Initial State Interventions for Deconfounded Imitation Learning

Samuel Pfrommer, Yatong Bai, Hyunin Lee, Somayeh Sojoudi

***Abstract*— Imitation learning suffers from *causal confusion*. This phenomenon occurs when learned policies attend to features that do not causally influence the expert actions but are instead spuriously correlated. Causally confused agents produce low open-loop supervised loss but poor closed-loop performance upon deployment. We consider the problem of masking observed confounders in a disentangled representation of the observation space. Our novel masking algorithm leverages the usual ability to intervene in the initial system state, avoiding any requirement involving expert querying, expert reward functions, or causal graph specification. Under certain assumptions, we theoretically prove that this algorithm is *conservative* in the sense that it does not incorrectly mask observations that causally influence the expert; furthermore, intervening on the initial state serves to strictly reduce excess conservatism. The masking algorithm is applied to behavioral cloning for two illustrative control systems: CartPole and Reacher.**

I. INTRODUCTION

Imitation learning aims to train an intelligent agent to mimic expert demonstrations for a particular task. Various imitation learning instantiations, such as behavior cloning and inverse reinforcement learning, have been widely applied to fields including robotics [1, 2], autonomous driving [3, 4], and optimal navigation [5, 6]. Imitation learning enables agents to learn from high-quality samples instead of exploring from scratch, leading to significantly higher learning efficiency when compared with reinforcement learning methods [7]. This is especially important in safety-critical settings where reinforcement learning are difficult to execute [8, 9]. Even when the flexibility of reinforcement learning is desired, imitation learning can be used to accelerate the learning process [10].

Despite its broad applicability, imitation learning exhibits an issue known as *causal confusion* [11]: the learned policy misattributes features which are primarily *correlated* with expert actions as reflecting a *causal* relationship [12]. This can manifest itself both through the observed features which are spuriously correlated with the expert actions ("nuisance variables") as well as confounders which are available to the expert but not the imitator ("unobserved confounders"). We restrict ourselves to the former, although for completeness we include approaches addressing the latter in our work.

Consider an illustrative example of causal confusion adapted from [11]. The task at hand is learning to drive a car from expert demonstrations. A behavior cloning agent is provided video observations from the driver's point of view, including a brake light on the vehicle dashboard. Although the learned braking policy is excellent on the supervised dataset, upon deployment agent performance suffers: the agent has effectively learned the trivial policy of braking when the brake light is on, instead of attending to other pedestrians or vehicles. In this case, the brake light is a "nuisance variable," and we can dramatically improve the performance of the policy by covering the brake light and reducing information for the model.

Existing approaches for completely masking such nuisance variables generally require either a queryable expert or access to the expert reward function. The seminal work of [11] introduced a β-VAE decomposition the observation space along with a joint policy parameterized by hypothetical causal structures. The space of causal structures can then be searched with two distinct algorithms, one leveraging expert queries and the other based on policy evaluations and reward feedback. The existence of nuisance variables was also noted [13] as part of a broader issue with sequential models that can be addressed with Dagger-style expert queries [14]. The work of [15] partially addresses the nuisance variable problem by regularizing the learned policies to attend to multiple objects in the scene. While this approach does not require policy executions, it only weakens the learner's attention to a nuisance variable and does not eliminate it completely.

The complementary problem of unobserved confounders considers the setting where experts observe confounding variables that are inaccessible to the learner. In the car driving example, this might include a human driver listening to honking that is not detected with visual sensors. One exciting theoretical line of research in this area [16, 17] presents causal-model derived conditions for imitability and an algorithm for imitating the expert policy when possible. However, these works make the strong assumption that the causal graph is provided to the imitation learning agent. Other efforts to apply causal inference techniques to the unobserved confounder problem either require strong assumptions, such as the knowledge of the expert reward [18] and purely additive temporally correlated noise [19], or only evaluate simple multi-armed bandit problems [20].

