded blue). Input covariates of the policy scope $\mathcal{S}$ are shaded in light red.

$X(1-Z)$ and values of Z are drawn uniformly over $\{0, 1\}$. A human expert generates demonstrations following a behavior policy such that $P(X = 1 \mid Z = 0) = 0.6$ and $P(X = 0 \mid Z = 1) = 0.4$. Evaluating the expert's performance gives $\mathbb{E}[Y] = P(X = 1, Z = 0) + P(X = 0, Z = 1) = 0.5$. Now we apply standard IRL algorithms to learn a policy $X \leftarrow \pi(Z)$ so that the imitator's driving performance, denoted by $\mathbb{E}[Y \mid \text{do}(\pi)]$, is at least as good as the expert's performance $\mathbb{E}[Y]$. Detailed derivations of IRL policy are shown in Appendix A. Note that $\mathbb{E}[Y|z, x] = x + z - 2xz$ belongs to a family of reward functions $f_Y(x, z) = \alpha x + \beta z - \gamma xz$, where $0 < \alpha < \gamma$. A typical IRL imitator solves a minimax problem $\min_\pi \max_{f_Y} \mathbb{E}\left[f_Y(X, Z)\right] - \mathbb{E}\left[f_Y(X, Z) \mid \text{do}(\pi)\right]$. The inner step "guesses" a reward function being optimized by the expert; while the outer step learns a policy maximizing the learned reward function. Applying these steps leads to a policy $\pi^* : X \leftarrow \neg Z$ with the expected reward $\mathbb{E}[Y \mid \text{do}(\pi^*)] = 1$, which outperforms the sub-optimal expert.

Despite the performance guarantees provided by existing imitation methods, both BC and IRL rely on the assumption that the expert's input observations match those available to the imitator. More recently, there exists an emerging line of research under the rubric of *causal imitation learning* that augments the imitation paradigm to account for environments consisting of arbitrary causal mechanisms and the aforementioned mismatch between expert and imitator's sensory capabilities (de Haan et al., 2019; Zhang et al., 2020; Etesami & Geiger, 2020; Kumor et al., 2021). Closest to our work, Zhang et al. (2020); Kumor et al. (2021) derived graphical criteria that completely characterize when and how BC could lead to successful imitation even when the agents perceive reality differently. Still, it is unclear how to perform IRL-type training if some expert's observed states remain latent to the imitator, which leads to the presence of unobserved confounding (UCs) in expert's demonstrations. Perhaps surprisingly, naively applying IRL methods when UCs are present does not necessarily lead to satisfactory performance, even when the expert itself behaves optimally.

To witness, we now modify the previous highway driving scenario to demonstrate the challenges of UCs. In reality, covariates Z (i.e., velocities and location) are also affected by the car horn U_1 of surrounding vehicles and the wind condition U_2. However, due to the different perspectives of drones (recording from the top), such critical information (i.e, U_1, U_2) is not recorded by the camera and thus remains unobserved. Fig. 1b graphically describes this modified learning setting. More specifically, consider an instance where $Z \leftarrow U_1 \oplus U_2$, $Y \leftarrow \neg X \oplus Z \oplus U_2$; $\oplus$ is the *exclusive-or* operator; and values of U_1 and U_2 are drawn uniformly over $\{0, 1\}$. An expert driver, being able to hear the car horn U_1, follows a behavior policy $X \leftarrow U_1$ and achieves the optimal performance $\mathbb{E}[Y] = 1$. Meanwhile, observe that $\mathbb{E}[Y|z, x] = 1$ belongs to a family of reward functions $f_Y(x, z) = \alpha$ (where $\alpha > 0$). Solving $\min_\pi \max_{f_Y} \mathbb{E}\left[f_Y(X, Z)\right] - \mathbb{E}\left[f_Y(X, Z) \mid \text{do}(\pi)\right]$ leads to an IRL policy π^* with expected reward $\mathbb{E}[Y|\text{do}(\pi^*)] = 0.5$, which is far from the expert's optimal performance $\mathbb{E}[Y] = 1$.

After all, a question that naturally arises is, under what conditions an IRL imitator procedure can perform well when UCs are present, and there is a mismatch between the perception of the two agents? In this paper, we answer this question and, more broadly, investigate the challenge of performing IRL through causal lenses. In particular, our contributions are summarized as follows. (1) We provide a novel, causal formulation of the inverse reinforcement learning problem. This formulation allows one to formally study and understand the conditions under which an IRL policy is learnable, including in settings where UCs cannot be ruled out a priori. (2) We derive a new graphical condition for deciding whether an imitating policy can be computed from the available data and knowledge, which provides a robust generalization of current IRL algorithms to non-Markovian settings, including GAIL (Ho & Ermon, 2016) and MWAL (Syed & Schapire, 2008). (3) Finally, we move beyond this graphical condition and develop an effective IRL algorithm for structural causal models (Pearl, 2000) with arbitrary causal relationships. Due to the space constraints, all proofs are provided in Appendix B. For a more detailed survey on imitation learning and causal inference, we refer readers to Appendix E.

1.1 PRELIMINARIES

We use capital letters to denote random variables (X) and small letters for their values (x). $\mathscr{D}_X$ represents the domain of X and $\mathscr{P}_X$ the space of probability distributions over $\mathscr{D}_X$. For a set $\boldsymbol{X}$, let $|\boldsymbol{X}|$ denote its dimension. The probability distribution over variables $\boldsymbol{X}$ is denoted by $P(\boldsymbol{X})$. Similarly, $P(\boldsymbol{Y} \mid \boldsymbol{X})$ represents a set of conditional distributions $P(\boldsymbol{Y} \mid \boldsymbol{X} = \boldsymbol{x})$ for all realizations $\boldsymbol{x}$. We use abbreviations $P(\boldsymbol{x})$ for probabilities $P(\boldsymbol{X} = \boldsymbol{x})$; so does $P(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x}) = P(\boldsymbol{y} \mid \boldsymbol{x})$. Finally, indicator function $\mathbb{1}\{\boldsymbol{Z} = \boldsymbol{z}\}$ returns 1 if $\boldsymbol{Z} = \boldsymbol{z}$ holds true; otherwise 0.

The basic semantic framework of our analysis rests on *structural causal models* (SCMs) (Pearl, 2000, Ch. 7). An SCM M is a tuple $\langle \boldsymbol{U}, \boldsymbol{V}, \mathcal{F}, P(\boldsymbol{U}) \rangle$ with $\boldsymbol{V}$ the set of endogenous, and $\boldsymbol{U}$ exogenous variables. $\mathcal{F}$ is a set of structural functions s.t. for $f_V \in \mathcal{F}$, $V \leftarrow f_V(\boldsymbol{pa}_V, \boldsymbol{u}_V)$, with $\boldsymbol{PA}_V \subseteq \boldsymbol{V}, \boldsymbol{U}_V \subseteq \boldsymbol{U}$. Values of $\boldsymbol{U}$ are drawn from an exogenous distribution $P(\boldsymbol{U})$, inducing distribution $P(\boldsymbol{V})$ over endogenous variables $\boldsymbol{V}$. Since the learner can observe only a subset of endogenous variables, we split $\boldsymbol{V}$ into a partition $\boldsymbol{O} \cup \boldsymbol{L}$ where variable $\boldsymbol{O} \subseteq \boldsymbol{V}$ are observed and $\boldsymbol{L} = \boldsymbol{V} \setminus \boldsymbol{O}$ remain latent to the leaner. The marginal distribution $P(\boldsymbol{O})$ is thus referred to as the *observational distribution*. An *atomic intervention* on a subset $\boldsymbol{X} \subseteq \boldsymbol{V}$, denoted by $\text{do}(\boldsymbol{x})$, is an operation where values of $\boldsymbol{X}$ are set to constants $\boldsymbol{x}$, replacing the functions $f_{\boldsymbol{X}} = \{f_X : \forall X \in \boldsymbol{X}\}$ that would normally determine their values. For an SCM M, let $M_{\boldsymbol{x}}$ be a submodel of M induced by intervention $\text{do}(\boldsymbol{x})$. For a set $\boldsymbol{Y} \subseteq \boldsymbol{V}$, the interventional distribution $P(\boldsymbol{s}|\text{do}(\boldsymbol{x}))$ induced by $\text{do}(\boldsymbol{x})$ is defined as the distribution over $\boldsymbol{Y}$ in the submodel $M_{\boldsymbol{x}}$, i.e., $P_M(\boldsymbol{Y}|\text{do}(\boldsymbol{x})) \triangleq P_{M_{\boldsymbol{x}}}(\boldsymbol{Y})$. We leave M implicit when it is obvious from the context.

Each SCM M is associated with a causal diagram $\mathcal{G}$ which is a directed acyclic graph where (e.g., see Fig. 1) solid nodes represent observed variables $\boldsymbol{O}$, dashed nodes represent latent variables $\boldsymbol{L}$, and arrows represent the arguments $\boldsymbol{PA}_V$ of each function $f_V \in \mathcal{F}$. Exogenous variables $\boldsymbol{U}$ are not explicitly shown; a bi-directed arrow between nodes V_i and V_j indicates the presence of an unobserved confounder (UC) affecting both V_i and V_j. We will use family abbreviations to represent graphical relationships such as parents, children, descendants, and ancestors. For example, the set of parent nodes of $\boldsymbol{X}$ in $\mathcal{G}$ is denoted by $pa(\boldsymbol{X})_{\mathcal{G}} = \cup_{X \in \boldsymbol{X}} pa(X)_{\mathcal{G}}$; ch, de and an are similarly defined. Capitalized versions Pa, Ch, De, An include the argument as well, e.g. $Pa(\boldsymbol{X})_{\mathcal{G}} = pa(\boldsymbol{X})_{\mathcal{G}} \cup \boldsymbol{X}$. For a subset $\boldsymbol{X} \subseteq \boldsymbol{V}$, the subgraph obtained from $\mathcal{G}$ with edges outgoing from $\boldsymbol{X}$ / incoming into $\boldsymbol{X}$ removed is written as $\mathcal{G}_{\underline{\boldsymbol{X}}}$/$\mathcal{G}_{\overline{\boldsymbol{X}}}$ respectively. $\mathcal{G}_{[\boldsymbol{X}]}$ is a subgraph of $\mathcal{G}$ containing only nodes $\boldsymbol{X}$ and edges among them. A path from a node X to a node Y in $\mathcal{G}$ is a sequence of edges, which does not include a particular node more than once. Two sets of nodes $\boldsymbol{X}, \boldsymbol{Y}$ are said to be d-separated by a third set $\boldsymbol{Z}$ in a DAG $\mathcal{G}$, denoted by $(\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} | \boldsymbol{Z})_{\mathcal{G}}$, if every edge path from nodes in $\boldsymbol{X}$ to nodes in $\boldsymbol{Y}$ is "blocked" by nodes in $\boldsymbol{Z}$. The criterion of blockage follows (Pearl, 2000, Def. 1.2.3). For a more detailed survey on SCMs, we refer readers to (Pearl, 2000; Bareinboim et al., 2022).

2 CAUSAL INVERSE REINFORCEMENT LEARNING

We investigate the sequential decision-making setting concerning a set of actions $\boldsymbol{X}$, a series of covariates $\boldsymbol{Z}$, and a latent reward Y in an SCM M. An expert (e.g., a physician, driver), operating in SCM M, selects actions following a *behavior policy*, which is the collection of structural functions $f_{\boldsymbol{X}} = \{f_X \mid X \in \boldsymbol{X}\}$. The expert's performance is evaluated as the expected reward $\mathbb{E}[Y]$. On the other hand, a learning agent (i.e., the imitator) intervenes on actions $\boldsymbol{X}$ following an ordering $X_1 \prec \cdots \prec X_n$; each action X_i is associated with a set of features $\boldsymbol{PA}_i^* \subseteq \boldsymbol{O} \setminus \{X_i\}$. A policy $\boldsymbol{\pi}$ over actions $\boldsymbol{X}$ is a sequence of decision rules $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_n\}$. Each decision rule $\pi_i(X_i \mid \boldsymbol{Z}_i)$ is a probability distribution over an action $X_i \in \boldsymbol{X}$, conditioning on values of a set of covariates $\boldsymbol{Z}_i \subseteq \boldsymbol{PA}_i^*$. Such policies $\boldsymbol{\pi}$ are also referred to as dynamic treatment regimes (Murphy et al., 2001; Chakraborty & Murphy, 2014), which generalize personalized medicine to time-varying treatment settings in healthcare, in which treatment is repeatedly tailored to a patient's dynamic state.

A *policy intervention* on actions $\boldsymbol{X}$ following a policy $\boldsymbol{\pi}$, denoted by $\text{do}(\boldsymbol{\pi})$, entails a submodel $M_{\boldsymbol{\pi}}$ from a SCM M where structural functions $f_{\boldsymbol{X}}$ associated with $\boldsymbol{X}$ (i.e., the expert's behavior policy) are replaced with decision rules $X_i \sim \pi_i(X_i \mid \boldsymbol{Z}_i)$ for every $X_i \in \boldsymbol{X}$. A critical assumption throughout this paper is that submodel $M_{\boldsymbol{\pi}}$ does not contain any cycles. Similarly, the interventional distribution $P(\boldsymbol{V} \mid \text{do}(\boldsymbol{\pi}))$ induced by policy $\boldsymbol{\pi}$ is defined as the joint distribution over $\boldsymbol{V}$ in $M_{\boldsymbol{\pi}}$.

Throughout this paper, detailed parametrizations of the underlying SCM M are assumed to be *unknown* to the agent. Instead, the agent has access to the **input**: (1) a causal diagram $\mathcal{G}$ associated with M, and (2) the expert's demonstrations, summarized as the observational distribution $P(\boldsymbol{O})$. The goal of the agent is to **output** an imitating policy $\boldsymbol{\pi}^*$ that achieves the expert's performance.

Definition 1. For an SCM $M = \langle \boldsymbol{U}, \boldsymbol{V}, \mathcal{F}, P(\boldsymbol{U}) \rangle$, an *imitating policy* $\boldsymbol{\pi}^*$ is a policy such that its expected reward is lower bounded by the expert's reward, i.e., $\mathbb{E}_M[Y \mid \text{do}(\boldsymbol{\pi}^*)] \geq \mathbb{E}_M[Y]$.

In words, the right-hand side is the expert's performance that the agent wants to achieve, while the left-hand side is the real reward experienced by the agent. The challenge in imitation learning arises from the fact that the reward Y is not specified and latent, i.e., $Y \notin \boldsymbol{O}$. This precludes approaches that identify $\mathbb{E}[Y|\text{do}(\boldsymbol{\pi})]$ directly from the demonstration data (e.g., through the do- or soft-do-calculus Pearl (2000); Correa & Bareinboim (2020)).

There exist methods in the literature for finding an imitating policy in Def. 1. Before describing their details, we first introduce some necessary concepts. For any policy $\boldsymbol{\pi}$, we summarize its associated state-action domain using a sequence of pairs of variables called a policy scope $\mathcal{S}$.

Definition 2 (Lee & Bareinboim (2020))**.** For an SCM M, a policy scope $\mathcal{S}$ (for short, scope) over actions $\boldsymbol{X}$ is a sequence of tuples $\{\langle X_i, \boldsymbol{Z}_i \rangle\}_{i=1}^n$ where $\boldsymbol{Z}_i \subseteq \boldsymbol{PA}_i^*$ for every $X_i \in \boldsymbol{X}$.

We will consistently use $\boldsymbol{\pi} \sim \mathcal{S}$ to denote a policy $\boldsymbol{\pi}$ associated with scope $\mathcal{S}$. For example, consider a policy scope $\mathcal{S} = \{\langle X_1, \{Z_1\}\rangle, \langle X_2, \{Z_2\}\rangle\}$ over actions X_1, X_2 in Fig. 1c. A policy $\boldsymbol{\pi} \sim \mathcal{S}$ is a sequence of distributions $\boldsymbol{\pi} = \{\pi_1(X_1 \mid Z_1), \pi_2(X_2 \mid Z_2)\}$.

Zhang et al. (2020); Kumor et al. (2021) provide a graphical condition that is sufficient for learning an imitating policy via behavioral cloning (BC) provided with a causal diagram $\mathcal{G}$. For a policy scope $\mathcal{S} = \{\langle X_i, \boldsymbol{Z}_i \rangle\}_{i=1}^n$, let $\mathcal{G}^{(i)}$, $i = 1, \ldots, n$, denote a manipulated graph obtained from $\mathcal{G}$ by the following steps: for all $j = i+1, \ldots, n$, (1) remove arrows coming into every action X_j; and (2) add direct arrows from nodes in $\boldsymbol{Z}_j$ to X_j. Formally, the *sequential* π*-backdoor* criterion is defined as:

Definition 3 (Kumor et al. (2021))**.** Given a causal diagram $\mathcal{G}$, a policy scope $\mathcal{S} = \{\langle X_i, \boldsymbol{Z}_i \rangle\}_{i=1}^n$ is said to satisfy the *sequential* π*-backdoor* criterion in $\mathcal{G}$ (for short, π-backdoor admissible) if at each $X_i \in \boldsymbol{X}$, one of the following conditions hold: (1) X_i is not an ancestor of Y in $\mathcal{G}^{(i)}$, i.e., $X \notin An(Y)_{\mathcal{G}^{(i)}}$; or (2) $\boldsymbol{Z}_i$ blocks all backdoor path from X_i to Y in $\mathcal{G}^{(i)}$, i.e., $(Y \perp\!\!\!\perp X_i | \boldsymbol{Z}_i)$ in $\mathcal{G}^{(i)}_{\underline{X_i}}$.

(Kumor et al., 2021) showed that whenever a π-backdoor admissible scope $\mathcal{S}$ is available, one could learn an imitating policy $\boldsymbol{\pi}^* \sim \mathcal{S}$ by setting $\pi_i^*(x_i \mid \boldsymbol{z}_i) = P(x_i \mid \boldsymbol{z}_i)$ for every action $X_i \in \boldsymbol{X}$. For instance, consider the causal diagram $\mathcal{G}$ in Fig. 1c. Scope $\mathcal{S} = \{\langle X_1, \{Z_1\}\rangle, \langle X_2, \{Z_2\}\rangle\}$ is π-backdoor admissible since $(X_1 \perp\!\!\!\perp Y | Z_1)$ and $(X_2 \perp\!\!\!\perp Y | Z_2)$ hold in $\mathcal{G}$, which is a super graph containing both manipulated $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$. An imitating policy $\boldsymbol{\pi}^* = \{\pi_1^*, \pi_2^*\}$ is thus obtainable by setting $\pi_1^*(X_1 \mid Z_1) = P(X_1 \mid Z_1)$ and $\pi_2^*(X_2 \mid Z_2) = P(X_2 \mid Z_2)$. While impressive, a caveat of their results is that the performance of the imitator is restricted by that of the expert, i.e., $\mathbb{E}[Y \mid \text{do}(\boldsymbol{\pi}^*)] = \mathbb{E}[Y]$. In other words, causal BC provides an efficient way to mimic the expert's performance. If the expert's behavior is far from optimal, the same will hold for the learning agent.

2.1 Minimal Sequential Backdoor Criterion

To circumvent this issue, we take a somewhat different approach to causal imitation by incorporating the principle of inverse reinforcement learning (IRL) principle. Following the game-theoretic approach (Syed & Schapire, 2008), we formulate the problem as learning to play a two-player zero-sum game in which the agent chooses a policy, and the nature chooses an SCM instance. A key property of this algorithm is that it allows us to incorporate prior parametric knowledge about the latent reward signal. When such knowledge is informative, our algorithm is about to obtain a policy that could significantly outperform the expert with respect to the unknown causal environment, while at the same time are guaranteed to be no worse. Formally, let $\mathscr{M} = \{\forall M \mid \mathcal{G}_M = \mathcal{G}, P_M(\boldsymbol{O}) = P(\boldsymbol{O})\}$ denote the set of SCMs compatible with both the causal diagram $\mathcal{G}$ and the observational distribution $P(\boldsymbol{O})$. Fix a policy scope $\mathcal{S}$. Now consider the optimization problem defined as follows.

$$\nu^* = \min_{\boldsymbol{\pi} \sim \mathcal{S}} \max_{M \in \mathscr{M}} \mathbb{E}_M[Y] - \mathbb{E}_M[Y \mid \text{do}(\boldsymbol{\pi})]. \tag{1}$$

The inner maximization in the above equation can be viewed as an *causal IRL* step where we attempt to "guess" a worst-case SCM $\hat{M}$ compatible with $\mathcal{G}$ and $P(\boldsymbol{O})$ that prioritizes the expert's policy.