This work focuses on the problem of observed nuisance variables. Our approach, presented in Section III leverages initial state interventions to identify and completely mask causally confusing features without relying on expert queries or policy interventions. We provide *conservativeness* guarantees for our method in Section IV and present illustrate experiments in Section V.

All authors are with the Department of Electrical Engineering and Computer Sciences, University of California Berkeley, Berkeley, CA, 94720. sam.pfrommer@berkeley.edu; yatong_bai@berkeley.edu; hyunin@berkeley.edu; sojoudi@berkeley.edu

The full technical report, including proofs, can be found at: https://sam.pfrommer.us/wp-content/uploads/2023/03/main.pdf.

II. NOTATION AND BACKGROUND

We denote the set of real numbers by $\mathbb{R}$ and the set of natural numbers by $\mathbb{N}$. The set $\{1, \ldots, a\} \subset \mathbb{N}$ is denoted by $[a]$ for $a \in \mathbb{N}$, and similarly $a, \ldots, b \subset \mathbb{N}$ is denoted by $[a \,..\, b]$. For a pair of boolean variables x and y, the notation $\wedge$ denotes the "and" operator while $\vee$ denotes "or." For a set of boolean variables $\{x_1, x_2, \ldots, x_n\}$, the notations $\bigwedge_{i=1}^n x_i$ and $\bigvee_{i=1}^n x_i$ denote $x_1 \wedge x_2 \wedge \ldots \wedge x_n$ and $x_1 \vee x_2 \vee \ldots \vee x_n$, respectively. The logical negation of a boolean variable or vector x is denoted by $\neg x$. We denote the identically zero function on a domain by $\mathbf{0}$, and we write $f(\cdot) \not\equiv \mathbf{0}$ to mean that $f(\cdot)$ is not equivalent to the zero function over its argument—i.e., there exists an input where f is nonzero.

A. Measure theory and probability

For a random variable X, we introduce the notation $P(x)$ to represent a probability measure over the values x in the domain of X. The uniform measure over an interval $[a, b] \subset \mathbb{R}$ is denoted by $\mathcal{U}(a, b)$. For two measures μ and ν, we say that ν is absolutely continuous with respect to μ if for every μ-measurable set A, $\mu(A) = 0$ implies $\nu(A) = 0$. If ν is absolutely continuous with respect to μ, we let $d\nu/d\mu$ denote the Radon-Nikodym derivative of ν with respect to μ. Note that $d\nu/d\mu$ is a nonnegative function. The standard Lebesgue measure on $\mathbb{R}$ is denoted λ. For a measure μ which is absolutely continuous with respect to λ, we define its L_1 norm in the typical manner

$$\|\mu\|_1 := \int \left| \frac{d\mu}{d\lambda} \right| d\lambda,$$

which we take to be the default norm in the Banach space of measures on $\mathbb{R}$. We denote independence between two random variables using $\perp\!\!\!\perp$ and its negation by $\not\perp\!\!\!\perp$.

B. Causal graphs and structural causal models

We denote a directed acyclic graph by $\mathcal{G}$, with the presence of a direct edge between nodes X and Y denoted $X \to Y$. For a given node X in $\mathcal{G}$, we let $\mathcal{G}_{\underline{X}}$ denote the graph obtained by deleting outgoing edges from X. We denote sets of nodes in a graph using bold font (e.g., $\mathbf{Z}$). The set of parents of a node X in a graph is denoted by $\mathbf{pa}_X$. A path between two nodes X and Y can consist of arbitrarily directed edges and is said to be "blocked" by a set of nodes $\mathbf{Z}$ if the path contains a chain $I \to M \to J$ or a fork $I \leftarrow M \to J$ with $M \in \mathbf{Z}$ or a collider $I \to M \leftarrow J$ such that $M \notin \mathbf{Z}$ and no descendent of M is in $\mathbf{Z}$ [21]. Two nodes X and Y are said to be d-separated by $\mathbf{Z}$ if $\mathbf{Z}$ blocks every path between X and Y. We call a path with all edges oriented the same direction a directed path.