That is, the gap in the performance between the expert's and the imitator's policies is maximized. Meanwhile, since the expert's reward $\mathbb{E}_M[Y]$ is not affected by the imitator's policy $\boldsymbol{\pi}$, the outer minimization is equivalent to a planning step that finds a policy $\boldsymbol{\pi}^*$ optimizing the learned SCM $\hat{M}$. Obviously, the solution $\boldsymbol{\pi}^*$ is an imitating policy if gap $\nu^* = 0$. In cases where the expert is sub-optimal, i.e., $E_{\hat{M}}[Y] < E_{\hat{M}}[Y \mid \text{do}(\boldsymbol{\pi})]$ for some policies $\boldsymbol{\pi}$, we may have $\nu^* < 0$. That is, the policy $\boldsymbol{\pi}^*$ will dominate the expert's policy $f_{\boldsymbol{X}}$ regardless of parametrizations of SCM M in the worst-case scenario. In other words, $\boldsymbol{\pi}^*$ to some extent ignores the sub-optimal expert, and instead exploits prior knowledge about the underlying model.

Despite the clear semantics in terms of causal models, the optimization problem in Eq. (1) requires the learner to search over all possible SCMs compatible with the causal diagram $\mathcal{G}$ and observational distribution $P(\boldsymbol{O})$. In principle, it entails a quite challenging search since one does not have access to the parametric forms of the underlying structural functions $\mathcal{F}$ nor the exogenous distribution $P(\boldsymbol{U})$. It is not clear how the existing optimization procedures can be used.

In this paper, we will develop novel methods to circumvent this issue, thus leading to effective imitating policies. Our first algorithm relies on a refinement of the sequential π-backdoor, based on the concept of minimality. A subscope $\mathcal{S}'$ of a policy scope $\mathcal{S} = \{\langle X_i, \boldsymbol{Z}_i\rangle\}_{i=1}^n$, denoted by $\mathcal{S}' \subseteq \mathcal{S}$, is a sequence $\{\langle X_i, \boldsymbol{Z}'_i\rangle\}_{i=1}^n$ where $\boldsymbol{Z}'_i \subseteq \boldsymbol{Z}_i$ for every $X_i \in \boldsymbol{X}$. A proper subscope $\mathcal{S}' \subset \mathcal{S}$ is a subscope in $\mathcal{S}$ other than $\mathcal{S}$ itself. The minimal π-backdoor admissible scope is defined as follows.

Definition 4. Given a causal diagram $\mathcal{G}$, a π-backdoor admissible scope $\mathcal{S}$ is said to be *minimal* if there exists no proper subscope $\mathcal{S}' \subset \mathcal{S}$ satisfying the sequential π-backdoor in $\mathcal{G}$.

Theorem 1. *Given a causal diagram $\mathcal{G}$, if there exists a minimal π-backdoor admissible scope $\mathcal{S} = \{\langle X_i, \boldsymbol{Z}_i\rangle\}_{i=1}^n$ in $\mathcal{G}$, consider the following conditions:*

1. *Let effective actions $\boldsymbol{X}^* = \boldsymbol{X} \cap An(Y)_{\mathcal{G}_{\mathcal{S}}}$ and effective covariates $\boldsymbol{Z}^* = \bigcup_{X_i \in \boldsymbol{X}^*} \boldsymbol{Z}_i$;*
2. *For $i = 1, \ldots, n+1$, let $\boldsymbol{X}^*_{<i} = \{\forall X_j \in \boldsymbol{X}^* \mid j < i\}$ and $\boldsymbol{Z}^*_{<i} = \bigcup_{X_j \in \boldsymbol{X}^*_{<i}} \boldsymbol{Z}_j$.*

Then, for any policy $\boldsymbol{\pi} \sim \mathcal{S}$, the expected reward $\mathbb{E}[Y \mid do(\boldsymbol{\pi})]$ is computable from $P(\boldsymbol{O}, Y)$ as:

$$\mathbb{E}[Y \mid do(\boldsymbol{\pi})] = \sum_{\boldsymbol{x}^*, \boldsymbol{z}^*} \mathbb{E}[Y \mid \boldsymbol{x}^*, \boldsymbol{z}^*] \rho_{\boldsymbol{\pi}}(\boldsymbol{x}^*, \boldsymbol{z}^*) \tag{2}$$

where the occupancy measure $\rho_{\boldsymbol{\pi}}(\boldsymbol{x}^*, \boldsymbol{z}^*) = \prod_{X_i \in \boldsymbol{X}^*} P\left(\boldsymbol{z}_i \mid \boldsymbol{x}^*_{<i}, \boldsymbol{z}^*_{<i}\right) \pi_i(x_i \mid \boldsymbol{z}_i)$.

To illustrate, consider again the causal diagram $\mathcal{G}$ in Fig. 1c; the manipulated diagram $\mathcal{G}^{(2)} = \mathcal{G}$ and $\mathcal{G}^{(1)}$ is obtained from $\mathcal{G}$ by removing $Z_2 \leftrightarrow X_2$. While scope $\mathcal{S}_1 = \{\langle X_1, \{Z_1\}\rangle, \langle X_2, \{Z_2\}\rangle\}$ satisfies the sequential π-backdoor, it is not minimal since $(X_1 \perp\!\!\!\perp Y)$ in $\mathcal{G}^{(1)}_{\underline{X_1}}$. On the other hand, $\mathcal{S}_2 = \{\langle X_1, \emptyset\rangle, \langle X_2, \{Z_2\}\rangle\}$ is minimal π-backdoor admissible since $(X_2 \perp\!\!\!\perp Y \mid Z_2)$ holds true in $\mathcal{G}^{(2)}_{\underline{X_2}}$; and the covariate set $\{Z_2\}$ is minimal due to the presence of the backdoor path $X_2 \leftarrow Z_2 \rightarrow Y$.

Let us focus on the minimal π-backdoor admissible scope $\mathcal{S}_2$. Note that $\mathcal{G}_{\mathcal{S}_2}$ is a subgraph obtained from $\mathcal{G}$ by removing the bi-directed arrow $Z_2 \leftrightarrow X_2$. We must have effective actions $\boldsymbol{X}^* = \{X_1, X_2\}$ and effective covariates $\boldsymbol{Z}^* = \{Z_2\}$. Therefore, $\boldsymbol{Z}^*_{<1} = \boldsymbol{Z}^*_{<2} = \emptyset$ and $\boldsymbol{Z}^*_{<3} = \{Z_2\}$. For any policy $\boldsymbol{\pi} \sim \mathcal{S}_2$, Thm. 1 implies $\mathbb{E}[Y \mid \text{do}(\boldsymbol{\pi})] = \sum_{x_1, x_2, z_2} \mathbb{E}[Y \mid x_1, x_2, z_2] P(z_2|x_1) \pi_2(x_2|z_2) \pi(x_1)$. On the other hand, the same result in Thm. 1 does not necessarily hold for a non-minimal π-backdoor admissible scope. For instance, consider again the non-minimal scope $\mathcal{S}_1 = \{\langle X_1, \{Z_1\}\rangle, \langle X_2, \{Z_2\}\rangle\}$. The expected reward $\mathbb{E}[Y \mid \text{do}(\boldsymbol{\pi})]$ of a policy $\boldsymbol{\pi} \sim \mathcal{S}_2$ is not computable from Eq. (2), and is ultimately not identifiable from distribution $P(\boldsymbol{O}, Y)$ in $\mathcal{G}$ (Tian, 2008).

2.2 Imitation via Inverse Reinforcement Learning

Once a minimal π-backdoor admissible scope $\mathcal{S}$ is found, there exist effective procedures to solve for an imitating policy in Eq. (1). Let $\mathscr{R}$ be a hypothesis class containing all expected rewards $\mathbb{E}_M[Y \mid \boldsymbol{x}^*, \boldsymbol{z}^*]$ compatible with candidate SCMs $M \in \mathscr{M}$, i.e., $\mathscr{R} = \{\mathbb{E}_M[Y \mid \boldsymbol{x}^*, \boldsymbol{z}^*] \mid \forall M \in \mathscr{M}\}$. Applying the identification formula in Thm. 1 reduces the optimization problem in Eq. (1) as follows:

$$\nu^* = \min_{\boldsymbol{\pi} \sim \mathcal{S}} \max_{r \in \mathscr{R}} \sum_{\boldsymbol{x}^*, \boldsymbol{z}^*} r(\boldsymbol{x}^*, \boldsymbol{z}^*) \left(\rho(\boldsymbol{x}^*, \boldsymbol{z}^*) - \rho_{\boldsymbol{\pi}}(\boldsymbol{x}^*, \boldsymbol{z}^*)\right) \tag{3}$$

where the expert's occupancy measure $\rho(\boldsymbol{x}^*, \boldsymbol{z}^*) = P(\boldsymbol{x}^*, \boldsymbol{z}^*)$ and the agent's occupancy measure $\rho_{\boldsymbol{\pi}}(\boldsymbol{x}^*, \boldsymbol{z}^*)$ is given by Eq. (2). The above minimax problem is solvable using standard IRL algorithms. The identification result in Thm. 1 ensures that the learned policy applies to any SCM compatible with the causal diagram and the observational data, thus robust to the unobserved confounding bias in the expert's demonstrations. Henceforth, we will consistently refer to Eq. (3) as the ***canonical equation of causal IRL***. In this paper, we solve for an imitating policy $\boldsymbol{\pi}^*$ in Eq. (3) using state-of-the-art IRL algorithms, provided with common choices of parametric reward functions. These algorithms include the multiplicative-weights algorithm (MWAL) (Syed & Schapire, 2008) and the generative adversarial imitation learning (GAIL) (Ho & Ermon, 2016). We refer readers to Algs. 3 and 4 in Appendix C for more discussions on the pseudo-code and implementation details.

Causal MWAL (Abbeel & Ng, 2004; Syed & Schapire, 2008) study IRL in Markov decision processes where the reward function $r(\boldsymbol{x}^*, \boldsymbol{z}^*)$ is a linear combination of k-length *feature expectations* vectors $\boldsymbol{\phi}(\boldsymbol{x}^*, \boldsymbol{z}^*)$. Particularly, let $r(\boldsymbol{x}^*, \boldsymbol{z}^*) = \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}^*, \boldsymbol{z}^*)$ for a coefficient vector $\boldsymbol{w}$ contained in a convex set $\mathbb{S}^k = \left\{\boldsymbol{w} \in \mathbb{R}^k \mid \|\boldsymbol{w}\|_1 = 1 \text{ and } \boldsymbol{w} \succeq \boldsymbol{0}\right\}$. Let $\boldsymbol{\phi}^{(i)}$ be the i-th component of feature vector $\boldsymbol{\phi}$ and let deterministic policies with scope $\mathcal{S}$ be ordered by $\boldsymbol{\pi}^{(1)}, \ldots, \boldsymbol{\pi}^{(n)}$. The canonical equation in Eq. (3) is reducible to a two-person zero-sum matrix game under linearity.

Proposition 1. *For a hypothesis class $\mathscr{R} = \{r = \boldsymbol{w} \cdot \boldsymbol{\phi} \mid \boldsymbol{w} \in \mathbb{S}^k\}$, the solution ν^* of the canonical equation in Eq.* (3) *is obtainable by solving the following minimax problem:*

$$\nu^* = \min_{\boldsymbol{\pi} \sim \mathcal{S}} \max_{\boldsymbol{w} \in \mathbb{S}^k} \boldsymbol{w}^\top \boldsymbol{G} \boldsymbol{\pi}, \tag{4}$$

where $\boldsymbol{G}$ is a $k \times n$ matrix given by $\boldsymbol{G}(i, j) = \sum_{\boldsymbol{x}^*, \boldsymbol{z}^*} \boldsymbol{\phi}^{(i)}(\boldsymbol{x}^*, \boldsymbol{z}^*) \left(\rho(\boldsymbol{x}^*, \boldsymbol{z}^*) - \rho_{\boldsymbol{\pi}^{(j)}}(\boldsymbol{x}^*, \boldsymbol{z}^*)\right)$.

There exist effective multiplicative weights algorithms for solving the matrix game in Eq. (4), including MW (Freund & Schapire, 1999) and MWAL (Syed & Schapire, 2008).

Causal GAIL (Ho & Ermon, 2016) introduces the GAIL algorithm for learning an imitating policy in Markov decision processes with a general family of non-linear reward functions. In particular, $r(\boldsymbol{x}^*, \boldsymbol{z}^*)$ takes values in the real space $\mathbb{R}$, i.e., $r \in \mathbb{R}^{\boldsymbol{X}^*, \boldsymbol{Z}^*}$ where $\mathbb{R}^{\boldsymbol{X}^*, \boldsymbol{Z}^*} = \{r : \mathscr{D}_{\boldsymbol{X}^*} \times \mathscr{D}_{\boldsymbol{Z}^*} \mapsto \mathbb{R}\}$. The complexity of reward function r is penalized by a convex regularization function $\psi(r)$, i.e.,

$$\nu^* = \min_{\boldsymbol{\pi} \sim \mathcal{S}} \max_{r \in \mathbb{R}^{\boldsymbol{X} \times \boldsymbol{Z}}} \sum_{\boldsymbol{x}^*, \boldsymbol{z}^*} r(\boldsymbol{x}^*, \boldsymbol{z}^*) \left(\rho(\boldsymbol{x}^*, \boldsymbol{z}^*) - \rho_{\boldsymbol{\pi}}(\boldsymbol{x}^*, \boldsymbol{z}^*)\right) - \psi(r) \tag{5}$$

Henceforth, we will consistently refer to Eq. (5) as the *penalized canonical equation* of causal IRL. It is often preferable to solve its conjugate form. Formally,

Proposition 2. *For a hypothesis class $\mathscr{R} = \{r : \mathscr{D}_{\boldsymbol{X}^*} \times \mathscr{D}_{\boldsymbol{Z}^*} \mapsto \mathbb{R}\}$ regularized by ψ, the solution ν^* of the penalized canonical equation in Eq.* (5) *is obtainable by solving the following problem:*

$$\nu^* = \min_{\boldsymbol{\pi} \sim \mathcal{S}} \psi^*\left(\rho - \rho_{\boldsymbol{\pi}}\right) \tag{6}$$

where ψ^* be a conjugate function of ψ and is given by $\psi^* = \max_{r \in \mathbb{R}^{\boldsymbol{X} \times \boldsymbol{Z}}} a^\top r - \psi(r)$.

Eq. (6) seeks a policy $\boldsymbol{\pi}$ which minimizes the divergence of the occupancy measures between the imitator and the expert, as measured by the function ψ^*. The computational framework of generative adversarial networks (Goodfellow et al., 2014) provides an effective approach to solve such a matching problem, e.g., the GAIL algorithm (Ho & Ermon, 2016).

3 Causal Imitation without Sequential Backdoor

In this section, we investigate causal IRL beyond the condition of minimal sequential π-backdoor. Observe that the key to the reduction of the canonical causal IRL equation in Eq. (3) lies in the identification of expected rewards $\mathbb{E}[Y \mid \text{do}(\boldsymbol{\pi})]$ had the latent reward Y been observed. Next we will study general conditions under which $\mathbb{E}[Y \mid \text{do}(\boldsymbol{\pi})]$ is uniquely discernible from distribution $P(\boldsymbol{O}, Y)$ in the causal diagram $\mathcal{G}$, called the *identifiability* of causal effects (Pearl, 2000, Def. 3.2.4).

Definition 5 (Identifiability)**.** Given a causal diagram $\mathcal{G}$ and a policy $\boldsymbol{\pi} \sim \mathcal{S}$, the expected reward $\mathbb{E}[Y \mid \text{do}(\boldsymbol{\pi})]$ is said to be identifiable from distribution $P(\boldsymbol{O}, Y)$ in $\mathcal{G}$ if $\mathbb{E}[Y \mid \text{do}(\boldsymbol{\pi})]$ is uniquely computable from $P(\boldsymbol{O}, Y)$ in any SCM M compatible with $\mathcal{G}$.

We say a policy scope $\mathcal{S}$ is identifiable (from $P(\boldsymbol{O}, Y)$ in $\mathcal{G}$) if for all policies $\boldsymbol{\pi} \sim \mathcal{S}$, the corresponding expected rewards $\mathbb{E}[Y \mid \text{do}(\boldsymbol{\pi})]$ are identifiable from $P(\boldsymbol{O}, Y)$ in $\mathcal{G}$. Our next result shows that whenever an identifiable policy scope $\mathcal{S}$ is found, one could always reduce the causal IRL problem to the canonical optimization equation in Eq. (3).

Theorem 2. *Given a causal diagram $\mathcal{G}$, a policy scope $\mathcal{S}$ is identifiable from $P(\boldsymbol{O}, Y)$ in $\mathcal{G}$ if and only if for any policy $\boldsymbol{\pi} \sim \mathcal{S}$, the expected reward $\mathbb{E}[Y \mid do(\boldsymbol{\pi})]$ is computable from $P(\boldsymbol{O}, Y)$ as*

$$\mathbb{E}[Y \mid do(\boldsymbol{\pi})] = \sum_{\boldsymbol{x}^*, \boldsymbol{z}^*} \mathbb{E}[Y \mid \boldsymbol{x}^*, \boldsymbol{z}^*] \rho_{\boldsymbol{\pi}}(\boldsymbol{x}^*, \boldsymbol{z}^*) \tag{7}$$

where subsets $\boldsymbol{X}^* \subseteq \boldsymbol{X}$, $\boldsymbol{Z}^* \subseteq \boldsymbol{O} \setminus \boldsymbol{X}$; and the imitator's occupancy measure $\rho_{\boldsymbol{\pi}}(\boldsymbol{x}^*, \boldsymbol{z}^*)$ is a function of the observational distribution $P(\boldsymbol{O})$ and policy $\boldsymbol{\pi}$.

Thm. 2 suggests a general procedure to learn an imitating policy via causal IRL. Whenever an identifiable scope $\mathcal{S}$ is found, the identification formula in Eq. (7) permits one to reduce the optimization problem in Eq. (1) to the canonical equation in Eq. (3). One could thus obtain an imitating policy $\boldsymbol{\pi} \sim \mathcal{S}$ by solving Eq. (3) where the expert's occupancy measure $\rho(\boldsymbol{x}^*, \boldsymbol{z}^*) = P(\boldsymbol{x}^*, \boldsymbol{z}^*)$ and the imitator's occupancy measure $\rho_{\boldsymbol{\pi}}(\boldsymbol{x}^*, \boldsymbol{z}^*)$ is given by Eq. (7). As an example, consider the frontdoor diagram described in Fig. 2a and a policy scope $\mathcal{S} = \{\langle X, \emptyset \rangle\}$. The expected reward $\mathbb{E}[Y \mid \text{do}(\pi)] = \sum_{x'} \mathbb{E}[Y \mid \text{do}(x')]\pi(x')$ and $\mathbb{E}[Y \mid \text{do}(x')]$ is identifiable from $P(X, Y, Z)$ using the frontdoor adjustment formula (Pearl, 2000, Thm. 3.3.4). The expected reward $\mathbb{E}[Y \mid \text{do}(\pi)]$ of any policy $\pi(X)$ could be written as:

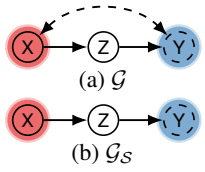


Figure 2: Frontdoor

$$\mathbb{E}[Y \mid \text{do}(\pi)] = \sum_{z,x} \mathbb{E}[Y \mid x, z] P(x) \sum_{x'} P(z|x')\pi(x'). \tag{8}$$

Let occupancy measures $\rho(x, z) = P(x, z)$ and $\rho_{\boldsymbol{\pi}}(x, z) = P(x) \sum_{x'} P(z|x')\pi(x')$. We could thus learn an imitating policy in the frontdoor diagram by solving the canonical equation given by:

$$\nu^* = \min_{\pi \sim \mathcal{S}} \max_{r \in \mathscr{R}} \sum_{x,z} r(x, z) \left(\rho(x, z) - \rho_{\boldsymbol{\pi}}(x, z)\right), \tag{9}$$

where $\mathscr{R}$ is a hypothesis class of the reward function $r(x, z) \triangleq \mathbb{E}[Y \mid x, z]$. The solution $\pi^*(X)$ is an imitating policy performing at least as well as the expert's behavior policy if the gap $\nu^* \leq 0$.

Next, we will describe how to obtain the identification formula in Eq. (7) provided with an identifiable scope $\mathcal{S}$. Without loss of generality, we will assume that the reward Y is the only endogenous variable that is latent in the causal diagram $\mathcal{G}$, i.e., $\boldsymbol{V} = \boldsymbol{O} \cup \{Y\}$.[1] We will utilize a special type of clustering of nodes in the causal diagram $\mathcal{G}$, called the *confounded component* (for short, c-component).

Definition 6 (C-component (Tian & Pearl, 2002))**.** For a causal diagram $\mathcal{G}$, a subset $\boldsymbol{C} \subseteq \boldsymbol{V}$ is a c-component if any pair $V_i, V_j \in \boldsymbol{C}$ is connected by a bi-directed path in $\mathcal{G}$.