We leverage Pearl's structural causal model (SCM) formalism [21]. An SCM $\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F} \rangle$ consists of endogenous variables $\mathbf{V}$, exogenous variables $\mathbf{U}$, and structural equations $\mathcal{F}$. Each $V \in \mathbf{V}$ is considered to be a node in the causal graph G with one associated exogenous variable $U_V \in \mathbf{U}$ which is independently distributed. The structural equations $f_V \in \mathcal{F}$ assign values of a particular node $V \in \mathbf{V}$ as a function $V := f_V(\mathbf{pa}_V, U_V)$ of its parents and associated exogenous variable. The SCM $\mathcal{M}$ induces a joint distribution $P(\mathbf{v})$ over the endogenous variables $\mathbf{V}$. We say that an SCM $\mathcal{M}$ is *faithful* to its causal graph $\mathcal{G}$ if the distribution $P(\mathbf{v})$ induced by $\mathcal{M}$ contains only the pairwise conditional independencies implied by $\mathcal{G}$; i.e. $X \perp\!\!\!\perp Y \mid \mathbf{Z}$ in the joint distribution from $\mathcal{M}$ iff X and Y are d-separated by $\mathbf{Z}$ in $\mathcal{G}$ [22]. As a notable special case, if $\mathbf{Z}$ is empty and there exists a path from X to Y with no colliders then $X \not\perp\!\!\!\perp Y$.

We define an intervention on a particular node V to be a reassignment of the associated structural equation f_V. This intervention can take the form of a constant intervention $V := v$, which we denote by $\mathrm{do}(V = v)$ for a constant v and may abbreviate to $\mathrm{do}(v)$. We also define a distributional intervention, denoted by $\mathrm{do}(V \sim \tilde{P}(v))$, where we assign V to be drawn from a specified distribution $\tilde{P}(v)$. We denote the post-intervention SCM by $\widetilde{\mathcal{M}}$, with an associated causal graph $\widetilde{\mathcal{G}}$ identical to $\mathcal{G}$ but with incoming edges to V removed. Note that reassigning the associated structural equation for any particular node V induces a new distribution generated by $\widetilde{\mathcal{M}}$ over the set of all endogenous variables $\mathbf{V}$, which we denote by $P(\mathbf{v} \mid \mathrm{do}(V = v))$ or $P(\mathbf{v} \mid \mathrm{do}(V \sim \tilde{P}(v)))$. We may abbreviate a constant intervention $\mathrm{do}(V = v)$ as simply $\mathrm{do}(v)$, where it is clear that v is associated with the uppercase V.

C. Imitation learning

The goal of imitation learning is to learn an agent that replicates some expert behavior. We specifically focus on behavior cloning, which uses collected expert trajectories from random initializations to train an imitating policy. Precise details are formalized in the remainder of this section.

For the system of interest, we use $d_{\mathcal{S}}$, $d_{\mathcal{I}}$, $d_{\mathcal{O}}$, and $d_{\mathcal{A}}$ to denote the dimensionality of the bounded state space $\mathcal{S} \subseteq \mathbb{R}^{d_{\mathcal{S}}}$, raw image observation space $\mathcal{I} \subseteq \mathbb{R}^{d_{\mathcal{I}}}$, disentangled observation space $\mathcal{O} \subseteq \mathbb{R}^{d_{\mathcal{O}}}$, and action space $\mathcal{A} \subseteq \mathbb{R}^{d_{\mathcal{A}}}$. Let S_t, I_t, O_t, and A_t be vector random variables taking on values in $\mathcal{S}$, $\mathcal{I}$, $\mathcal{O}$, and $\mathcal{A}$, respectively, for a discrete time step $t \in \mathbb{N}$. States variables S_t represent the intrinsic low-dimensional dynamics of the system (e.g. simulator variables) while observations O_t are distilled using a β-VAE style framework from high-dimensional image measurements I_t, with typically $d_{\mathcal{I}} \gg d_{\mathcal{O}}$—although we specify images for concreteness, our approach generalizes to any high-dimensional observation space with a low-dimensional disentangled structure. The system dynamics follow the typical assumption that S_{t+1} is strictly a function of S_t and A_t, excluding I_t and O_t, with initial time step $t = 1$.