For instance, the frontdoor diagram in Fig. 2a contains two c-components $\boldsymbol{C}_1 = \{X, Y\}$ and $\boldsymbol{C}_2 = \{Z\}$. We will utilize a sound and complete procedure IDENTIFY (Tian, 2002; 2008) for identifying causal effects $\mathbb{E}[Y \mid \text{do}(\boldsymbol{\pi})]$ of an arbitrary policy $\boldsymbol{\pi} \sim \mathcal{S}$. Particularly, IDENTIFY takes as input the causal diagram $\mathcal{G}$, a reward Y, and a policy scope $\mathcal{S}$. It returns an identification formula for $\mathbb{E}[Y \mid \text{do}(\boldsymbol{\pi})]$ from $P(\boldsymbol{O}, Y)$ if expected rewards of all policies $\boldsymbol{\pi} \sim \mathcal{S}$ are identifiable. Otherwise, IDENTIFY$(\mathcal{G}, Y, \mathcal{S})$ = "FAIL". Details of IDENTIFY are shown in (Zhang et al., 2020, Appendix B). Recall that $\mathcal{G}_{\mathcal{S}}$ is the causal diagram of submodel $M_{\boldsymbol{\pi}}$ induced by policy $\boldsymbol{\pi} \sim \mathcal{S}$. Fig. 2b shows diagram $\mathcal{G}_{\mathcal{S}}$ obtained from the frontdoor graph $\mathcal{G}$ and scope $\mathcal{S} = \{\langle X, \emptyset \rangle\}$ described in Fig. 2a. Let $\boldsymbol{Z}_Y = An(Y)$ be ancestors of Y in $\mathcal{G}_{\mathcal{S}}$. Our next result shows that IDENTIFY$(\mathcal{G}, Y, \mathcal{S})$ is ensured to find an identification formula of the form in Eq. (7) when it is identifiable.

Lemma 1. *Given a causal diagram $\mathcal{G}$, a policy scope $\mathcal{S}$ is identifiable from $P(\boldsymbol{O}, Y)$ in $\mathcal{G}$ if and only if* IDENTIFY$(\mathcal{G}, Y, \mathcal{S}) \neq$ "FAIL". *Moreover,* IDENTIFY$(\mathcal{G}, Y, \mathcal{S})$ *returns an identification formula of the form in Eq. (7) where $\boldsymbol{X}^* = Pa(\boldsymbol{C}_Y) \cap \boldsymbol{X}$ and $\boldsymbol{Z}^* = Pa(\boldsymbol{C}_Y) \setminus (\{Y\} \cup \boldsymbol{X})$; and $\boldsymbol{C}_Y$ is a c-component containing reward Y in subgraph $\mathcal{G}_{[An(\boldsymbol{Z}_Y)]}$.*

[1] Otherwise, one could always simplify the diagram $\mathcal{G}$ and project other latent variables $\boldsymbol{L} \setminus \{Y\}$ using the projection algorithm (Tian, 2002, Sec. 4.5), without affecting the identifiability of target query $E[Y \mid \text{do}(\boldsymbol{\pi})]$.

For example, for the frontdoor diagram $\mathcal{G}$ in Fig. 2a, the manipulated diagram $\mathcal{G}_{\mathcal{S}}$ with scope $\mathcal{S} = \{\langle X, \emptyset\rangle\}$ is described in Fig. 2b. Since $\boldsymbol{Z}_Y = An(Y)_{\mathcal{G}_{\mathcal{S}}} = \{X, Z, Y\}$, $\boldsymbol{C}_Y$ is thus given by $\{X, Y\}$. Lem. 1 implies that $\boldsymbol{X}^* = Pa(\{X, Y\}) \cap \{X\} = \{X\}$ and $\boldsymbol{Z}^* = Pa(\{X, Y\}) \setminus \{X, Y\} = \{Z\}$. Applying IDENTIFY$(\mathcal{G}, Y, \{\langle X, \emptyset\rangle\})$ returns the frontdoor adjustment formula in Eq. (8).

3.1 SEARCHING FOR IDENTIFIABLE POLICY SCOPES

The remainder of this section describes an effective algorithm to find identifiable policy scopes $\mathcal{S}$ had the latent reward signal Y been observed. Let $\mathbb{S}$ denote the collection of all identifiable policy scopes $\mathcal{S}$ from distribution $P(\boldsymbol{O}, Y)$ in the causal diagram $\mathcal{G}$. Our algorithm LISTIDSCOPE, described in Alg. 1, enumerates elements in $\mathbb{S}$. It takes as input a causal diagram $\mathcal{G}$, a reward signal Y, and subsets $\boldsymbol{L} = \emptyset$ and $\boldsymbol{R} = \bigcup_{i=1}^{n} \boldsymbol{PA}_i^*$. More specifically, LISTIDSCOPE maintains two scopes $\mathcal{S}_l \subseteq \mathcal{S}_r$ (Step 2). It performs backtrack search to find identifiable scopes $\mathcal{S}$ in $\mathcal{G}$ such that $\mathcal{S}_l \subseteq \mathcal{S} \subseteq \mathcal{S}_r$. It aborts branches that either (1) all subscopes in $\mathcal{S}_r$ are identifiable (Step 3); or (2) all subscopes containing $\mathcal{S}_l$ are non-identifiable (Step 6). The following proposition supports our aborting criterion.

Lemma 2. *Given a causal diagram $\mathcal{G}$, for policy scopes $\mathcal{S}' \subseteq \mathcal{S}$, $\mathcal{S}'$ is identifiable from distribution $P(\boldsymbol{O}, Y)$ in $\mathcal{G}$ if $\mathcal{S}$ is identifiable from $P(\boldsymbol{O}, Y)$ in $\mathcal{G}$.*

At Step 7, LISTIDSCOPE picks an arbitrary variable V that is included in input covariates $\boldsymbol{R}$ but not in $\boldsymbol{L}$. It then recursively returns all identifiable policy scopes $\mathcal{S}$ in $\mathcal{G}$: the first recursive call returns scopes taking V as an input for some actions $X_i \in \boldsymbol{X}$ and the second call return all scopes that do not consider V when selecting values for all actions $\boldsymbol{X}$. We say a policy $\boldsymbol{\pi}$ is associated with a collection of policy scopes $\mathbb{S}$, denoted by $\boldsymbol{\pi} \sim \mathbb{S}$, if there exists $\mathcal{S} \in \mathbb{S}$ so that $\boldsymbol{\pi} \sim \mathcal{S}$. It is possible to show that LISTIDSCOPE produces a collection of identifiable scopes that is sufficient for the imitation task.

Algorithm 1: LISTIDSCOPE

1: **Input:** $\mathcal{G}, Y$ and subsets $\boldsymbol{L} \subseteq \boldsymbol{R}$
2: **Output:** a set of identifiable policy scopes $\mathbb{S}$
3: Let scopes $\mathcal{S}_r = \{\langle X_i, \boldsymbol{R} \cap \boldsymbol{PA}_i^*\rangle\}_{i=1}^n$ and $\mathcal{S}_l = \{\langle X_i, \boldsymbol{L} \cap \boldsymbol{PA}_i^*\rangle\}_{i=1}^n$.
4: **if** IDENTIFY$(\mathcal{G}, Y, \mathcal{S}_r) \neq$ "FAIL" **then**
5: Output $\mathcal{S}_r$.
6: **end if**
7: **if** IDENTIFY$(\mathcal{G}, Y, \mathcal{S}_l) \neq$ "FAIL" **then**
8: Pick an arbitrary $V \in \boldsymbol{R} \setminus \boldsymbol{L}$.
9: LISTIDSCOPE$(\mathcal{G}, Y, \boldsymbol{L} \cup \{V\}, \boldsymbol{R})$.
10: LISTIDSCOPE$(\mathcal{G}, Y, \boldsymbol{L}, \boldsymbol{R} \setminus \{V\})$.
11: **end if**

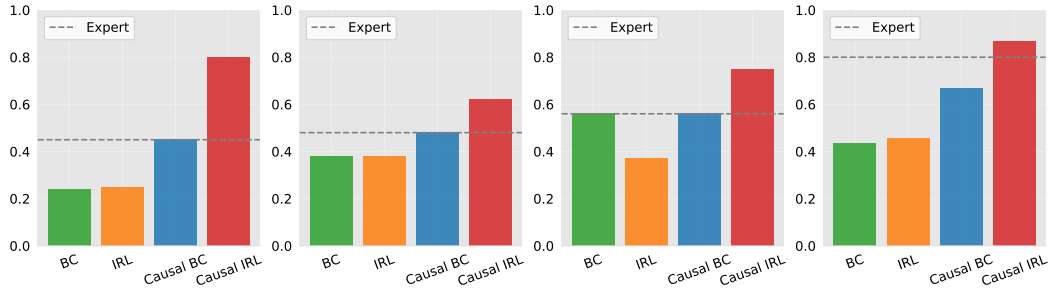
Theorem 3. *For a causal diagram $\mathcal{G}$ and a reward Y,* LISTIDSCOPE$(\mathcal{G}, Y, \emptyset, \bigcup_{i=1}^{n} \boldsymbol{PA}_i^*)$ *enumerates a subset $\mathbb{S}^* \subseteq \mathbb{S}$ so that for any $\boldsymbol{\pi} \sim \mathbb{S}$, there is $\boldsymbol{\pi}^* \sim \mathbb{S}^*$ where $\mathbb{E}[Y \mid do(\boldsymbol{\pi})] = \mathbb{E}[Y \mid do(\boldsymbol{\pi}^*)]$.*

Moreover, LISTIDSCOPE outputs identifiable policy scopes with a polynomial delay. This follows from the observation that LISTIDSCOPE searches over a tree of policy scopes with height at most $|\bigcup_{i=1}^{n} \boldsymbol{PA}_i^*|$ and IDENTIFY$(\mathcal{G}, Y, \mathcal{S})$ terminates in polynomial steps w.r.t. the size of diagram $\mathcal{G}$.

4 EXPERIMENTS

In this section, we demonstrate our framework on various imitation learning tasks, ranging from synthetic causal models to real-world datasets, including highway driving (Krajewski et al., 2018) and images (LeCun, 1998). We find that our approach is able to incorporate parametric knowledge about the reward function and achieve effective imitating policies across different causal diagrams. For all experiments, we evaluate our proposed `Causal-IRL` based on the *canonical equation* formulation in Eq. (3). As a baseline, we also include: (1) standard `BC` mimicking the expert's nominal behavior policy; (2) standard `IRL` utilizing all observed covariates preceding every $X_i \in \boldsymbol{X}$ while being blind to causal relationships in the underlying model; and (3) `Causal-BC` (Zhang et al., 2020; Kumor et al., 2021) that learn an imitating policy with the sequential π-backdoor criterion. We refer readers to Appendix D for additional experiments and more discussions on the experimental setup.

Backdoor Consider an SCM instance compatible with Fig. 1c including binary observed variables $Z_1, X_1, Z_2, X_2, Y \in \{0, 1\}$. `Causal-BC` utilizes a sequential π-backdoor admissible scope $\{\langle X_1, \{Z_1\}\rangle, \langle X_2, \{Z_2\}\rangle\}$; while `Causal-IRL` utilizes the scope $\{\langle X_1, \emptyset\rangle, \langle X_2, \{Z_2\}\rangle\}$ satisfying the minimal sequential π-backdoor. Simulation results, shown in Fig. 3a, reveal that `Causal-IRL` consistently outperforms the expert's policy and other imitation strategies by exploiting additional parametric knowledge about the expected reward $\mathbb{E}[Y \mid X_1, X_2, Z_2]$; `Causal-BC` is able to achieve the expert's performance. Unsurprisingly, neither `BC` nor `IRL` is able to obtain an imitating policy.



(a) Backdoor (b) Highway Driving (c) MNIST Digits (d) Infinite MDPUC

Figure 3: Simulation results (a, b, c, d) for our experiments, where y-axis represents the expected reward of learned policies in the actual causal model; the grey dashed line denotes the expert's reward.

Highway Driving We consider a learning scenario where the agent learns a driving policy from the observed trajectories of a human expert. Causal diagram of this example is provided in Fig. 4 (Appendix D) where X_1 is the accelerations of the ego vehicle at the previous step; Z_1 is both longitudinal and lateral historical accelerations of the ego vehicle two steps ago; X_2 is the velocity of the ego vehicle; Z_2 is the velocity of the preceding vehicle; W indicates the information from surrounding vehicles. Values of X_1, X_2, Z_1, Z_2 are drawn from a real-world driving dataset `HighD` Krajewski et al. (2018). The reward Y is decided by a non-linear function $f_Y(X_2, Z_2, U_Y)$. Both `Causal-IRL` and `Causal-BC` utilize the scope $\{\langle X_1, \emptyset\rangle, \langle X_2, \{Z_2\}\rangle\}$. `Causal-IRL` also exploits the additional knowledge that the expected reward $\mathbb{E}[Y \mid X_1, X_2, Z_2]$ is a monotone function via reward augmentation (Li et al., 2017). Simulation results are shown in Fig. 3b. We found that `Causal-IRL` performs the best among all strategies. `Causal-BC` is able to achieve the expert's performance. `BC` and `IRL` perform the worst among all and fail to obtain an imitating policy.

MNIST Digits Consider again the frontdoor diagram in Fig. 2a. To evaluate the performance of our proposed approach in high-dimensional domains, we now replace variable Z with sampled images drawn from `MNIST` digits dataset (LeCun, 1998). The reward Y is decided by a linear function taking Z and an unobserved confounder $U_{X,Y}$ as input. The `Causal-IRL` formulates the imitation problem as a two-person zero-sum game through the frontdoor adjustment described in Eq. (9), which can be solved by the MW algorithm (Freund & Schapire, 1999; Syed & Schapire, 2008). As shown in Fig. 3c, simulation results reveal that `Causal-IRL` outperforms `Causal-BC` and `BC`; while `IRL` performs the worst among all the algorithms.

Infinite MDPUC To demonstrate our proposed framework in the sequential decision-making setting with an infinite horizon, we consider a generalized Markov decision process incorporating unobserved confounders (Ruan & Di, 2022), called the MDPUC (Zhang & Bareinboim, 2022). This sequential model simulates real-world driving dynamics. By exploiting the Markov property over time steps, we are able to decompose the causal diagram over the infinite horizon into a collection of sub-graphs, one for each time step $i = 1, 2, \ldots$. Fig. 1d shows the causal diagram spanning time steps $i = 1, 2, 3$. As a comparison, `BC` and `IRL` still utilize the stationary policy $\{\langle X_i, \{Z_i\}\rangle\}$. By applying Thm. 1 at each time step, we obtain a π-backdoor admissible policy scope $\{\langle X_i, \{Z_i, X_{i-1}, Z_{i-1}\}\rangle\}$ for `Causal-IRL` and `Causal-BC`. Simulation results are shown in Fig. 3d. One could see by inspection that `Causal-IRL` performs the best and achieves the expert's performance.

5 Conclusion

This paper investigates imitation learning via inverse reinforcement learning (IRL) in the semantical framework of structural causal models. The goal is to find an effective imitating policy that performs at least as well as the expert's behavior policy from combinations of demonstration data, qualitative knowledge the data-generating mechanisms represented as a causal diagram, and quantitative knowledge about the reward function. We provide a graphical criterion (Thm. 1) based on the sequential backdoor, which allows one to obtain an imitating policy by solving a canonical optimization equation of causal IRL. Such a canonical formulation addresses the challenge of the presence of unobserved confounders (UCs), and is solvable by leveraging standard IRL algorithms (Props. 1 and 2). Finally, we move beyond the backdoor criterion and show that the canonical equation is achievable whenever expected rewards of policies are identifiable had the reward also been observed (Thms. 2 and 3).

ETHICS STATEMENT

This paper investigates the theoretical framework of causal inverse RL from the natural trajectories of an expert demonstrator, even when the reward signal is unobserved. Input covariates used by the expert to determine the original values of the action are unknown, introducing unobserved confounding bias in demonstration data. Our framework may apply to various fields in reality, including autonomous vehicle development, industrial automation, and chronic disease management. A positive impact of this work is that we discuss the potential risk of training IRL policy from demonstrations with the presence of unobserved confounding (UC). Our formulation of causal IRL is inherently robust against confounding bias. For example, solving the causal IRL problem in Eq. (1) requires the imitator to learn an effective policy that maximizes the reward in a worst-case causal model where the performance gap between the expert and imitator is the largest possible. More broadly, automated decision systems using causal inference methods prioritize safety and robustness during their decision-making processes. Such requirements are increasingly essential since black-box AI systems are prevalent, and our understandings of their potential implications are still limited.

REPRODUCIBILITY STATEMENT

The complete proof of all theoretical results presented in this paper, including Thms. 1 and 2, is provided in Appendix B. Details on the implementation of the proposed algorithms are included Appendix C. Finally, Appendix D provides a detailed description of the experimental setup. Readers could find all appendices as part of the supplementary text after "References" section. We provided references to all existing datasets used in experiments, including HIGHD (Krajewski et al., 2018) and MNIST (LeCun, 1998). Other experiments are synthetic and do not introduce any new assets.

REFERENCES

Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 1. ACM, 2004.

Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

Paul Atchley, Stephanie Atwood, and Aaron Boulton. The choice to text and drive in younger drivers: Behavior may shape attitude. *Accident Analysis & Prevention*, 43(1):134–142, 2011.

E. Bareinboim and J. Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113:7345–7352, 2016.

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, 1st edition, 2022.

Richard Bellman. Dynamic programming. *Science*, 153(3731):34–37, 1966.

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemyslaw Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

Aude Billard, Sylvain Calinon, Ruediger Dillmann, and Stefan Schaal. Survey: Robot programming by demonstration. *Handbook of robotics*, 59, 2008.

Léon Bottou, Jonas Peters, Denis Xavier Charles, Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Y Simard, and Ed Snelson. Counterfactual reasoning and learning systems: the example of computational advertising. *Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

Bibhas Chakraborty and EE Moodie. Statistical methods for dynamic treatment regimes. *Springer-Verlag. doi*, 10:978–1, 2013.

Bibhas Chakraborty and Susan A Murphy. Dynamic treatment regimes. *Annual review of statistics and its application*, 1:447, 2014.

J. Correa and E. Bareinboim. A calculus for stochastic interventions: Causal effect identification and surrogate experiments. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, NY, 2020. AAAI Press.

Juan D Correa and Elias Bareinboim. From statistical transportability to estimating the effect of stochastic interventions. In *IJCAI*, pp. 1661–1667, 2019.

A. Philip Dawid and Vanessa Didelez. Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview. *Statist. Surv.*, 4:184–231, 2010. doi: 10.1214/10-SS081. URL `https://doi.org/10.1214/10-SS081`.

Pim de Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. In *Advances in Neural Information Processing Systems*, pp. 11693–11704, 2019.

Jalal Etesami and Philipp Geiger. Causal transfer for imitation learning and decision making under sensor-shift. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, NY, 2020. AAAI Press.

Michael Fix, Raymond Struyk, et al. Clear and convincing evidence: Measurement of discrimination in america. Technical report, The Field Experiments Website, 1993.

Yoav Freund and Robert E Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1-2):79–103, 1999.

Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*, 2017.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Bryan Higgs, Montasir Abbas, and Alejandra Medina. Analysis of the wiedemann car following model over different speeds using naturalistic data. In *Procedia of RSS Conference*, pp. 1–22, 2011.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29, 2016.

R.A. Howard. *Dynamic Programming and Markov Processes*. MIT Press, Cambridge, MA, 1960.

Y. Huang and M. Valtorta. Identifiability in causal bayesian networks: A sound and complete algorithm. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI 2006)*, pp. 1149–1156. AAAI Press, Menlo Park, CA, 2006.

Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)*, 50(2):1–35, 2017.

Robert Krajewski, Julian Bock, Laurent Kloeker, and Lutz Eckstein. The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pp. 2118–2125, 2018. doi: 10.1109/ITSC.2018.8569552.

Daniel Kumor, Junzhe Zhang, and Elias Bareinboim. Sequential causal imitation learning with unobserved confounders. *Advances in Neural Information Processing Systems*, 2021.

Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Sanghack Lee and Elias Bareinboim. Characterizing optimal mixed policies: Where to intervene and what to observe. *Advances in Neural Information Processing Systems*, 33, 2020.

Yunzhu Li, Jiaming Song, and Stefano Ermon. Infogail: Interpretable imitation learning from visual demonstrations. *Advances in Neural Information Processing Systems*, 30, 2017.

Jeffrey Mahler and Ken Goldberg. Learning deep policies for robot bin picking by simulating robust grasping sequences. In *Conference on robot learning*, pp. 515–524, 2017.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Urs Muller, Jan Ben, Eric Cosatto, Beat Flepp, and Yann L Cun. Off-road obstacle avoidance through end-to-end learning. In *Advances in neural information processing systems*, pp. 739–746, 2006.

Katharina Mülling, Jens Kober, Oliver Kroemer, and Jan Peters. Learning to select and generalize striking movements in robot table tennis. *The International Journal of Robotics Research*, 32(3): 263–279, 2013.

S. A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, 2003.

S A Murphy, M J van der Laan, J M Robins, and Conduct Problems Prevention Research Group. Marginal Mean Models for Dynamic Regimes. *Journal of the American Statistical Association*, 96 (456):1410–1423, December 2001.

Susan A Murphy. A generalization error for Q-learning. *Journal of Machine Learning Research*, 6 (Jul):1073–1097, 2005.

Allen Newell, Herbert Alexander Simon, et al. *Human problem solving*, volume 104. Prentice-hall Englewood Cliffs, NJ, 1972.

Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pp. 278–287, 1999.

Andrew Y Ng, Stuart J Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, pp. 663–670, 2000.

Stefanos Nikolaidis, Swaprava Nath, Ariel D Procaccia, and Siddhartha Srinivasa. Game-theoretic modeling of human adaptation in human-robot collaboration. In *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*, pp. 323–331, 2017.

Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J Andrew Bagnell, Pieter Abbeel, Jan Peters, et al. An algorithmic perspective on imitation learning. *Foundations and Trends in Robotics*, 7(1-2): 1–179, 2018.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000.

Dean A Pomerleau. Alvinn: An autonomous land vehicle in a neural network. In *Advances in neural information processing systems*, pp. 305–313, 1989.

Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., 1994.

Jette Randløv and Preben Alstrøm. Learning to drive a bicycle using reinforcement learning and shaping. In *ICML*, volume 98, pp. 463–471. Citeseer, 1998.

Kangrui Ruan and Xuan Di. Learning human driving behaviors with sequential causal imitation learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(4):4583–4592, Jun. 2022. doi: 10.1609/aaai.v36i4.20382. URL `https://ojs.aaai.org/index.php/AAAI/article/view/20382`.

I. Shpitser. *Complete Identification Methods for Causal Inference*. PhD thesis, Computer Science Department, University of California, Los Angeles, CA, April 2008.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*. MIT press, 1998.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Gokul Swamy, Sanjiban Choudhury, J Andrew Bagnell, and Steven Wu. Of moments and matching: A game-theoretic framework for closing the imitation gap. In *International Conference on Machine Learning*, pp. 10022–10032. PMLR, 2021.

Gokul Swamy, Sanjiban Choudhury, Drew Bagnell, and Steven Wu. Causal imitation learning under temporally correlated noise. In *International Conference on Machine Learning*, pp. 20877–20890. PMLR, 2022.

Umar Syed and Robert E Schapire. A game-theoretic approach to apprenticeship learning. In *Advances in neural information processing systems*, pp. 1449–1456, 2008.

Guy Tennenholtz, Assaf Hallak, Gal Dalal, Shie Mannor, Gal Chechik, and Uri Shalit. On covariate shift of latent confounders in imitation and reinforcement learning. *arXiv preprint arXiv:2110.06539*, 2021.

J. Tian. *Studies in Causal Reasoning and Learning*. PhD thesis, Computer Science Department, University of California, Los Angeles, CA, November 2002.

Jin Tian. Identifying dynamic sequential plans. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, UAI'08, pp. 554–561, Arlington, Virginia, United States, 2008. AUAI Press. ISBN 0-9749039-4-9.

Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI 2002)*, pp. 567–573, Menlo Park, CA, 2002. AAAI Press/The MIT Press.

Martin Treiber, Ansgar Hennecke, and Dirk Helbing. Congested traffic states in empirical observations and microscopic simulations. *Physical review E*, 62(2):1805, 2000.

John N Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE transactions on automatic control*, 42(5):674–690, 1997.

Benito van der Zander, Maciej Liśkiewicz, and Johannes Textor. Constructing separators and adjustment sets in ancestral graphs. In *Proceedings of the UAI 2014 Conference on Causal Inference: Learning and Prediction-Volume 1274*, pp. 11–24, 2014.

T. Verma and J. Pearl. Causal networks: Semantics and expressiveness. In *Proceedings of the Fourth Workshop on Uncertainty in Artificial Intelligence*, pp. 352–359, Mountain View, CA, 1988.

B Widrow. Pattern-recognizing control systems. *Compurter and Information Sciences*, 1964.

Xingrui Yu, Yueming Lyu, and Ivor Tsang. Intrinsic reward driven imitation learning via generative model. In *International Conference on Machine Learning*, pp. 10925–10935. PMLR, 2020.

J. Zhang and E. Bareinboim. Markov decision processes with unobserved confounders: A causal approach. Technical Report R-23, Purdue AI Lab, 2016.

J. Zhang and E. Bareinboim. Can humans be out of the loop? Technical Report R-64, Causal Artificial Intelligence Lab, Columbia University, 2020. Also, to appear: Proc. of the 1st Conference on Causal Learning and Reasoning (CLeaR), 2022.

J. Zhang, D. Kumor, and E. Bareinboim. Causal imitation learning with unobserved confounders. *Advances in Neural Information Processing Systems*, 33:12263–12274, 2020.

Yao Zhang and Mihaela van der Schaar. Gradient regularized v-learning for dynamic treatment regimes. *Advances in Neural Information Processing Systems*, 33:2245–2256, 2020.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pp. 1433–1438. Chicago, IL, USA, 2008.

A Derivation of IRL Policies

In this section, we provide detailed derivation of IRL policies in Introduction (Sec. 1).

Fig. 1a Let $\mathscr{F}$ be a set of reward function $\mathscr{F} = \{f_Y(x,z) = \alpha x + \beta z - \gamma xz \mid 0 < \alpha < \gamma\}$. Fix any reward function $f_Y \in \mathbb{R}$. We have

$$\min_{\pi} \mathbb{E}[f_Y(X,Z)] - \mathbb{E}[f_Y(X,Z) \mid \text{do}(\pi)] = \mathbb{E}[f_Y(X,Z)] - \max_{\pi} \mathbb{E}[f_Y(X,Z) \mid \text{do}(\pi)]. \tag{10}$$

That is, once a reward function f_Y is fixed, the IRL policy is an optimal policy $X \leftarrow \pi^*(Z)$ maximizing the reward f_Y. Note that for any policy $X \leftarrow \pi(Z)$,

$$\mathbb{E}[f_Y(X,Z) \mid \text{do}(\pi)] = \mathbb{E}[\alpha X + \beta Z - \gamma XZ \mid \text{do}(\pi)] \tag{11}$$

$$= \mathbb{E}[\alpha\pi(Z) + \beta Z - \gamma\pi(Z)Z] \tag{12}$$

For $Z = 0$, since $\alpha > 0$, the optimal action $\pi^*(Z=0)$ is thus given by

$$\pi^*(Z=0) = \arg\max_{x \in \{0,1\}} \mathbb{E}[\alpha x] = 1 \tag{13}$$

Similarly, for $Z = 1$, the optimal action $\pi^*(Z=1)$ is given by

$$\pi^*(Z=1) = \arg\max_{x \in \{0,1\}} \mathbb{E}[\alpha x + \beta - \gamma x] = \arg\max_{x \in \{0,1\}} \mathbb{E}[(\alpha - \gamma)x + \beta] = 0. \tag{14}$$

The above solution follows from $\alpha < \gamma$. The above equations together imply that the optimal policy $\pi^* : X \leftarrow \neg Z$. Finally, since π^* is the solution of the outer minimization step for any reward function $f_Y \in \mathscr{F}$, it is verifiable that π^* must be the policy obtained by IRL algorithms.

Fig. 1b For any reward function in the hypothesis class $\mathscr{F} = \{f_Y(x,z) = \alpha \mid \forall \alpha > 0\}$,

$$\mathbb{E}[f_Y(X,Z)] - \mathbb{E}[f_Y(X,Z) \mid \text{do}(\pi)] = \mathbb{E}[\alpha] - \mathbb{E}[\alpha \mid \text{do}(\pi)] = 0. \tag{15}$$

This means that any policy π^* mapping from X to Z could be a solution for the minimax problem $\min_\pi \max_{f_Y \in \mathscr{F}} \mathbb{E}[f_Y(X,Z)] - \mathbb{E}[f_Y(X,Z) \mid \text{do}(\pi)]$. For an arbitrary policy $\pi(X|Z)$,

$$\mathbb{E}[Y \mid \text{do}(\pi)] = \sum_{x,z} \mathbb{E}[Y \mid \text{do}(x), z]\pi(x|z)P(z). \tag{16}$$

Among quantities in the above equation, for any x, z,

$$\mathbb{E}[Y \mid \text{do}(x), z] = \mathbb{E}[\neg X \oplus Z \oplus U_2 \mid \text{do}(x), z] = \mathbb{E}[\neg x \oplus z \oplus U_2] = 0.5. \tag{17}$$

The last step follows since U_1, U_2 are both uniformly drawn over $\{0,1\}$ and $Z \leftarrow U_1 \oplus U_2$. This implies that the expected reward of any policy π is equal to

$$\mathbb{E}[Y \mid \text{do}(\pi)] = 0.5 \sum_{x,z} \pi(x|z)P(z) = 0.5. \tag{18}$$

Therefore, we must have $\mathbb{E}[Y \mid \text{do}(\pi^*)] = 0.5$ where π^* is a solution obtained by IRL algorithms.

B Proofs

In this section, we provide proofs for the theoretical results presented in the paper. Throughout this paper, detailed parametrizations of the underlying SCM M are assumed to be unknown to the agent. Instead, the agent has access to a causal diagram $\mathcal{G}$ associated with M, and observed trajectories of the expert's demonstrations, summarized as the observational distribution $P(\boldsymbol{O})$. These theoretical assumptions are required for the our proposed causal IRL algorithm, which delineate the technical limitations of the work.

Theorem 1. *Given a causal diagram $\mathcal{G}$, if there exists a minimal π-backdoor admissible scope $\mathcal{S} = \{\langle X_i, \boldsymbol{Z}_i\rangle\}_{i=1}^n$ in $\mathcal{G}$, consider the following conditions:*

1. *Let effective actions $\boldsymbol{X}^* = \boldsymbol{X} \cap An(Y)_{\mathcal{G}_{\mathcal{S}}}$ and effective covariates $\boldsymbol{Z}^* = \bigcup_{X_i \in \boldsymbol{X}^*} \boldsymbol{Z}_i$;*
2. *For $i = 1, \ldots, n+1$, let $\boldsymbol{X}^*_{<i} = \{\forall X_j \in \boldsymbol{X}^* \mid j < i\}$ and $\boldsymbol{Z}^*_{<i} = \bigcup_{X_j \in \boldsymbol{X}^*_{<i}} \boldsymbol{Z}_j$.*

Then, for any policy $\boldsymbol{\pi} \sim \mathcal{S}$, the expected reward $\mathbb{E}[Y \mid do(\boldsymbol{\pi})]$ is computable from $P(\boldsymbol{O}, Y)$ as:

$$\mathbb{E}[Y \mid do(\boldsymbol{\pi})] = \sum_{\boldsymbol{x}^*, \boldsymbol{z}^*} \mathbb{E}[Y \mid \boldsymbol{x}^*, \boldsymbol{z}^*] \rho_{\boldsymbol{\pi}}(\boldsymbol{x}^*, \boldsymbol{z}^*) \tag{2}$$

where the occupancy measure $\rho_{\boldsymbol{\pi}}(\boldsymbol{x}^*, \boldsymbol{z}^*) = \prod_{X_i \in \boldsymbol{X}^*} P\left(\boldsymbol{z}_i \mid \boldsymbol{x}^*_{<i}, \boldsymbol{z}^*_{<i}\right) \pi_i(x_i \mid \boldsymbol{z}_i)$.

Proof. Without loss of generality, assume that $\boldsymbol{X} = \boldsymbol{X}^*$, i.e., all actions in $\boldsymbol{X}$ are active after policy $\boldsymbol{\pi} \sim \mathcal{S}$ is deployed. Otherwise, we could always reduce the action set $\boldsymbol{X}$ to the active actions $\boldsymbol{X}^*$ since $P(Y \mid \text{do}(\boldsymbol{\pi})) = P(Y \mid \text{do}(\boldsymbol{\pi}_{\boldsymbol{X}^*}))$ where $\boldsymbol{\pi}_{\boldsymbol{X}^*}$ is sequence of decision rules $\{\pi_i \mid X_i \in \boldsymbol{X}^*\}$ with restriction to $\boldsymbol{X}^*$.

For any $i = 1, \ldots, n$, let $\boldsymbol{X}_i = \{X_1, \ldots, X_{i-1}\}$ and $\boldsymbol{X}_{\leq i} = \boldsymbol{X}_{<i} \cup \{X_i\}$. Similarly, we define $\boldsymbol{Z}_{<i} = \bigcup_{j<i} \boldsymbol{Z}_j$ and $\boldsymbol{Z}_{\leq i} = \boldsymbol{Z}_{<i} \cup \boldsymbol{Z}_i$. Let $\boldsymbol{\pi}_{\geq i}$ be a sequence of decision rules $\{\pi_i, \pi_{i+1}, \ldots, \pi_n\}$ and $\boldsymbol{\pi}_{>i} = \boldsymbol{\pi}_{\geq i} \setminus \{\pi_i\}$. It is sufficient to show that for any $i = 1, \ldots, n$,

$$P(y, \boldsymbol{x}_{<i}, \boldsymbol{z}_{<i} \mid \text{do}(\boldsymbol{\pi}_{\geq i})) = \sum_{x_i, \boldsymbol{z}_i} P(y, \boldsymbol{x}_{\leq i}, \boldsymbol{z}_{\leq i} \mid \text{do}(\boldsymbol{\pi}_{>i})) \frac{\pi_i(x_i|\boldsymbol{z}_i)}{P(x_i|\boldsymbol{x}_{<i}, \boldsymbol{z}_{\leq i})} \tag{19}$$

By basic probabilistic operations,

$$P(y, \boldsymbol{x}_{<i}, \boldsymbol{z}_{<i} \mid \text{do}(\boldsymbol{\pi}_{\geq i})) = \sum_{x_i, \boldsymbol{z}_i} P(y, \boldsymbol{x}_{\leq i}, \boldsymbol{z}_{\leq i} \mid \text{do}(\boldsymbol{\pi}_{\geq i})) \tag{20}$$

$$= \sum_{x_i, \boldsymbol{z}_i} P(y \mid \boldsymbol{x}_{\leq i}, \boldsymbol{z}_{\leq i}, \text{do}(\boldsymbol{\pi}_{\geq i})) P(x_i \mid \boldsymbol{x}_{<i}, \boldsymbol{z}_{\leq i}, \text{do}(\boldsymbol{\pi}_{\geq i}) P(\boldsymbol{x}_{<i}, \boldsymbol{z}_{\leq i} \mid \text{do}(\boldsymbol{\pi}_{\geq i})) \tag{21}$$

$$= \sum_{x_i, \boldsymbol{z}_i} P(y \mid \text{do}(x_i), \boldsymbol{x}_{<i}, \boldsymbol{z}_{\leq i}, \text{do}(\boldsymbol{\pi}_{>i})) \pi_i(x_i|\boldsymbol{z}_i) P(\boldsymbol{x}_{<i}, \boldsymbol{z}_{\leq i}) \tag{22}$$

In the last step, $P(\boldsymbol{x}_{<i}, \boldsymbol{z}_{\leq i} \mid \text{do}(\boldsymbol{\pi}_{\geq i})) = P(\boldsymbol{x}_{<i}, \boldsymbol{z}_{\leq i})$ since all causal diagrams $\mathcal{G}^{(i)}$ are directed acyclic graph; and $\boldsymbol{X}_{<i}, \boldsymbol{Z}_{\leq i}$ are non-descendants of $X_i, X_{i+1}, \ldots, X_n$. Next we will show that

$$P(y \mid \text{do}(x_i), \boldsymbol{x}_{<i}, \boldsymbol{z}_{\leq i}, \text{do}(\boldsymbol{\pi}_{>i})) = P(y \mid x_i, \boldsymbol{x}_{<i}, \boldsymbol{z}_{\leq i}, \text{do}(\boldsymbol{\pi}_{>i})) \tag{23}$$

This is equivalent to show that

$$(Y \perp X_i | \boldsymbol{X}_{<i}, \boldsymbol{Z}_{\leq i}) \text{ in } \mathcal{G}^{(i)}_{\underline{X_i}} \tag{24}$$

First, by definition of sequential π-backdoor,

$$(Y \perp X_i | \boldsymbol{Z}_i) \text{ in } \mathcal{G}^{(i)}_{\underline{X_i}} \tag{25}$$

Suppose Eq. (24) does not hold. We must have a node $W \in \boldsymbol{Z}_{<i} \cup \{X_i\}$ opening a collider path from X_i to Y. Assume that there exists a directed path from W to Y in $\mathcal{G}^{(i)}_{\underline{X_i}}$. Then one could construct backdoor path from X_i to Y via node W, which contradicts Eq. (25).

What remains is to show that there must exist a directed path from W to Y in $\mathcal{G}^{(i)}_{\underline{X_i}}$. Since $\mathcal{S}$ is minimal π-backdoor admissible, it is verifiable from (van der Zander et al., 2014, Lem. 3.4) that $W \in An(\{Y\})$ in $\mathcal{G}^{(i)}$. Consequently, the only case that $W \notin An(\{Y\})$ in $\mathcal{G}^{(i)}_{\underline{X_i}}$ is when $W \in An(\{X_i\})$ in $\mathcal{G}^{(i)}$. Again, this allows us to construct a backdoor path between X_i and Y via W in $\mathcal{G}^{(i)}$, which contradicts Eq. (25). □

Proposition 1. *For a hypothesis class $\mathscr{R} = \{r = \boldsymbol{w} \cdot \boldsymbol{\phi} \mid \boldsymbol{w} \in \mathbb{S}^k\}$, the solution ν^* of the canonical equation in Eq.* (3) *is obtainable by solving the following minimax problem:*

$$\nu^* = \min_{\boldsymbol{\pi} \sim \mathcal{S}} \max_{\boldsymbol{w} \in \mathbb{S}^k} \boldsymbol{w}^\top \boldsymbol{G} \boldsymbol{\pi}, \tag{4}$$

where $\boldsymbol{G}$ is a $k \times n$ matrix given by $\boldsymbol{G}(i, j) = \sum_{\boldsymbol{x}^*, \boldsymbol{z}^*} \boldsymbol{\phi}^{(i)}(\boldsymbol{x}^*, \boldsymbol{z}^*) \left(\rho(\boldsymbol{x}^*, \boldsymbol{z}^*) - \rho_{\boldsymbol{\pi}^{(j)}}(\boldsymbol{x}^*, \boldsymbol{z}^*)\right)$.

Proof. Note that $\boldsymbol{\pi}$ is a *mixed policy* which is a probability distribution over deterministic policies compatible with scope $\mathcal{S}$. With a slight abuse of notation, we will treat $\boldsymbol{\pi}$ as a vector where $\boldsymbol{\pi}(i)$ is the probability assigned to the i-th deterministic policy $\boldsymbol{\pi}^{(i)}$. The claimed statement follows immediately from the definition of the game matrix $\boldsymbol{G}$. □

Proposition 2. *For a hypothesis class $\mathscr{R} = \{r : \mathscr{D}_{\boldsymbol{X}^*} \times \mathscr{D}_{\boldsymbol{Z}^*} \mapsto \mathbb{R}\}$ regularized by ψ, the solution ν^* of the penalized canonical equation in Eq.* (5) *is obtainable by solving the following problem:*

$$\nu^* = \min_{\boldsymbol{\pi} \sim \mathcal{S}} \psi^* \left(\rho - \rho_{\boldsymbol{\pi}}\right) \tag{6}$$

where ψ^* be a conjugate function of ψ and is given by $\psi^* = \max_{r \in \mathbb{R}^{\boldsymbol{x} \times \boldsymbol{z}}} a^\top r - \psi(r)$.

Proof. Let the vector $\alpha = \rho - \rho_{\boldsymbol{\pi}}$. The claimed statement follows from the definition of the conjugate function ψ^*. □

Theorem 2. *Given a causal diagram $\mathcal{G}$, a policy scope $\mathcal{S}$ is identifiable from $P(\boldsymbol{O}, Y)$ in $\mathcal{G}$ if and only if for any policy $\boldsymbol{\pi} \sim \mathcal{S}$, the expected reward $\mathbb{E}[Y \mid do(\boldsymbol{\pi})]$ is computable from $P(\boldsymbol{O}, Y)$ as*

$$\mathbb{E}[Y \mid do(\boldsymbol{\pi})] = \sum_{\boldsymbol{x}^*, \boldsymbol{z}^*} \mathbb{E}[Y \mid \boldsymbol{x}^*, \boldsymbol{z}^*] \rho_{\boldsymbol{\pi}}(\boldsymbol{x}^*, \boldsymbol{z}^*) \tag{7}$$

where subsets $\boldsymbol{X}^* \subseteq \boldsymbol{X}$, $\boldsymbol{Z}^* \subseteq \boldsymbol{O} \setminus \boldsymbol{X}$; and the imitator's occupancy measure $\rho_{\boldsymbol{\pi}}(\boldsymbol{x}^*, \boldsymbol{z}^*)$ is a function of the observational distribution $P(\boldsymbol{O})$ and policy $\boldsymbol{\pi}$.