The imitation learning agent is executed using only the observed images I_t and not the full state S_t, although we assume state variables are available for the training dataset. This naturally models a typical sim-to-real transfer scenario or the setting where a training-time sensor suite is reduced at test time due to budget constraints. Note that although we assume the input to our imitation learning policies lies in the high-dimensional images space $\mathcal{I}$, in our approach the policy first applies a β-VAE style compression on an input image to mask in the latent space of disentangled observations O_t.

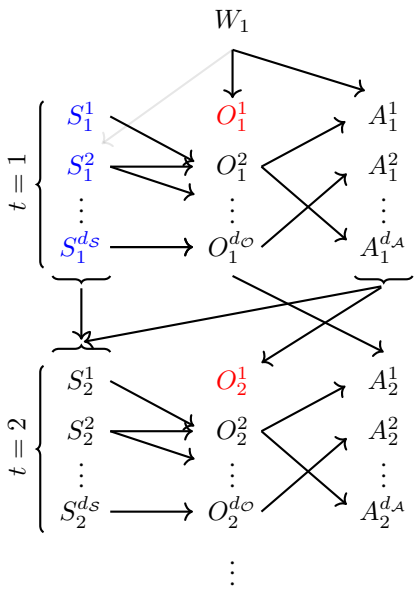


Fig. 1: An example (unknown) system causal graph $\mathcal{G}_s$. We hope to mask O^1 (e.g. brake light observation), which has no causal edge to any expert action but is correlated with A^1 through the confounding random "seed" W_1 and future spurious correlations. In $\mathcal{G}_s$, W_1 also causally influences S_1^2; however, if we intervene on S_1 (blue) this edge is removed in $\widetilde{\mathcal{G}}_s$ (light shading). This enables our masking algorithm to more reliably leverage state initialization to detect potential causes between observations and actions (Section III-B).

Lower-case script letters $\mathfrak{s} \in [d_\mathcal{S}]$, $\mathfrak{o} \in [d_\mathcal{O}]$, and $\mathfrak{a} \in [d_\mathcal{A}]$ denote specific indices in the state, observation, and action vectors. For example, $S_1^\mathfrak{s}$ refers to the real-valued random variable corresponding to the $\mathfrak{s}^{\text{th}}$ state variable at the first time step. We let W_t denote the concatenation of all states, observations, and actions up to but not including the t^{th} time step:

$$W_t = [S_1, \ldots, S_{t-1}, O_1, \ldots, O_{t-1}, A_1, \ldots, A_{t-1}] \cup W_1,$$

where we model $W_1 \sim \mathcal{U}(a, b)$ to be an unobserved variable capturing initialization stochasticity which is outside of our control (i.e. a random "seed").

The collection of states, observations, and actions, along with W_1, can be considered exogenous variables in an SCM defining our system. We denote the system SCM by $\mathcal{M}_s$ and denote the corresponding faithful causal graph by $\mathcal{G}_s$. Note that the SCM depends on the choice of policy. Since we aim to infer causalities regarding the expert policy, we generally let any causal relationships be relative to the $\mathcal{M}_s$ and $\mathcal{G}_s$ induced by the expert policy. We denote our system SCM to be the tuple $\langle \mathcal{M}_s, \mathcal{G}_s \rangle$. Although nodes in $\mathcal{G}_s$ are individual elements in our vector-valued random variables (i.e., $S_t^\mathfrak{s}$ is a node, not S_t), with some abuse of notation, we let the edge symbol $S_t \to X$ signify that $S_t^\mathfrak{s} \to X$ for some $\mathfrak{s} \in [d_\mathcal{S}]$. Similarly, $X \to S_t$ denotes that $X \to S_t^\mathfrak{s}$ for some $\mathfrak{s}$.