Proof. the claimed statement follows from Lem. 1. □

Lemma 1. *Given a causal diagram $\mathcal{G}$, a policy scope $\mathcal{S}$ is identifiable from $P(\boldsymbol{O}, Y)$ in $\mathcal{G}$ if and only if* IDENTIFY$(\mathcal{G}, Y, \mathcal{S}) \neq$ "FAIL". *Moreover,* IDENTIFY$(\mathcal{G}, Y, \mathcal{S})$ *returns an identification formula of the form in Eq.* (7) *where $\boldsymbol{X}^* = Pa(\boldsymbol{C}_Y) \cap \boldsymbol{X}$ and $\boldsymbol{Z}^* = Pa(\boldsymbol{C}_Y) \setminus (\{Y\} \cup \boldsymbol{X})$; and $\boldsymbol{C}_Y$ is a c-component containing reward Y in subgraph $\mathcal{G}_{[An(\boldsymbol{Z}_Y)]}$.*

Proof. We first review the identification algorithm for effects of policies with scope $\mathcal{S}$ in the causal diagram $\mathcal{G}$ (Tian, 2008). Let $\boldsymbol{D} = An(\{Y\}) \setminus \boldsymbol{X}$ in the induced diagram $\mathcal{G}$. For any policy $\boldsymbol{\pi} \sim \mathcal{S}$, $P(Y \mid \text{do}(\boldsymbol{\pi}))$ could be written as, following (Tian, 2008, Eq. 15):

$$P(Y \mid \text{do}(\boldsymbol{\pi})) = \sum_{\boldsymbol{x}, \boldsymbol{d} \setminus \{y\}} Q[\boldsymbol{D}] \prod_{X_i \in \boldsymbol{X}} \pi_i(x_i \mid \boldsymbol{z}_i). \tag{26}$$

It has been shown that effects $P(Y \mid \text{do}(\boldsymbol{\pi}))$ of all policies $\boldsymbol{\pi} \sim \mathcal{S}$ are identifiable if and only if $Q[\boldsymbol{D}]$ is identifiable from $P(\boldsymbol{O}, Y)$ in $\mathcal{G}$ (Tian, 2008; Correa & Bareinboim, 2019). Let $\boldsymbol{D}_1, \ldots, \boldsymbol{D}_m$ denote c-components in $\mathcal{G}_{[\boldsymbol{D}]}$; and let $\boldsymbol{D}_Y$ denote the c-component in $\mathcal{G}_{[\boldsymbol{D}]}$ containing Y and let $\boldsymbol{D}_1, \ldots, \boldsymbol{D}_m$ be other c-components. $Q[\boldsymbol{D}]$ could be written as ((Tian, 2002, Lemma 11)):

$$Q[\boldsymbol{D}] = Q[\boldsymbol{D}_Y] \prod_{i=1,\ldots,m} Q[\boldsymbol{D}_i]. \tag{27}$$

Similarly, let $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_n$ denote c-components in the diagram $\mathcal{G}$. The algorithm then attempts to identify each c-factor $Q[\boldsymbol{D}_i]$ from $Q[\boldsymbol{S}_j]$ where $\boldsymbol{S}_j$ is a c-component in $\mathcal{G}$ containing variables in $\boldsymbol{D}_i$. This procedure could be done by applying the IDENTIFY algorithm introduced in (Tian, 2002), which was later shown to be complete (Huang & Valtorta, 2006; Shpitser, 2008). Details of algorithm IDENTIFY are described in Alg. 2.

Algorithm 2: IDENTIFY

Input: c-components $\boldsymbol{C} \subseteq \boldsymbol{T}$, a causal diagram $\mathcal{G}$.
Output: expression for $Q[\boldsymbol{C}]$ in terms of $Q[\boldsymbol{T}]$ or fail to determine.
Let $\boldsymbol{A} = An(\boldsymbol{C})_{\mathcal{G}[\boldsymbol{T}]}$.
if $\boldsymbol{A} = \boldsymbol{T}$ **then** Output $Q[\boldsymbol{C}] = \sum_{\boldsymbol{t} \setminus \boldsymbol{c}} Q[\boldsymbol{T}]$.
end if
if $\boldsymbol{A} = \boldsymbol{C}$ **then** Output "FAIL".
end if
if $\boldsymbol{C} \subset \boldsymbol{A} \subset \boldsymbol{T}$ **then**
    Assume that in $\mathcal{G}[\boldsymbol{A}]$, $\boldsymbol{C}$ is contained in a c-component $\boldsymbol{T}'$.
    Compute $Q[\boldsymbol{T}']$ from $Q[\boldsymbol{A}] = \sum_{\boldsymbol{t} \setminus \boldsymbol{a}} Q[\boldsymbol{T}]$ by (Tian, 2002, Lemma 11).
    Output IDENTIFY$(\boldsymbol{C}, \boldsymbol{T}', \mathcal{G})$.
end if

Since all variables in $\boldsymbol{D} \setminus \{Y\}$ are non-descendant of Y in $\mathcal{G}$, it is verifiable that the identification formula for each $Q[\boldsymbol{D}_i]$ is a function of distribution $P(\boldsymbol{O})$, not including reward Y. Let $\boldsymbol{S}_Y$ denote a c-component in $\mathcal{G}$ that contains $\boldsymbol{D}_Y$. Therefore, it is sufficient to show that c-factor $Q[\boldsymbol{D}_Y]$ is identifiable from $Q[\boldsymbol{S}_Y]$ as follows:

$$Q[\boldsymbol{D}_Y] = \sum_{\boldsymbol{s}} P(y \mid \boldsymbol{x}^*, \boldsymbol{z}^*) g(P(\boldsymbol{O})), \tag{28}$$

where $\boldsymbol{S}$ is a certain subset in $\boldsymbol{O} \setminus Pa(\boldsymbol{D}_Y)$; and $g(P(\boldsymbol{O}))$ is a function consisting of a series of arithmetic operation on the observational distribution $P(\boldsymbol{O})$.

We next show that $\text{IDENTIFY}(\boldsymbol{D}_Y, \boldsymbol{S}_Y, \mathcal{G})$ returns an identification formula of the form Eq. (28). If $\boldsymbol{S}_Y = An(\boldsymbol{D}_Y)_{\mathcal{G}_{[\boldsymbol{S}_Y]}}$ (Step 3), the algorithm returns an expression $Q[\boldsymbol{D}_Y] = \sum_{\boldsymbol{s}_Y \setminus \boldsymbol{d}_Y} Q[\boldsymbol{S}_Y]$. (Tian, 2002, Lemma 7) shows that $Q[\boldsymbol{S}_Y]$ is given by, following a topological ordering $\prec$ in $\mathcal{G}$:

$$Q[\boldsymbol{S}_Y] = \prod_{V \in \boldsymbol{S}_Y} P(v \mid \boldsymbol{pa}_V^+) \tag{29}$$

where $\boldsymbol{pa}^+ = Pa(\boldsymbol{C}_V) \setminus \{V\}$; and $\boldsymbol{C}_V$ is a c-component in subgraph $\mathcal{G}_{[\{V' \in \boldsymbol{V} | V' \prec V\}]}$ containing V. Summing over variables in $\boldsymbol{S}_Y \setminus \boldsymbol{D}_Y$ we obtain Eq. (28).

We next consider cases where the condition in Step 3 is not satisfied; and $\text{IDENTIFY}(\boldsymbol{D}_Y, \boldsymbol{S}_Y, \mathcal{G})$ has to enter the recursive step in Step 10. We will utilize the following claim.

Claim 1. *At each recursive call in* $\text{IDENTIFY}(\boldsymbol{D}_Y, \boldsymbol{S}_Y, \mathcal{G})$ *(Step 10),* $Q[\boldsymbol{T}']$ *could be written as:*

$$Q[\boldsymbol{T}'] = \sum_{\boldsymbol{s}} P(y \mid \boldsymbol{x}^*, \boldsymbol{z}^*) g(P(\boldsymbol{O})), \tag{30}$$

where $\boldsymbol{S}$ *is a certain subset in* $\boldsymbol{O} \setminus Pa(\boldsymbol{D}_Y)$*; and* $g(P(\boldsymbol{O}))$ *function consisting of a series of arithmetic operation on* $P(\boldsymbol{O})$*. We note that* $\boldsymbol{S}$ *and* g *could be different for each different* $\boldsymbol{T}'$*.*

The above claim allows us to derive the identification formula of the form Eq. (28). Since if $Q[\boldsymbol{D}_Y]$ is identifiable, it must enter Step 3 after some recursive steps. IDENTIFY algorithm returns a formula $Q[\boldsymbol{D}_Y] = \sum_{\boldsymbol{t} \setminus \boldsymbol{c}} Q[\boldsymbol{T}]$ where $\boldsymbol{A} = An(\boldsymbol{D}_Y)_{\mathcal{G}[\boldsymbol{T}]}$. Since $Q[\boldsymbol{T}]$ takes the form of Eq. (30), we obtain the claimed statement Eq. (28).

We will prove Claim 1 by induction on the number of recursive call m in $\text{IDENTIFY}(\boldsymbol{C}, \boldsymbol{T}', \mathcal{G})$.

Base case $m = 1$**.** This follows immediately from (Tian, 2002, Lems. 10 and 11). At Step 9, $Q[\boldsymbol{A}] = \sum_{\boldsymbol{s}_Y \setminus \boldsymbol{a}} Q[\boldsymbol{S}_Y]$ and $Q[\boldsymbol{S}_Y]$ is given by $Eq.$ (29). This implies $Q[\boldsymbol{S}]$ could be written in the form of:

$$Q[\boldsymbol{A}] = \sum_{\boldsymbol{s}} P(y \mid \boldsymbol{x}^*, \boldsymbol{z}^*) g(P(\boldsymbol{O})) \tag{31}$$

For any variable $V \in \boldsymbol{A}$, let $\boldsymbol{A}_{\prec V}$ denote variables in $\boldsymbol{A}$ that precedes V following the topological ordering $\prec$ in $\mathcal{G}$. (Tian, 2002, Lemma 11) shows that

$$Q[\boldsymbol{T}'] = \prod_{V \in \boldsymbol{T}'} \frac{Q[V, \boldsymbol{A}_{\prec V}]}{Q[\boldsymbol{A}_{\prec V}]} = \prod_{V \in \boldsymbol{T}'} \frac{Q[V, \boldsymbol{A}_{\prec V}]}{\sum_v Q[V, \boldsymbol{A}_{\prec V}]} \tag{32}$$

Since all variables in $\boldsymbol{A} \setminus \{Y\}$ precedes Y according to the topological ordering $\prec$, we have $Y \notin \boldsymbol{A}_{\prec V}$ for every $V \in \boldsymbol{A} \setminus \{Y\}$. This implies $Q[\boldsymbol{T}']$ could be written as

$$Q[\boldsymbol{T}'] = \frac{Q[Y, \boldsymbol{A}_{\prec Y}]}{\sum_y Q[Y, \boldsymbol{A}_{\prec Y}]} \prod_{V \in \boldsymbol{T}' \setminus \{Y\}} \frac{Q[V, \boldsymbol{A}_{\prec V}]}{\sum_v Q[V, \boldsymbol{A}_{\prec V}]} \tag{33}$$

$$= \frac{\sum_{a \setminus (\{y\} \cup \boldsymbol{a}_{\prec Y})} Q[\boldsymbol{A}]}{\sum_{a \setminus \boldsymbol{a}_{\prec Y}} Q[\boldsymbol{A}]} g'(P(\boldsymbol{O})) \tag{34}$$

The last step follows from (Tian, 2002, Lemma 10). Replacing $Q[\boldsymbol{A}]$ with Eq. (31) we obtain

$$Q[\boldsymbol{T}'] = \frac{\sum_{a\setminus(\{y\}\cup \boldsymbol{a}_{\prec Y})}\sum_{\boldsymbol{s}} P(y \mid \boldsymbol{x}^*, \boldsymbol{z}^*)g(P(\boldsymbol{O}))}{\sum_{a\setminus \boldsymbol{a}_{\prec Y}}\sum_{\boldsymbol{s}} P(y \mid \boldsymbol{x}^*, \boldsymbol{z}^*)g(P(\boldsymbol{O}))} g'(P(\boldsymbol{O})) \tag{35}$$

$$= P(y \mid \boldsymbol{x}^*, \boldsymbol{z}^*)\frac{\sum_{a\setminus(\{y\}\cup \boldsymbol{a}_{\prec Y})}\sum_{\boldsymbol{s}} g(P(\boldsymbol{O}))}{\sum_{a\setminus \boldsymbol{a}_{\prec Y}}\sum_{\boldsymbol{s}} P(y \mid \boldsymbol{x}^*, \boldsymbol{z}^*)g(P(\boldsymbol{O}))} g'(P(\boldsymbol{O})) \tag{36}$$

$$= P(y \mid \boldsymbol{x}^*, \boldsymbol{z}^*)\frac{\sum_{a\setminus(\{y\}\cup \boldsymbol{a}_{\prec Y})}\sum_{\boldsymbol{s}} g(P(\boldsymbol{O}))}{\sum_{a\setminus(\{y\}\cup \boldsymbol{a}_{\prec Y})}\sum_{\boldsymbol{s}} g(P(\boldsymbol{O}))} g'(P(\boldsymbol{O})) \tag{37}$$

$$= P(y \mid \boldsymbol{x}^*, \boldsymbol{z}^*)g''(P(\boldsymbol{O})) \tag{38}$$

where g'' is a series of arithmetic operations of $P(\boldsymbol{O})$. Eq. (37) holds since $g(P(\boldsymbol{O}))$ is not a function of reward signal Y. This proves Eq. (30) for the base case $m = 1$.

Induction case $m = k + 1$**.** Suppose the result holds for $m \leq k$ where $k \geq 1$ and consider the case $m = k + 1$. At Step 9, $Q[\boldsymbol{A}] = \sum_{\boldsymbol{t}\setminus \boldsymbol{a}} Q[\boldsymbol{T}]$. By the induction assumption, $Q[\boldsymbol{T}]$ could be written as Eq. (30). This implies that $Q[\boldsymbol{S}]$ could be written in the form of Eq. (31). Again, following the derivation in the base case $m = 1$, we prove that Eq. (30) for the induction case $m = k + 1$.

Let $g_i(P(\boldsymbol{O}))$ denote the identification formula of $Q[\boldsymbol{D}_i]$ for every c-component $\boldsymbol{D}_i$ in $\mathcal{G}_{[\boldsymbol{D}]}$. Eqs. (26) to (28) together imply that for any policy $\boldsymbol{\pi} \sim \mathcal{S}$,

$$\mathbb{E}[Y \mid \text{do}(\boldsymbol{\pi})] = \sum_{\boldsymbol{x},\boldsymbol{d}\setminus\{y\}}\sum_{\boldsymbol{s}} \mathbb{E}[Y \mid \boldsymbol{x}^*, \boldsymbol{z}^*]g(P(\boldsymbol{O}))\prod_{i=1,\dots,m} g_i(P(\boldsymbol{O}))\prod_{X_i\in\boldsymbol{X}} \pi_i(x_i \mid \boldsymbol{z}_i) \tag{39}$$

Simplifying the above equation gives

$$\mathbb{E}[Y \mid \text{do}(\boldsymbol{\pi})] = \sum_{\boldsymbol{s}'} \mathbb{E}[Y \mid \boldsymbol{x}^*, \boldsymbol{z}^*]\rho_{\boldsymbol{\pi}}(\boldsymbol{x}^*, \boldsymbol{z}^*) \tag{40}$$

where $\boldsymbol{S}'$ is a certain subset of $\boldsymbol{X}^* \cup \boldsymbol{Z}^*$; and $\rho_{\boldsymbol{\pi}}$ is a function of policy $\boldsymbol{\pi}$ and the observational distribution $P(\boldsymbol{O})$. Note that it is possible that $\boldsymbol{T} = (\boldsymbol{X}^* \cup \boldsymbol{Z}^*) \setminus \boldsymbol{S}' \neq \emptyset$. In that case, one could fix $\boldsymbol{T}$ to an arbitrary constant $\boldsymbol{t}$ by multiplying the indicator function $\mathbb{1}\{\boldsymbol{T} = \boldsymbol{t}\}$. □

Lemma 2. *Given a causal diagram $\mathcal{G}$, for policy scopes $\mathcal{S}' \subseteq \mathcal{S}$, $\mathcal{S}'$ is identifiable from distribution $P(\boldsymbol{O}, Y)$ in $\mathcal{G}$ if $\mathcal{S}$ is identifiable from $P(\boldsymbol{O}, Y)$ in $\mathcal{G}$.*

Proof. If a scope $\mathcal{S}$ is identifiable, then the expected rewards $\mathbb{E}[Y \mid \text{do}(\boldsymbol{\pi})]$ of all policies $\boldsymbol{\pi}$ with scope $\mathcal{S}$ are identifiable from $P(\boldsymbol{O}, Y)$ in $\mathcal{G}$. This subsumes all policies associated with the subscope $\mathcal{S}' \subseteq \mathcal{S}$. Therefore, $\mathcal{S}'$ is identifiable from $P(\boldsymbol{O}, Y)$ in $\mathcal{G}$. □

Theorem 3. *For a causal diagram $\mathcal{G}$ and a reward Y,* ListIDScope$(\mathcal{G}, Y, \emptyset, \bigcup_{i=1}^{n} \boldsymbol{PA}_i^*)$ *enumerates a subset $\mathbb{S}^* \subseteq \mathbb{S}$ so that for any $\boldsymbol{\pi} \sim \mathbb{S}$, there is $\boldsymbol{\pi}^* \sim \mathbb{S}^*$ where $\mathbb{E}[Y \mid do(\boldsymbol{\pi})] = \mathbb{E}[Y \mid do(\boldsymbol{\pi}^*)]$.*

Proof. The recursive calls at Steps 8 and 9 guarantee that ListIDScope generates every identifiable scope $\mathcal{S}$ in $\mathbb{S}^*$ exactly once where

$$\mathbb{S}^* = \left\{ \mathcal{S} = \{\langle X_i, \boldsymbol{Z}_i\rangle\}_{i=1}^{n} \text{ is identifiable} \mid \boldsymbol{Z}_i = \boldsymbol{PA}_i^* \cap \boldsymbol{Z},\ \forall \boldsymbol{Z} \subseteq \bigcup_{i=1}^{n} \boldsymbol{PA}_i^* \right\} \tag{41}$$

The pruning criteria in Steps 3 and 6 only remove branches containing exclusively only identifiable (or non-identifiable) scopes (Lem. 2). In that case, it is sufficient to discard all such subscopes (non-identifiable) or only return the root scope $\mathcal{S}_r$ of the corresponding branch (identifiable).

What remains is to show the following claim:

Claim 2. *For any scope $\mathcal{S} \in \mathbb{S}$, there exists $\mathcal{S}^* \in \mathbb{S}^*$ such that $\mathcal{S}^*$* perfectly mimics *$\mathcal{S}$: that is, for any $\boldsymbol{\pi} \sim \mathcal{S}$, one could find $\boldsymbol{\pi}^* \sim \mathcal{S}^*$ so that $\mathbb{E}[Y \mid do(\boldsymbol{\pi})] = \mathbb{E}[Y \mid do(\boldsymbol{\pi}^*)]$.*

Algorithm 3: Causal MWAL

Input: $\mathcal{G}$, Expert demonstrations τ_E
Output: an imitating policy $\boldsymbol{\pi}^*$
Apply Thm. 1 (or Thm. 2) to obtain formulas for the expert's $\rho(\boldsymbol{x}^*, \boldsymbol{z}^*)$ and imitator's occupancy measure $\rho_{\pi_j}(\boldsymbol{x}^*, \boldsymbol{z}^*)$
Let $\boldsymbol{\mu}(i) = \sum_{\boldsymbol{x}^*, \boldsymbol{z}^*} \phi^{(i)}(\boldsymbol{x}^*, \boldsymbol{z}^*)\rho(\boldsymbol{x}^*, \boldsymbol{z}^*)$ and let $\boldsymbol{\mu}_{\boldsymbol{\pi}}(i) = \sum_{\boldsymbol{x}^*, \boldsymbol{z}^*} \phi^{(i)}(\boldsymbol{x}^*, \boldsymbol{z}^*)\rho_{\boldsymbol{\pi}}(\boldsymbol{x}^*, \boldsymbol{z}^*)$
Let $\widetilde{\boldsymbol{G}}(i, \boldsymbol{\mu}_{\boldsymbol{\pi}}) = ((\hat{\boldsymbol{\mu}}(i) - \boldsymbol{\mu}_{\boldsymbol{\pi}}(i)) - 2)/4$
Let $\beta = \left(1 + \sqrt{\frac{2\ln(k)}{J}}\right)^{-1}$
Initialize $\boldsymbol{w}^1(i) = 1$ for $i = 1, ..., k$
for iteration $j = 0, 1, 2, \ldots, J$ **do**
    Set $\boldsymbol{w}^j(i) = \frac{\boldsymbol{w}^j(i)}{\sum_i \boldsymbol{w}^j(i)}$ for $i = 1, ..., k$
    Compute the policy $\hat{\boldsymbol{\pi}}_j$ by $\arg\min_{\boldsymbol{\pi}} \boldsymbol{w}^\top \widetilde{\boldsymbol{G}}\boldsymbol{\pi}$, where $\boldsymbol{w} := \boldsymbol{w}^j$
    Compute $\hat{\boldsymbol{\mu}}_j = \boldsymbol{\mu}_{\hat{\boldsymbol{\pi}}_j}$
    $\boldsymbol{w}^{j+1}(i) = \boldsymbol{w}^j(i) \cdot \exp(\ln(\beta) \cdot \widetilde{\boldsymbol{G}}(i, \hat{\boldsymbol{\mu}}_j))$ for $i = 1, ..., k$
end for
return The mixed policy that has a probability $\frac{1}{J}$ of choosing $\hat{\boldsymbol{\pi}}_j$, for all $t \in \{1, \ldots, J\}$

Algorithm 4: Causal GAIL

Input: $\mathcal{G}$, Expert demonstrations τ_E
Output: an imitating policy $\boldsymbol{\pi}^*$
Apply Thm. 1 (or Thm. 2) to obtain formulas for the expert's $\rho(\boldsymbol{x}^*, \boldsymbol{z}^*)$ and imitator's occupancy measure $\rho_{\boldsymbol{\pi}_j}(\boldsymbol{x}^*, \boldsymbol{z}^*)$
for iteration $j = 0, 1, 2, \ldots$ **do**
    Collect trajectories from distributions $\rho(\boldsymbol{x}^*, \boldsymbol{z}^*)$ and $\rho_{\boldsymbol{\pi}_j}(\boldsymbol{x}^*, \boldsymbol{z}^*)$
    Update the parameters w of discriminator D_j with gradient
$\hat{\mathbb{E}}[\nabla_w \log(D_j(\boldsymbol{x}^*, \boldsymbol{z}^*))] + \hat{\mathbb{E}}_{\boldsymbol{\pi}_j}[\nabla_w \log(1 - D_j(\boldsymbol{x}^*, \boldsymbol{z}^*))]$
    Update the policy $\boldsymbol{\pi}_j = \arg\min_\pi \mathbb{E}_{\boldsymbol{\pi}}[\log(1 - D(\boldsymbol{x}^*, \boldsymbol{z}^*))]$ with policy optimization for DTR
end for
return The learned policy $\boldsymbol{\pi}^*$

Fix any $\mathcal{S} \in \mathbb{S} \setminus \mathbb{S}^*$. Let $\boldsymbol{X}_{\mathcal{S}} = An(Y)_{\mathcal{G}_{\mathcal{S}}} \cap \boldsymbol{X}$ and let $\boldsymbol{Z}_{\mathcal{S}} = An(Y)_{\mathcal{G}_{\mathcal{S}}} \setminus \boldsymbol{X}$. We construct a scope $\mathcal{S}^* = \{\langle X_i, \boldsymbol{Z}_i^* \rangle\}_{i=1}^n$ where $\boldsymbol{Z}_i^* = \boldsymbol{PA}_i^* \cap \left(\bigcup_{X \in \boldsymbol{X}_{\mathcal{S}}} pa(X)_{\mathcal{G}_{\mathcal{S}}}\right)$. Similarly, define $\boldsymbol{X}_{\mathcal{S}^*} = An(Y)_{\mathcal{G}_{\mathcal{S}^*}} \cap \boldsymbol{X}$ and $\boldsymbol{Z}_{\mathcal{S}^*} = An(Y)_{\mathcal{G}_{\mathcal{S}^*}} \setminus \boldsymbol{X}$.

We will next show that $\boldsymbol{X}_{\mathcal{S}} = \boldsymbol{X}_{\mathcal{S}^*}$ and $\boldsymbol{Z}_{\mathcal{S}} = \boldsymbol{Z}_{\mathcal{S}^*}$. Suppose that $\boldsymbol{X}_{\mathcal{S}} \subset \boldsymbol{X}_{\mathcal{S}^*}$. Let X be an action in the set difference $\boldsymbol{X}_{\mathcal{S}^*} \setminus \boldsymbol{X}_{\mathcal{S}}$. Then by the construction of scope $\mathcal{S}^*$, there must exist a directed path from X to Y in $\mathcal{G}_{\mathcal{S}^*}$ via a node in $Pa(X)_{\mathcal{G}_{\mathcal{S}}}$ for some $X \in \boldsymbol{X}_{\mathcal{S}}$. However, such a path must also exist in the diagram $\mathcal{G}_{\mathcal{S}}$. That is, $X \in \boldsymbol{X}_{\mathcal{S}}$, which is a contradiction. Since $\mathcal{G}_{\mathcal{S}^*}$ does not add any new ancestor of Y, we also have $\boldsymbol{Z}_{\mathcal{S}} = \boldsymbol{Z}_{\mathcal{S}^*}$.

It follows immediately from the factorization in (Tian, 2008, Eq. 15) that $\mathcal{S}^*$ perfectly mimics $\mathcal{S}$. Moreover, since $\mathcal{S}$ is identifiable, the identification condition in (Tian, 2008, Thm. 1) (which is later shown to be complete in (Correa & Bareinboim, 2019)) implies that $Q[\boldsymbol{Z}_{\mathcal{S}}]$ must be identifiable from $P(\boldsymbol{O}, Y)$ in $\mathcal{G}$. Since $\boldsymbol{Z}_{\mathcal{S}} = \boldsymbol{Z}_{\mathcal{S}^*}$, applying the identification condition in (Tian, 2008, Thm. 1) again shows that $\mathcal{S}^*$ is identifiable from $P(\boldsymbol{O}, Y)$ in $\mathcal{G}$. That is, $\mathcal{S}^* \in \mathbb{S}^*$, which completes the proof. □

C CAUSAL MWAL AND CAUSAL GAIL

In this section, we provide details of causal MWAL and causal GAIL, including pseudo-code and practical considerations in their implementations.

Causal MWAL Similar to (Syed & Schapire, 2008), we are interested in learning a *mixed policy* $\boldsymbol{\pi}$ compatible with the scope $\mathcal{S}$. Particularly, a mixed policy $\boldsymbol{\pi}^*$ is a probability distribution over deterministic policies compatible with scope $\mathcal{S}$. Details of our Causal MWAL are described in Alg. 3. Step 8 amounts to finding the optimal policy in an SCM with a known expected $\mathbb{E}[Y \mid \boldsymbol{x}^*, \boldsymbol{z}^*]$. There are a huge array of techniques available for this, such as dynamic programming or policy gradient. Step 9 is the same as computing the feature expectations of a given policy. These can be computed exactly by solving k systems of linear equations, or they can be approximated using iterative techniques. The algorithm is essentially the MW algorithm (Freund & Schapire, 1999), applied to a game matrix $\widetilde{\boldsymbol{G}}$ defined in Step 3. Compared to the original game matrix $\boldsymbol{G}$ defined in Prop. 1, $\widetilde{\boldsymbol{G}}$ utilizes the estimates $\hat{\boldsymbol{\mu}}$ of the expert's feature expectations $\boldsymbol{\mu}$ from the observational data, rather than requiring that they be computed exactly. Following the strategy in (Syed & Schapire, 2008), we estimate the feature expectations $\boldsymbol{\mu}$ using its empirical mean from demonstrations.

Causal GAIL When the reward is non-linear, given expert demonstrations and the underlying graph $\mathcal{G}$, we can formalize the optimization problem based on the canonical equation Eq. (3) and Prop. 2. More specifically, when we utilize a regularizer $\psi(r)$ similar to (Ho & Ermon, 2016), the convex conjugate function ψ^* in Eq. (6) is given by:

$$\min_{\pi} \psi^* \left(\rho - \rho_\pi\right) = \min_{\pi} \max_{D \in (0,1)^{\boldsymbol{X}^* \times \boldsymbol{Z}^*}} E[\log(D(\boldsymbol{X}^*, \boldsymbol{Z}^*))] + E_\pi[\log(1 - D(\boldsymbol{X}^*, \boldsymbol{Z}^*))], \quad (42)$$

where function $D \in \Omega_{\boldsymbol{X}^*} \times \Omega_{\boldsymbol{Z}^*} \mapsto (0, 1)$ is a discriminator classifier (e.g, a neural network). The above optimization problem is in the form of two neural networks competing against each other in a zero-sum game.

Details of our proposed Causal GAIL is summarized in Alg. 4. Step 4 optimizes parameters of the discriminator D using standard back-propagation. Step 5 computes a policy that minimizes a fixed reward $\log(1 - D)$. For discrete domains, transition probabilities $P(z_i \mid \boldsymbol{x}^*_{<i}, \boldsymbol{z}^*_{<i})$ are estimated using empirical means; no value function approximation is used. The optimal policy is then computed via standard dynamic programming (Howard, 1960; Puterman, 1994). For continuous domains, the optimal policy $\boldsymbol{\pi}^*$ is computed using policy gradient (Sutton et al., 1999). We note that effective actor-critic algorithms exist for optimizing policies in SCMs via value function approximation (Dawid & Didelez, 2010; Murphy, 2003; Zhang & van der Schaar, 2020).

D Details on the Experimental Setup

In this section, we describe details of the experimental setup and processing of datasets. We also provide in Appendix D.1 additional experiments. For all experiments, we evaluate 4 separate imitation learning strategies summarized as follows:

1. `BC`: standard behavior cloning approach that learns a series of policy mappings by utilizing all observed covariates preceding every action $X_i \in \boldsymbol{X}$.
2. `IRL`: standard inverse reinforcement learning approach utilizes all the available covariates preceding every action $X_i \in \boldsymbol{X}$ up to each time step. We will apply the MWAL algorithm (Syed & Schapire, 2008) when the expected reward $\mathbb{E}[Y \mid \boldsymbol{O}]$ is a linear combination of feature functions. Otherwise, the GAIL algorithm (Ho & Ermon, 2016) is used to obtain an imitating policy due to the expressive power of neural networks.
3. `Causal-BC`: the causal behavior cloning approach Zhang et al. (2020); Kumor et al. (2021) first selects a set of covariates $\boldsymbol{Z}_i$ for every action $X_i \in \boldsymbol{X}$ following the sequential π-backdoor criterion (Def. 3). It then learns an imitating policy $\boldsymbol{\pi}$ with scope $\mathcal{S} = \{\langle X_i, \boldsymbol{Z}_i \rangle\}_{i=1}^n$ using standard BC algorithms.
4. `Causal-IRL`: Our proposed causal inverse reinforcement learning approach obtains an imitating policy by solving the canonical causal IRL equation in Eq. (3). We will utilize the Causal MWAL algorithm (Prop. 1) when the expected reward $\mathbb{E}[Y \mid \boldsymbol{X}^*, \boldsymbol{Z}^*]$ is a linear combination of feature functions. Otherwise, we resort to the Causal GAIL algorithm (Prop. 2) due to the expressive power of neural networks.

For discrete domains (Experiments 1, 5, 6 and 7), we estimate transition distributions using empirical means and solve for an optimal policy in the RL step using dynamic programming. For an environment

Table 1: Expected rewards $\mathbb{E}[Y \mid \text{do}(\boldsymbol{\pi})]$ for all imitation strategies. $\mathbb{E}[Y]$ measures the expert's performance. For each experiment, we highlight policies with the optimal performance.

	Experiment	BC	IRL	$\boldsymbol{\pi}_{c\text{-}bc}$	$\boldsymbol{\pi}_{c\text{-}irl}$	$\mathbb{E}[Y]$
1	Backdoor	0.24 ± 0.0042	0.25 ± 0.017	0.45 ± 0.0057	$\mathbf{0.80} \pm 0.0048$	0.45
2	HighD + RA	0.38 ± 0.0073	0.38 ± 0.0815	0.49 ± 0.0066	$\mathbf{0.62} \pm 0.0026$	0.48
3	MNIST	0.56 ± 0.0046	0.37 ± 0.0035	0.56 ± 0.0048	$\mathbf{0.75} \pm 0.0027$	0.56
4	MDPUC	0.435 ± 7.160	0.457 ± 3.718	0.669 ± 1.789	$\mathbf{0.868} \pm 0.327$	0.8
5	Frontdoor	0.52 ± 0.0046	0.51 ± 0.077	0.52 ± 0.0054	$\mathbf{0.63} \pm 0.0045$	0.60
6	Backdoor (Linear)	0.70 ± 0.0044	0.72 ± 0.013	0.75 ± 0.0036	$\mathbf{0.98} \pm 0.0003$	0.75
7	Frontdoor (Linear)	0.62 ± 0.0037	0.50 ± 0.0040	0.62 ± 0.0036	$\mathbf{0.75} \pm 0.0030$	0.62
8	HighD	0.38 ± 0.0073	0.38 ± 0.082	0.49 ± 0.0066	$\mathbf{0.49} \pm 0.051$	0.48

with continuous states and actions space (Experiments 2, 3, 4 and 8), we solve for the optimal policy using stochastic policy gradient (Sutton et al., 1999). Expected rewards of policies are estimated by a Monte-Carlo method computing empirical means over the agent's trajectories. Effective function approximation methods also exist to evaluate the value function of policies in high-dimensional domains (Tsitsiklis & Van Roy, 1997; Murphy, 2005).

Each experiment is repeated for 100 times; we measure the expected reward $\mathbb{E}[Y \mid \text{do}(\boldsymbol{\pi})]$ of all algorithms averaging over 100 repetitions. All experiments were performed on Intel Cascade Lake processors with 30 vCPUs and 120 GB memory in Ubuntu 18.04, implemented in Python. We will release the source code with the camera-ready version of the paper if the manuscript is accepted.

Experiment 1: Backdoor We study the problem of imitation learning in an SCM compatible with the causal diagram in Fig. 1c. X_1, X_2, Z_1, Z_2 are binary variables in $\{0, 1\}$. The reward signal Y is determined by a non-linear function. Detailed parametrization of this SCM is provided as follows:

$$\begin{aligned}
&U_{Z_1,Z_2} \sim \texttt{Bern}(0.8), && U_{Z_2,X_2} \sim \texttt{Bern}(0.8), && U_{Z_1,Y} \sim \texttt{Bern}(0.2)\\
&U_{Z_2} \sim \texttt{Bern}(0.1), && Z_1 \leftarrow U_{Z_1,Z_2} \oplus U_{Z_1,Y}, && X_1 \sim \texttt{Bern}(0.68)\\
&Z_2 \leftarrow U_{Z_2} \oplus U_{Z_1,Z_2} \oplus U_{Z_2,X_2}, && X_2 \leftarrow U_{Z_2,X_2} \oplus Z_2 &&\\
&Y \leftarrow U_{Z_1,Y} \oplus X_1 \oplus X_2 \oplus Z_2 \oplus Z_2 &&&&
\end{aligned} \tag{43}$$

Both `BC` and `IRL` learn a policy associated with the scope $\{\langle X_1, Z_1\rangle, \langle X_2, \{Z_1, X_1, Z_2\}\rangle\}$, which we denote by $\boldsymbol{\pi}_{bc}$ and $\boldsymbol{\pi}_{irl}$ respectively. `Causal-BC` utilizes the sequential π-backdoor admissible scope $\{\langle X_1, \{Z_1\}\rangle, \langle X_2, \{Z_2\}\rangle\}$ and learns an imitating policy $\boldsymbol{\pi}_{c\text{-}bc}$. Our proposed `Causal-IRL` learner obtains an imitating policy $\boldsymbol{\pi}_{c\text{-}irl}$ associated with scope $\{\langle X_1, \emptyset\rangle, \langle X_2, \{Z_2\}\rangle\}$, which satisfies the minimal sequential π-backdoor in Def. 4. Since the expected reward $\mathbb{E}[Y \mid X_1, X_2, Z_2]$ is non-linear, we solve for $\boldsymbol{\pi}_{c\text{-}irl}$ using the Causal GAIL algorithm described in Prop. 2. Reward augmentation (RA) is performed to incorporates the parametric knowledge that $\mathbb{E}[Y \mid X_1, X_2, Z_2]$ is a monotone function with regard to values of X_1, X_2 (Li et al., 2017). This is done by adding an additional regularization function in Eq. (6) to encourage the `Causal-IRL` agent to take higher values of actions X_1, X_2.

Simulation results are shown in Fig. 3a. Values of expected rewards for all imitation strategies are provided in Table 1. The analysis reveals that `Causal-IRL` consistently outperforms the expert's policy and other imitation strategies by exploiting additional parametric knowledge about the expected reward $\mathbb{E}[Y \mid X_1, X_2, Z_2]$; `Causal-BC` can learn a policy that mimics the expert's performance. As expected, `BC` and `IRL` fail to obtain an imitating policy performing as well as the expert's policy.

Experiment 2: Highway Driving This experiment evaluates our proposed `Causal-IRL` on a real-world `HighD` dataset Krajewski et al. (2018) which consists of humans' natural driving trajectories on highway. We study the problem of imitation learning in an SCM compatible with the

causal diagram in Fig. 4; where reward signal Y is determined by a non-linear function. Detailed construction of this SCM is described as follows.

1. Z_1 is both longitudinal and lateral historical accelerations of the ego vehicle two steps ago.
2. X_1 is the accelerations of the ego vehicle at the previous step.
3. Z_2 is the velocity of the preceding vehicle.
4. X_2 is the velocity of the ego vehicle.
5. We bootstrap samples of Z_1, X_1, Z_2, X_2 from the `HighD` dataset.
6. L was constructed such that X_2 satisfies the relation $L = I_{\{X_2 - Z_2 > -0.4\}}$. This leads to a distribution $P(L = 1) = 0.62$.
7. Values of Y, W are decided by the following causal mechanisms:

$$\begin{aligned} &W \leftarrow L \wedge U_{W,Y}, \quad U_{W,Y} \sim \texttt{Bern}(0.62) \\ &Y \leftarrow \left(U_{W,Y} \wedge I_{\{X_2 - Z_2 \leq -0.4\}}\right) \vee \left(\neg U_{W,Y} \wedge I_{\{X_2 - Z_2 > -0.4\}}\right). \end{aligned} \tag{44}$$

We also perform the standard causal discovery algorithm (Verma & Pearl, 1988; Spirtes et al., 2000) to the `HighD` data over variables Z_1, X_1, Z_2, X_2. The analysis reveals the independence relationship $(Z_2 \perp\!\!\!\perp X_1, Z_1)$. In practice, it is common that the leading vehicle's velocity will not be affected by the follower's past accelerations (Treiber et al., 2000; Higgs et al., 2011). We do not impose any constraints among variables Z_1, X_1, X_2 belonging to the ego vehicle. That is, the above construction procedure is compatible with the causal diagram in Fig. 4.

Both `BC` and `IRL` learn a policy associated with the scope $\{\langle X_1, \{Z_1\}\rangle, \langle X_2, \{Z_1, X_1, W, Z_2\}\rangle\}$, which we denote by $\boldsymbol{\pi}_{bc}$ and $\boldsymbol{\pi}_{irl}$ respectively. On the other hand, `Causal-BC` and `Causal-IRL` utilizes scope $\{\langle X_1, \emptyset\rangle, \langle X_2, \{Z_2\}\rangle\}$ which satisfies the minimal sequential π-backdoor (Def. 4). `Causal-BC` learns a policy $\boldsymbol{\pi}_{c\text{-}bc}$ via behavior cloning. Meanwhile, for any policy $\boldsymbol{\pi}$ with scope $\{\langle X_1, \emptyset\rangle, \langle X_2, \{Z_2\}\rangle\}$, the expected reward $\mathbb{E}[Y \mid \text{do}(\boldsymbol{\pi})]$ decomposes as, following Thm. 1,

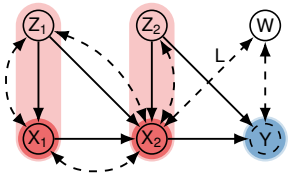


Figure 4

$$\mathbb{E}[Y \mid \text{do}(\boldsymbol{\pi})] = \sum_{x_1, x_2, z_2} \mathbb{E}[Y \mid x_1, x_2, z_2] P(z_2 \mid x_1) \pi_1(x_1) \pi_2(x_2 \mid z_2). \tag{45}$$

Using the above identification formula, `Causal-IRL` translates the imitation problem into the canonical equation of Eq. (3) and obtains an imitating policy $\boldsymbol{\pi}_{c\text{-}irl}$ via the Causal GAIL algorithm (Prop. 2 and Alg. 4).

We evaluate `Causal-IRL` in the `HightD` dataset provided with additional parametric knowledge about the reward signal. Therefore, `Causal-IRL` is able to exploit the fact that the expected reward $\mathbb{E}[Y \mid X_1, X_2, Z_2]$ is a monotone function with regard to X_2, Z_2. Reward augmentation (Li et al., 2017) is performed to account for such parametric knowledge in the Causal GAIL algorithm. Particularly, we add an additional regularization function in the optimization problem of Eq. (6) to encourage the agent to take higher values of action X_2 and achieve higher values of state Z_2.

We show simulation results in Fig. 3b. Details of expected rewards for all imitation strategies are described in Table 1. We found that `Causal-BC` learner is able to achieve the expert's performance. `Causal-IRL` is able to dominate the expert and other imitation strategies. As expected, `BC` and `IRL` perform the worst among all algorithms and fail to obtain an effective imitating policy matching the expert's behavior policy.

Experiment 3: MNIST Digits This experiment evaluates our proposed `Causal-IRL` on the `MNIST` digit dataset (LeCun, 1998), which consists of images of handwritten digits. Again, we study the problem of imitation learning in an SCM compatible with the frontdoor diagram in Fig. 2a. However, values of mediator Z are now replaced with an image bootstrapped from `MNIST` dataset containing digit 0 or 1. More specifically, the data-generating mechanisms of this SCM are described as follows:

$$\begin{aligned} &P(U = 1) = 0.75, \quad P(X = 1 \mid U = 0) = 0.5, \quad P(X = 1 \mid U = 1) = 0.5 \\ &P(Z = 0 \mid X = 0) = 0.2, \quad P(Z = 0 \mid X = 1) = 0.8, \quad Y \leftarrow 0.25U + 0.75Z \end{aligned} \tag{46}$$

Both BC and IRL learn a policy associated with the scope $\{\langle X, \emptyset\rangle\}$, which we denote by $\boldsymbol{\pi}_{bc}$ and $\boldsymbol{\pi}_{irl}$ respectively. Since there exists no sequential π-backdoor admissible scope in Fig. 2a, Causal-BC fails and resorts to a policy obtained by standard BC, i.e., $\boldsymbol{\pi}_{c\text{-}bc} = \boldsymbol{\pi}_{bc}$. Finally, Causal-IRL obtains an imitating policy $\boldsymbol{\pi}_{c\text{-}irl}$ with scope $\{\langle X, \emptyset\rangle\}$. It formulates the canonical equation in Eq. (9) as a two-person zero-sum game, which can be solved by the Causal MWAL algorithm described in Prop. 1. Simulation results are shown in Fig. 3c. Details of expected rewards for all imitation strategies are shown in Table 1. We found that Causal-IRL outperforms Causal-BC and BC; while IRL performs the worst among all algorithms.

Experiment 4: Infinite MDPUC This experiment evaluates our proposed Causal-IRL on an MDPUC model with the infinite horizon (Ruan & Di, 2022), which is a Markov decision process explicitly incorporating the existence of unobserved confounders (Zhang & Bareinboim, 2016; 2022). Fig. 1d shows a simplified causal diagram of this MDPUC model for $i = 1, 2, 3$. A more detailed causal diagram was described in (Ruan & Di, 2022).

More specifically, at the time step i, the expert takes an action X_i based on the current state Z_i, U_π and past action X_{i-1}. Covariates $U_i^{(Z,Y)}$, and U_π are unobserved to the imitator, thus summarized as exogenous variables. At every time step i, only information that the imitator could utilize is the past observed states and actions $\{Z_0, X_0, Z_1, X_1, \ldots, Z_{i-1}, X_{i-1}\}$. Detailed semantics of the SCM is described as follows.

1. $U_i^{(Z,Y)}$ is the road conditions.
2. U_π denotes the level of expert driving skills.
3. $U_i^{(A)}$ is the action of the leading vehicle, which cannot be observed by the imitator or the expert.
4. Z_i contains the velocities of the follower vehicle and the leading vehicle, and the gap between them, which is affected by $U_{i-1}^{(A)}$, Z_{i-1}, X_{i-1} and $U_i^{(Z,Y)}$.
5. X_i is the acceleration or deceleration of the follower vehicle.
6. Y_i is the reward, which takes multiple goals into consideration, e.g., efficiency, comfort, and safety. When the agent is involved in a collision, the environment will be stopped, and the agent will receive a negative penalty.

Both BC and IRL learn a policy associated with the scope $\mathcal{S} = \{\langle X_i, Z_i\rangle\}_{i=1}^{\infty}$. On the other hand, Causal-BC and Causal-IRL utilize scope $\mathcal{S}^* = \{\langle X_i, \{Z_i, X_{i-1}, Z_{i-1}\}\rangle\}_{i=1}^{\infty}$ which satisfies the minimal sequential π-backdoor (Def. 4). Causal-BC learns a stationary policy $\boldsymbol{\pi}_{c\text{-}bc} \sim \mathcal{S}^*$ via standard behavioral cloning. Meanwhile, Causal-IRL exploits the Markov property up to time step i, which is,

$$P(z_i \mid \boldsymbol{x}_{<i}, \boldsymbol{z}_{<i}) = P(z_i \mid x_{i-1}, z_{i-1}) = P(z_2 \mid x_1, z_1) \tag{47}$$

$$\mathbb{E}[Y_i \mid \boldsymbol{x}_{\leq i}, \boldsymbol{z}_{\leq i}) = \mathbb{E}[Y_i \mid x_i, x_{i-1}, z_i, z_{i-1}] = \mathbb{E}[Y_2 \mid x_2, x_1, z_2, z_1]. \tag{48}$$

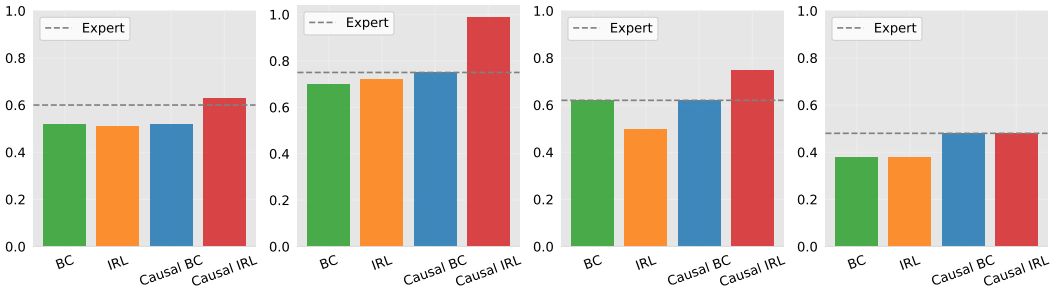
By doing so, Causal-IRL translates the imitation problem into the canonical equation of Eq. (3) and obtains an imitating policy $\boldsymbol{\pi}_{c\text{-}irl}$ via the Causal GAIL algorithm (Prop. 2 and Alg. 4). Fig. 3d shows the results, where Causal-IRL outperforms Causal-BC, IRL and BC.

D.1 Additional Experiments

We also evaluate our imitation approach on various synthetic SCM instances with other forms of reward functions. Overall, we found that simulation results match our findings in the main manuscript. Our causal IRL approach consistently dominates state-of-art imitation learning methods across various causal diagrams. Furthermore, it is able to incorporate parametric knowledge about the reward function and achieve effective imitating policies from high-dimensional demonstration data.

Experiment 5: Frontdoor We study the problem of imitation learning in an SCM compatible with the frontdoor diagram in Fig. 2a. X, Y, Z are binary variables in $\{0, 1\}$. The reward signal Y is determined by a non-linear function. Detailed parametrization of this SCM is provided as follows:

$$\begin{aligned} &U \sim \texttt{Bern}(0.75), \quad U_X \sim \texttt{Bern}(0.7), \quad U_Z \sim \texttt{Bern}(0.25) \\ &X \leftarrow U \oplus U_X, \qquad Z \leftarrow X \oplus U_Z, \qquad Y \leftarrow U \oplus Z \end{aligned} \tag{49}$$



(a) Experiment 5 (b) Experiment 6 (c) Experiment 7 (d) Experiment 8

Figure 5: Simulation results for additional experiments that are not included in the main manuscript.

Both BC and IRL learn a policy associated with the scope $\{\langle X, \emptyset\rangle\}$, which we denote by $\boldsymbol{\pi}_{bc}$ and $\boldsymbol{\pi}_{irl}$ respectively. Since there exists no sequential π-backdoor admissible scope in Fig. 2a, Causal-BC fails and resorts to a policy obtained by standard BC, i.e., $\boldsymbol{\pi}_{c\text{-}bc} = \boldsymbol{\pi}_{bc}$. Finally, Causal-IRL obtains an imitating policy $\boldsymbol{\pi}_{c\text{-}irl}$ with $\{\langle X, \emptyset\rangle\}$ by solving the canonical equation in Eq. (9). It performs reward augmentation (Li et al., 2017) to incorporate the parametric knowledge that $\mathbb{E}[Y \mid X, Z]$ is a monotone function with regard to values of X, Z. Simulation results are shown in Fig. 5a. Details of expected rewards for all imitation strategies are provided in Table 1. The analysis reveals that Causal-IRL consistently outperforms the expert's policy by exploiting additional parametric knowledge about the expected reward $\mathbb{E}[Y \mid X, Z]$; while other strategies fail to match the sub-optimal expert.

Experiment 6: Backdoor with Linear Reward We now consider an SCM instance compatible with the causal diagram in Fig. 6 where reward signal Y is determined by a linear function; X_1, X_2, Z_1, Z_2 are binary variables in $\{0, 1\}$. Detailed parametrization of this SCM is given by:

$$\begin{aligned} &U_{Z_1,Z_2}, U_{Z_1,X_2} \sim \texttt{Bern}(0.5), \quad X_1 \sim \texttt{Bern}(0.75), \quad Z_1 \leftarrow U_{Z_1,Z_2} \wedge U_{Z_1,X_2} \\ &X_2 \leftarrow U_{Z_1,X_2}, \qquad\qquad Z_2 \leftarrow X_2 \vee U_{Z_1,Z_2}, \quad Y \leftarrow 0.5X_1 + 0.5Z_2 \end{aligned} \tag{50}$$

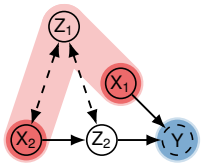


Figure 6

Both BC and IRL learn a policy associated with the scope $\{\langle X_1, \{Z_1\}\rangle, \langle X_2, \{Z_1, X_1\}\rangle\}$, which we denote by $\boldsymbol{\pi}_{bc}$ and $\boldsymbol{\pi}_{irl}$ respectively. Causal-BC utilizes the sequential π-backdoor admissible scope $\{\langle X_1, \{Z_1\}\rangle, \langle X_2, \emptyset\rangle\}$ and learns an imitating policy $\boldsymbol{\pi}_{c\text{-}bc}$. Our proposed Causal-IRL learner obtains an imitating policy $\boldsymbol{\pi}_{c\text{-}irl}$ associated with scope $\{\langle X_1, \emptyset\rangle, \langle X_2, \emptyset\rangle\}$, which satisfies the minimal sequential π-backdoor in Def. 4. Since the expected reward $\mathbb{E}[Y \mid X_1, X_2]$ is a linear function, we solve for $\boldsymbol{\pi}_{c\text{-}irl}$ using the Causal MWAL algorithm described in Prop. 1. Simulation results are shown in Fig. 5c. Expected rewards for all imitation strategies are provided in Table 1. Our analysis reveals that Causal-IRL consistently outperforms the expert's policy and other imitation strategies by exploiting the linearity of the expected reward $\mathbb{E}[Y \mid X_1, X_2]$; Causal-BC is able to learn a policy that mimics the expert's performance; while BC and IRL perform the worst among all.

Experiment 7: Frontdoor with Linear Reward Consider again the frontdoor diagram in Fig. 2a with binary variables $X, Y, Z \in \{0, 1\}$. The reward signal Y is now determined by a linear function. Detailed parametrization of the underlying SCM is provided as follows:

$$\begin{aligned} &U \sim \texttt{Bern}(0.75), \quad U_X \sim \texttt{Bern}(0.5), \quad U_Z \sim \texttt{Bern}(0.75) \\ &X \leftarrow U \oplus U_X, \qquad Z \leftarrow X \oplus U_Z, \qquad Y \leftarrow 0.5U + 0.5Z \end{aligned} \tag{51}$$

Both BC and IRL learn a policy associated with the scope $\{\langle X, \emptyset\rangle\}$, which we denote by $\boldsymbol{\pi}_{bc}$ and $\boldsymbol{\pi}_{irl}$ respectively. Causal-BC fails to find a π-backdoor admissible scope and resorts to a standard BC policy $\boldsymbol{\pi}_{c\text{-}bc} = \boldsymbol{\pi}_{bc}$. Causal-IRL obtains an imitating policy $\boldsymbol{\pi}_{c\text{-}irl}$ with scope $\{\langle X, \emptyset\rangle\}$. It formulates the canonical equation in Eq. (9) as a two-person zero-sum game, which can be solved by the Causal MWAL algorithm described in Prop. 1. Simulation results are shown in Fig. 5c. Expected rewards for all imitation strategies are provided in Table 1. The analysis reveals that Causal-IRL consistently outperforms the expert's policy and other imitation strategies by exploiting the linearity of the expected reward $\mathbb{E}[Y \mid X, Z]$; while IRL performs the worst among all strategies.

Experiment 8: HighD without Reward Augmentation Finally, we evaluate `Causal-IRL` in the `HightD` dataset without utilizing any additional parametric knowledge about the reward signal. For this experiment, the simulation setup and training process remain the same as in Experiment 2. The only difference here is that `Causal-IRL` now does not exploit the fact that the expected reward $\mathbb{E}[Y \mid X_1, X_2, Z_2]$ is a monotone function with regard to X_2, Z_2.

We show in Fig. 5d simulation results for this experiment. Table 1 provides details of expected rewards for all imitation strategies. We could see by inspection that the performance of `BC`, `IRL`, `Causal-BC` and the expert remains the same. The performance of `Causal-IRL` decreases since it is now unable to exploit the monotonicity of the expected reward $\mathbb{E}[Y \mid X_1, X_2, Z]$. However, it is still able to achieve the expert's performance and obtain an effective imitating policy.

E Related Work and Contributions

Imitation Learning (IL) paradigm investigates the problem of an imitator learning how to behave in an environment with an unknown reward function by observing demonstrations from a human expert (Argall et al., 2009; Billard et al., 2008; Hussein et al., 2017; Osa et al., 2018). There are two major learning modalities that implements IL – *behavioral cloning* (BC) (Widrow, 1964; Pomerleau, 1989; Muller et al., 2006; Mülling et al., 2013; Mahler & Goldberg, 2017) and *inverse reinforcement learning* (IRL) Ng et al. (2000); Ziebart et al. (2008); Ho & Ermon (2016); Fu et al. (2017). BC methods directly mimic the expert's behavior policy by learning a mapping from observed states to the expert's action via supervised learning. Alternatively, IRL methods first learn a potential reward function under which the expert's behavior policy is optimal. The imitator then obtains a policy by employing standard RL methods to maximize the learned reward function. Under some common assumptions, both BC and IRL are able to obtain policies that achieve the expert's performance (Kumor et al., 2021; Swamy et al., 2021). Moreover, when additional parametric knowledge about the reward function is provided, IRL may produce a policy that outperforms the expert's in the underlying environment (Syed & Schapire, 2008; Li et al., 2017; Yu et al., 2020).

Despite the performance guarantees provided by existing imitation methods, both BC and IRL rely on the assumption that the expert's input observations match those available to the imitator. While this assumption may certainly hold in some settings, it is quite stringent in others, where agents may differ in their sensory capabilities, have distinct goals in mind, and evolve their reasoning (e.g., "software") and attention independently. When the expert and imitator's input covariates disagree, the expert's demonstration could be contaminated with the presence of unobserved confounders (UCs). In this case, naively applying standard imitation methods may not necessarily lead to satisfactory performance, even when the expert him or herself behaves optimally (Zhang et al., 2020).

Causal Inference (CI) addresses the challenge of unobserved confounding bias in the observational data by exploiting causal assumptions about the data-generating mechanisms (commonly through causal graphs and potential outcomes). Several criteria and algorithms have been developed (Pearl, 2000; Spirtes et al., 2000; Bareinboim & Pearl, 2016). More recently, there exists an emerging line of research under the rubric of *causal imitation learning* that augments the imitation paradigm to account for environments consisting of arbitrary causal mechanisms and the aforementioned mismatch between expert and imitator's sensory capabilities (de Haan et al., 2019; Zhang et al., 2020; Etesami & Geiger, 2020; Kumor et al., 2021). Closest to our work, Zhang et al. (2020); Kumor et al. (2021) derive graphical criteria that completely characterize when and how BC could lead to successful imitation even when the agents perceive reality differently. However, most of these causal imitation methods focus on the BC setting, where the performance of the imitator is limited by the expert's policy. (de Haan et al., 2019) studies the problem of "causal confusion" by assuming that there does not exist any UC. There are several other works that investigate this problem based on the framework of (or variations of) Markov decision processes (MDP) (Puterman, 1994), including contextual MDPs (Tennenholtz et al., 2021), MDPs with temporally correlated noises (Swamy et al., 2022), and MDPs with the presence of UCs (Ruan & Di, 2022). Still, it is unclear how to perform IRL-type training beyond MDP settings, especially when expert policies are non-stationary, the underlying environment is non-Markovian, and UCs are generally present.

Contributions This paper investigates the settings where the expert's demonstrations are contaminated with biases induced by UCs. When UCs are present, it is possible to construct a simple

causal model with binary variables such that existing IRL methods cannot learn an effective imitating policy from demonstrations alone, regardless of how advanced the feature extraction procedures are and what the sample size is. In a nutshell, our contribution is to empower imitation learning to be robust against confounding bias through knowledge augmentation via a causal lens. Accordingly, we propose the first causal IRL method that could perform at least as well as the expert while being robust against UCs by exploiting causal relationships encoded in the underlying environment. The learned IRL policy may significantly outperform the expert's policy by leveraging additional parametric knowledge about the reward signal. Moreover, our method reduces the causal IRL to the canonical form, which allows us to apply existing IRL methods to find an effective imitating policy, despite the presence of unobserved confounding bias.

Finally, we would like to clarify that there are many challenges in reinforcement learning, including estimation in the high-dimensional domain. In this paper, we investigate a probably undermined but crucial one, which is, when the expert's demonstrations are contaminated with biases induced by UCs, potentially leading to suboptimal or even undesired imitating behaviors. It is orthogonal to many challenges that state-of-art IRL methods concern, e.g., high-dimensionality, and we believe that it is as equally urgent to address as other existing challenges in IRL. Our goal in this paper is to enhance the existing IRL methods to be robust against confounding bias through knowledge augmentation, provided with a causal perspective about the environment. This is where we believe that the communities of causality and imitation learning could and should converge.

F Frequently Asked Questions

Q1. Are causal assumptions necessary for causal imitation learning?

Answer. Yes, causal assumptions are necessary for practical imitation learning when some input covariates of the expert's policy are latent to the imitator, and unobserved confounders (UCs) are present in the expert's demonstrations. The departing point of our work is the realization that an imitating policy that performs as well as the expert is generally underdetermined by the observational data alone. Indeed, this observation is not new, and was first made in (Zhang et al., 2020).For concreteness, consider models M_1, M_2, unknown to researchers, where in M_1, $X \leftarrow U, Y \leftarrow X$; in M_2, $X \leftarrow U, Y \leftarrow X \oplus U$; in $M_i, i = 1, 2$, $P(U = 0) = P(U = 1) = 0.5$. We assume that Y, U are unobserved; Y is the reward. In both M_1 and M_2, the observational distribution $P(X = 0) = P(X = 1) = 0.5$. In M_1, $P(y)$ is imitable with policy $\pi(x) = P(x)$; while in M_2, $P(Y = 0|do(\pi)) = 0.5$ for any $\pi(x)$, which is far from $P(Y = 0) = 1$. This example shows that when UCs are present, imitation learning from observations alone is generally impossible.

A common approach to addressing the challenges of UCs is to explore causal relationships among variables. Such relationships could be compactly represented in the form of a causal diagram (Pearl, 2000). Consider again the previous example. Note that the bi-directed arrow $X \leftrightarrow Y$ exists in $\mathcal{G}_2$ but not $\mathcal{G}_1$. One could distinguish SCMs M_1, M_2 using their corresponding causal diagrams $\mathcal{G}_1, \mathcal{G}_2$, and obtain an effective imitating policy when it is possible (i.e., M_1).

Finally, we would like to point out that most standard imitation learning methods, including both BC and IRL, focus on the Markov decision process (MDP) model. It (perhaps implicitly) assumes the structural constraint of *unconfoundedness*, i.e., UCs do not exist in the expert's demonstrations. In other words, existing BC and IRL methods are developed based on a common set of causal assumptions, which we could compactly represent using a canonical causal diagram (e.g., see (Zhang & Bareinboim, 2022)). This canonical MDP graph does not include the presence of UCs, thus not representative of the more generalized setting studied in this paper.

Q2. Is the requirement of causal diagram a limitation?

Answer. We note that requiring a causal diagram as input is not a limitation of our work. Instead, as explained in Q1, causal assumptions are necessary for practical IRL that aims to provide guarantees and understand the conditions algorithms may succeed in many real-world applications. Our reasoning and arguments are summarized below:

1. Unobserved confounders generally exist in demonstration data when the sensory capabilities of the imitator and the expert differ, e.g., HighD (Krajewski et al., 2018).
2. Without any causal assumption, existing imitation learning algorithms, including BC and IRL, are unable to learn an effective policy from the demonstration data alone that performs as well as the expert (Etesami & Geiger, 2020; Kumor et al., 2021; Zhang et al., 2020).
3. That is, when UCs are present, causal assumptions about the environment are necessary to achieve reasonable imitation performance via IRL.
4. Existing IRL methods assume the underlying environment to be an MDP model, which satisfies the Markov property and precludes the analysis of UCs.
5. A causal diagram (Pearl, 2000) represents an arbitrary collection of causal relationships in the environment. This flexibility allows one to encode the existence of UCs explicitly.

To sum up, this paper extends existing IRL methods, via simple augmentation, to the generalized setting where UCs generally exist, and causal relationships are encoded in the form of a causal diagram. In a very general sense, the structural assumptions required to perform causal inferences are inevitable, as shown in (Bareinboim et al., 2022, Theorem 1). We then note that in some compelling real-world applications, one may have a qualitative understanding of the causal relationships of the environment, e.g., the ad-placement engine (Bottou et al., 2013). Also, there are quite general and systematic methods under the rubrics of "causal discovery" capable of learning a causal diagram (or its equivalence class) from observational data (Spirtes et al., 2000). Our methods could be combined with causal discovery algorithms to learn an imitating policy after the graph is obtained. This means that the dependence on domain experts could be minimized.

Q3. The learned policy requires a separate decision rule at each time step. How does it differ from stationary policies in MDPs?

Answer. A policy that remains invariant at each step is said to be a *stationary* policy. It is a special case of non-stationary policies where decision rules are different at every time step (Puterman, 1994; Fix et al., 1993). Stationary policies are optimal in MDP models that satisfy the Markov property. However, its optimality guarantee no longer holds in non-Markovian environments, where the transition distributions differ over every stage of action. This paper focuses on settings where Markovianity cannot be ascertained.

Generally, policies considered in this paper are also referred to as dynamic treatment regimes (DTR) (Murphy et al., 2001) in the causal inference literature. DTRs are a popular model, and have been widely applied for medical decision-making. Unlike standard MDP settings, DTR does not assume that the underlying environment satisfies the Markov property and the transition functions are generally different between different stages of intervention (action). Consequently, the optimal DTR policy is generally non-stationary, requiring a separate policy for every action stage.

We would like to highlight that the most challenging aspect of our investigation is confounding biases in the expert's demonstrations induced by unobserved confounders (UCs). UCs are pervasive in practical scenarios since not all information available to the decision-maker is recorded and available to the AI system, including the aforementioned medical treatment. Proposing a causal framework to imitate non-stationary policies for non-Markovian is more general, and can be naturally adapted to stationary policies in a Markovian environment.

Q4. What is the difference between GAIL and Causal GAIL?

Answer. GAIL was original developed in the Markov decision process (MDP) (Ho & Ermon, 2016), which encodes the assumption of *unconfoundedness*, i.e., unobserved confounders do not exist in the expert's demonstrations. On the other hand, Causal GAIL does not rely on such an assumption and is robust to the bias introduced by the presence of unobserved confounding.

Details of Causal GAIL are provided in Alg. 4 (Appendix C). During the initialization (Step 1), it first applies the identification procedure Thms. 1 and 2 to reduce the causal imitation problem to its corresponding canonical equation (Eq. (3)). Starting there, the learning procedure is essentially the GAIL algorithm applied to non-stationary DTR policies (Murphy et al., 2001). Step 4 updates the reward function D, parameterized as a neural network

discriminator, using the standard gradient descent. Step 5 computes an optimal policy in an SCM given a fixed reward function. There are a huge array of techniques available for this. As for discrete domains, the transition distributions are computed using empirical means, and we then solve for an optimal policy using standard dynamic programming (Bellman, 1966). For an environment with continuous states and actions space, we solve for the optimal policy using stochastic policy gradient (Sutton et al., 1999). Expected rewards of policies are estimated by a Monte-Carlo method computing empirical means over the agent's trajectories. Effective function approximation methods also exist to evaluate the value function of policies in high-dimensional domains (Tsitsiklis & Van Roy, 1997; Murphy, 2005).

Q5. Could the proposed method be scaled up to high-dimensional, complex environments?

Answer. Yes, our causal-IRL method naturally scales to general sequential decision-making settings with higher dimensional domains. A key observation in our proposed approach is the reduction to the canonical equation of causal IRL in Eq. (3). Once the canonical equation is obtained, the learner could then solves for an effective imitating policy using the standard IRL method, including MWAL and GAIL. Details of the implementation are described in Sec. 2.2 and Appendix C. This paper introduces novel algorithms (Thms. 1 and 2) to reduce the causal imitation learning problem to its corresponding canonical form (when it is possible). The reduction procedure takes at most a polynomial-number of steps with regard to the size of the causal diagram. This allows us to benefit from the state-of-art computational framework for IRL in MDP models while ascertaining the validity of the procedure through proper causal tools. For instance, we demonstrate the validity of our algorithm in Experiment 3 using the MNIST digits (LeCun, 1998).

Here, we would also like to clarify that there are many challenges in IRL. Scalability to the high-dimensional domain is one of them. However, in this paper, we investigate another challenge, a probably undermined but crucial one, which is, the settings where the expert's demonstrations are contaminated with biases induced by unobserved confounders (UCs). It is orthogonal to the challenge of high-dimensionality that state-of-art IRL methods concern, and we strongly believe that it is as equally urgent to address as other existing challenges in IRL. Indeed, when UCs are present, it is possible to construct a simple model with binary variables such that existing IRL methods cannot learn an effective imitating policy from demonstrations alone, regardless of how advanced the feature extraction procedures are and what the sample size is. Our goal in this paper is to enhance the existing IRL methods to be robust against confounding bias through knowledge augmentation, provided with a causal perspective about the environment. This is where we believe that the communities of causality and imitation learning could and should converge. Thus, it is not our purpose to beat existing IRL methods in another benchmark with high-dimensional domains that match the assumptions they expect (i.e., UCs are assumed away).

Q6. Could the proposed method be scaled to long-sequence decision problems with infinite horizons?

Answer. Yes, the proposed methods could be generalized to sequential decision problems with infinite horizons by incorporating common assumptions of temporal models. In Experiment 4, we evaluate the proposed algorithm in a generalized Markov decision process incorporating unobserved confounders, called the MDPUC (Ruan & Di, 2022; Zhang & Bareinboim, 2022). The graph underlying this experiment properly simulates the real-world driving decision process. By exploiting the Markov property over time, we can decompose the causal diagram over the infinite horizon into a collection of sub-graphs, one for each pair of the action X_i and the reward Y_i. Fig. 1d shows the causal diagram spanning over time step $i = 1, 2, 3$. As a comparison, `BC` and `IRL` still utilize the standard 1-step information scope $\{\langle X_i, \{Z_i\}\rangle\}$. By applying Thm. 1 at each time step, we obtain a π-backdoor admissible policy scope $\{\langle X_i, \{Z_i, X_{i-1}, Z_{i-1}\}\rangle\}$ for `Causal-IRL` and `Causal-BC`. Simulation results are shown in Fig. 3d. One could see by inspection that `Causal-IRL` performs the best among all algorithms and achieve the experts' performance.

Q7. How is Causal IRL able to outperform Causal BC and the expert?

Answer. When the imitator has no prior knowledge about the reward function, methods like GAIL model the reward function as a general discriminator (i.e., a neural network),

and are able to converge to the equilibrium point where the imitator exactly matches the expert's distributions. In other words, causal IRL generally coincides with causal BC without additional parametric knowledge about the latent reward signal.

In general, causal IRL consistently dominates causal BC since it can exploit parametric knowledge about the reward function. This observation is not new and was noted in the literature, e.g., see (Syed & Schapire, 2008; Li et al., 2017). Effective methods exist to encode prior knowledge about the reward signal when training GAIL-type imitators, which are referred to as reward augmentation (Li et al., 2017). For instance, in Experiment 2, we train causal-IRL imitator using reward augmentation that encourages the agent to accelerate when the velocity of the preceding vehicle is high. Simulation results in Fig. 3b show that causal-IRL can consistently outperform the expert and causal BC by exploiting the additional parametric knowledge about reward. In Experiment 8 (highway driving), we trained another causal-IRL imitator without additional knowledge about the reward. Simulation results in Fig. 5d show that the performance of causal-IRL and causal-BC coincide, exactly matching the expert's performance.

Q8. Could the proposed method be applicable to settings with multiple reward signals?

Answer. Yes, following the conventions of the dynamic treatment regime (DTR) literature (Murphy et al., 2001), the reward signal is represented as a single random variable. In a nutshell, policy optimization methods for DTR extend immediately to multiple reward signals, one per every stage of action (Chakraborty & Moodie, 2013). Theoretically, Thms. 1 and 2 hold when the reward Y is a set of variables. The reward function $r(\boldsymbol{x}^*, \boldsymbol{z}^*)$ is defined as the cumulative reward $\sum_{Y \in \boldsymbol{Y}} \mathbb{E}[Y \mid \boldsymbol{x}^*, \boldsymbol{z}^*]$. For concreteness, we demonstrate our algorithm in Experiment 4 to optimize the cumulative reward in a sequential decision-making problem with an infinite horizon.

Q9. What is the projection algorithm, and why do we need to perform such projections?

Answer. The projection algorithm (Tian, 2002, Sec. 4.5) marginalizes unobserved endogenous variables in a causal diagram and allows us to focus on the causal relationships among observed endogenous variables. It is particularly useful for causal identification since the identifiability of causal effects remains invariant across the projection operation. Details fo the projection algorithm is provided in Alg. 5. It transforms an arbitrary causal diagram $\mathcal{G}$ with observed endogenous variables $\boldsymbol{O}$ into a simplified causal diagram $\mathcal{H}$ such that unobserved endogenous variables $\boldsymbol{V} \setminus \boldsymbol{O}$ are omitted.

For instance, consider a causal diagram $\mathcal{G}$ with directed arrows $X \to Z \to Y$ and a bi-directed arrow $X \leftrightarrow Y$; variable Z is unobserved so the observational distribution $P(X, Y)$. In this case, one could simplify the diagram by projecting out node Z, i.e., $\mathcal{H} = \textsc{Project}(\mathcal{G})$. The resulting diagram $\mathcal{H}$ is given by $X \to Y$ and $X \leftrightarrow Y$. It can be shown that the interventional distribution $P(y|\text{do}(x))$ is not identifiable in the projected diagram $\mathcal{H}$, thus is not identifiable from $P(X, Y)$ in the original diagram G.

Due to these nice properties of the projection algorithm, in Sec. 3, we consistently assume the causal diagram $\mathcal{G}$ to be the outcome of the projection algorithm, only consisting of endogenous nodes $\boldsymbol{O}, Y$. The identifiablity results in Thm. 2 and Lem. 1 hold without loss of generality.

Q10. Why does the proposed method mainly focus on $\nu^* \leq 0$?

Answer. The condition $\nu^* \leq 0$ indicates that the expert's policy might be sub-optimal and the Causal IRL algorithm is able to find a policy that improves over the expert.

Many real-world applications exist where the expert generating the demonstrations is possibly sub-optimal and could be improved. For instance, consider the development of autonomous vehicles, where the reinforcement signal is never fully known, and imitation learning methods are widely used. To collect demonstration data, it often requires a human demonstrator, possibly a driving expert, to drive the vehicle around the target environment (e.g., a city, a highway) and record the natural driving trajectories. No matter how clean their driving record is, the human demonstrator is naturally error-prone and cannot guarantee optimal driving behaviors at all times (Atchley et al., 2011). Indeed, in many complicated learning tasks, it might be difficult for the demonstrator to be consistently optimal, even for

Algorithm 5: PROJECT (Tian, 2002, Def. 5)

Input: A causal diagram $\mathcal{G}$.
Output: A causal diagram $\mathcal{H}$ where all endogenous variables are observed, i.e., $\boldsymbol{V} = \boldsymbol{O}$.
Let $\boldsymbol{O}, \boldsymbol{L}$ be, respectively, observed endogenous variables and latent endogenous variables in $\mathcal{G}$.
Let $\mathcal{H}$ be a causal diagram constructed as follows.
for each observed $V \in \boldsymbol{O}$ in $\mathcal{G}$ **do**
  Add an observed node V in $\mathcal{H}$.
end for
for each pair $S, E \in \boldsymbol{O}$ in $\mathcal{G}$ s.t. $S \neq E$ **do**
  if there exists a directed path $S \to E$ in $\mathcal{G}$ **then**
    Add an edge $S \to E$ in $\mathcal{H}$.
  else if there exists a path $S \to V_1 \to \cdots \to V_n \to E$ in $\mathcal{G}$ s.t. $V_1, \cdots V_n \in \boldsymbol{L}$ **then**
    Add an edge $S \to E$ in $\mathcal{H}$.
  else if there exists a bidirected edge $S \leftrightarrow E$ in $\mathcal{G}$ **then**
    Add a bidirected edge $S \leftrightarrow E$ in $\mathcal{H}$.
  else if there exists a path $S \leftarrow V_{l,1} \leftarrow \cdots \leftarrow V_{l,n} \leftrightarrow V_{r,m} \to \cdots \to V_{r,1} \to E$ in $\mathcal{G}$ s.t. $V_{l,1}, \cdots, V_{l,n}, V_{r,1}, \cdots, V_{r,m} \in \boldsymbol{L}$ **then**
    Add a bidirected edge $S \leftrightarrow E$ in $\mathcal{H}$.
  else if there exists a path $S \leftarrow V_{l,1} \leftarrow \cdots \leftarrow V_{l,n} \leftarrow V_c \to V_{r,m} \to \cdots \to V_{r,1} \to E$ in $\mathcal{G}$ s.t. $V_{l,1}, \cdots, V_{l,n}, V_c, V_{r,1}, \cdots, V_{r,m} \in \boldsymbol{L}$ **then**
    Add a bidirected edge $S \leftrightarrow E$ in $\mathcal{H}$.
  end if
end for
Return $\mathcal{H}$.

relatively veteran experts, e.g., playing complex video games or stock trading (Newell et al., 1972; Nikolaidis et al., 2017).

Behavioral cloning methods mimic the demonstrator's policy and thus never outperform the sub-optimal expert. On the other hand, by exploiting additional parametric knowledge about the latent reward signal, inverse RL methods are able to obtain an effective policy that consistently dominates the expert's policy. This observation is not new, and was first proposed in (Syed & Schapire, 2008), who also formalized the use of performance gap ν^*. (Zhang et al., 2020) studied the behavioral cloning from the combination of confounded observations and causal constraints about the environment. In this work, we take inspiration from inverse RL approaches and develop non-trivial imitation learning methods that are robust to both unobserved confounding and a sub-optimal expert.

Q11. How about $\nu^* > 0$?

Answer. When the performance gap $\nu^* > 0$, there exists at least one causal model such that the imitator can never achieve the expert's performance.

For instance, consider a bow graph consisting of a directed arrow $X \to Y$ and a bi-directed arrow $X \leftrightarrow Y$. We could construct a causal model M compatible with this causal constraint such that no intervention $\text{do}(x)$ on action X could outperform the expert's performance $E[Y]$. Let an exogenous variable U be uniformly drawn over a binary domain $\{0, 1\}$. Let the expert's policy be $X \leftarrow U$ and the reward function $Y \leftarrow X \oplus \neg U$. Fig. 7 below describes the causal diagram associated with SCM M.

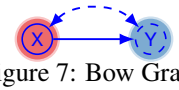


Figure 7: Bow Graph

Evaluating the expert's reward in SCM M gives

$$E[Y] = E[X \oplus \neg U] = E[U \oplus \neg U] = 1$$

On the other hand, for any intervention $do(x)$, the imitator's reward is given by

$$E[Y|do(x)] = E[x \oplus \neg U] = 0.5$$

The last step holds since U is a uniform binary variable. In this environment, for any action, the imitator's reward $E[Y|do(x)] = 0.5$ is far from the expert's performance $E[Y]$. This means that in the bow graph, the performance gap is at least $\nu^* > E[Y] - E[Y|do(x)] = 0.5$. Such settings are referred to as "non-imitable" in (Zhang et al., 2020, Lemma 1).

Q12. How does causality help inverse RL to get better performance than the expert?

Answer. By utilizing causal relationships embedded in the environment, the learner is able to address the bias/distribution drift due to the presence of unobserved confounding in demonstration data. Once the adjustment formula is obtained (i.e., Thms. 1 and 2), IRL algorithms could produce a policy that substantially outperforms the expert by exploiting parametric knowledge about the reward, e.g., linearity (Syed & Schapire, 2008) or through reward augmentation (Li et al., 2017). In this work, one of our main contributions is to extend existing IRL approaches to more generalized settings where the expert's demonstrations are imperfect, and the phenomenon of unobserved confounding exists (i.e., Markovianity cannot be assumed). To further ground our contributions, we summarize the current literature on imitation learning as follows, in order to highlight that our work fills a critical literature gap where unobserved confounders exist in the sequential imitating processes and the expert is sub-optimal:

	Optimal Expert	**Sub-optimal Expert**
Unconfounded	Behavior Cloning (BC): Widrow (1964); Pomerleau (1989); Muller et al. (2006); Mülling et al. (2013); Mahler & Goldberg (2017)	Invserse Reinforcement Learning (IRL): Ng et al. (2000); Ziebart et al. (2008); Ho & Ermon (2016); Fu et al. (2017)
Confounded	Causal BC: Zhang et al. (2020); Kumor et al. (2021)	Causal IRL: **our work**

To sum up, behavior cloning (BC) is able to achieve satisfactory performance when the demonstration data is unconfounded, and the expert is (near) optimal. Inverse RL (IRL) can outperform the suboptimal expert by exploiting additional parametric knowledge about the reward signal. However, both BC and IRL are fragile to data drift due to the presence of unobserved confounders. Causal BC (Zhang et al., 2020; Kumor et al., 2021) produces a confounding-robust policy but is still limited by the performance of a suboptimal expert. Finally, this paper empowers IRL methods with the causal inference theory so that the IRL imitator could produce a policy robust to both unobserved confounding and the suboptimal expert.

Q13. What is the problem formulation of causal IRL?

Answer. Details of the problem formulation have been described in Sec. 2. Formally, the problem of causal IRL can be summarized as follows:

- **Input:** (1) the expert's demonstrations, summarized as the observational distribution $P(\boldsymbol{O})$, and (2) a causal diagram $\mathcal{G}$ associated with the underlying SCM.
- **Output:** a policy $\boldsymbol{\pi}^* = \{\pi_1^*(X_1 \mid \boldsymbol{Z}_1), \ldots, \pi_n^*(X_n \mid \boldsymbol{Z}_n)\}$ determining actions $\boldsymbol{X}$.
- **Objective:** to learn an *imitating policy* $\boldsymbol{\pi}^*$ that achieves the expert's performance, i.e., $\mathbb{E}[Y \mid \text{do}(\boldsymbol{\pi}^*)] \geq \mathbb{E}[Y]$.

Q14. Could the authors provide a table of notations?

Answer. Yes, we provide in Table 2 a summary of important notations used in this paper.

Table 2: Summary of Notations

Notations	**Meanings**
Preliminaries:	
M	Structural Causal Model
$\boldsymbol{O}$	observed variables
$\boldsymbol{L}$	latent variables
$\boldsymbol{U}$	exogenous variables
Causal IRL:	
$\boldsymbol{X}$	actions
$\boldsymbol{X}^*$	effective actions
$\boldsymbol{Z}$	covariates
$\boldsymbol{Z}^*$	effective covariates
Y	latent reward
Expert:	
$f_{\boldsymbol{X}}$	behavior policy or expert policy
$\rho(\boldsymbol{x}^*, \boldsymbol{z}^*)$	occupancy measure of the expert
$\mathbb{E}_M[Y]$	expert performance
Imitator:	
$M_{\boldsymbol{\pi}}$	a submodel from M where $f_{\boldsymbol{X}}$ replaced by $\boldsymbol{\pi}$
$\mathcal{S}$	policy scope
$\text{do}(\boldsymbol{x})$	*atomic intervention* on $\boldsymbol{X}$ by constants $\boldsymbol{x}$
$\text{do}(\boldsymbol{\pi})$	*policy intervention* on actions $\boldsymbol{X}$ following $\boldsymbol{\pi}$
$\rho_{\boldsymbol{\pi}}(\boldsymbol{x}^*, \boldsymbol{z}^*)$	occupancy measure of the imitator
$\mathbb{E}_M[Y \mid \text{do}(\boldsymbol{\pi}^*)]$	imitator performance
Graph Notations:	
$\mathcal{G}$	causal diagram associated with M
$\mathcal{G}_{\mathcal{S}}$	causal diagram associated with submodel $M_{\boldsymbol{\pi}}$
pa, ch, de, an	parents, children, descendants, and ancestors
Pa, Ch, De, An	parents with the argument, children with the argument, descendants with the argument, and ancestors with the argument