This work evaluates the importance of intervening on the initial state to assign it to a particular distribution $S_1 \sim \tilde{P}(s_1)$. This intervention yields a modified SCM $\widetilde{\mathcal{M}}_s$ with a corresponding (not necessarily faithful) causal graph $\widetilde{\mathcal{G}}_s$, which removes the edge $W_1 \to S_1$ in $\mathcal{G}_s$ (Figure 1). We collect N arbitrary-length expert rollout trajectories of states, observations, and actions from $\widetilde{\mathcal{M}}_s$. The collection of all such trajectories is denoted ${}^{(1..N)}\boldsymbol{\tau}$. Among these N trajectories, the i^{th} trajectory consists of the tuple

$${}^{i}\boldsymbol{\tau} = \langle s_1, \ldots, s_T;\ \imath_1, \ldots, \imath_T;\ o_1, \ldots, o_T;\ a_1, \ldots, a_T \rangle,$$

where T is the length of this trajectory, and the lowercase letter corresponding to a random variable represents the concrete value (to avoid confusion with indices, for raw image observations, we use $\imath_t$ to denote a value of I_t). Note that implicit in this characterization of a trajectory is the existence of an *encoder* $\psi_e : \mathcal{I} \to \mathcal{O}$ mapping each image $\imath_t$ to a disentangled observation o_t. While we characterize trajectories as containing both images and disentangled observations to simplify the exposition, in practice, the real-world or the simulator data collection process only provides the images $\imath_t$, and the extraction of disentangled observations o_t is internal to methods that benefit from this representation. As our masking algorithm operates strictly on disentangled representations, we remove from focus the raw images $\imath_t$ unless discussing the policy training.

When training imitation learning agents on ${}^{(1..N)}\boldsymbol{\tau}$, we parameterize policies as a neural network $f_\theta : \mathcal{I}^L \to \mathcal{A}$. The neural policy maps some history of observations to the an action a_t via the calculation

$$a_t = f_\theta(\imath_t, \imath_{t-1}, \ldots, \imath_{t-L+1}). \tag{1}$$

We then train f_θ via standard behavior cloning by randomly sampling batches of images and expert actions from ${}^{(1..N)}\boldsymbol{\tau}$ and performing supervised regression.

D. Statistical independence tests

Our method relies on identifying whether two random variables are statistically dependent. While this is a challenging problem with a rich literature [23], in this paper, we only briefly introduce a well-known independence test for continuous distributions based on Hoeffding's D statistic [24, 25]. Consider two real-valued random variables X and Y with a joint cumulative distribution function $F(x, y) = P\big(X \leq x, Y \leq y\big)$. Hoeffding's D statistic operates on N_{Hoeff} independent pairs of observations $\{(X_1, Y_1), \ldots (X_{N_{\text{Hoeff}}}, Y_{N_{\text{Hoeff}}})\}$ and outputs a real number D in the range $[-0.5, 1]$, with $D > 0$ indicating dependence. The computational complexity of calculating this statistic is $\mathcal{O}(N_{\text{Hoeff}} \log N_{\text{Hoeff}})$. For absolutely continuous joint distributions, the D statistic is unbiased and consistent as $N_{\text{Hoeff}} \to \infty$, meaning that the dependence is correctly represented with probability arbitrarily close to 1. Subsequent variations of the D statistic maintain consistency even for non-absolutely continuous joint distributions [26], although these complications are outside the scope of our work. We refer to the independence test based on the Hoeffding's D statistic as Hoeffding's independence test.

We invoke Hoeffding's independence test over a dataset of trajectories ${}^{(1..N)}\boldsymbol{\tau}$, extracting exactly one pair of variables from each trajectory ($N_{\text{Hoeff}} = N$). For concreteness, consider the call HOEFFDING($S_1^2 \not\perp\!\!\!\perp A_3^4$ in ${}^{(1..N)}\boldsymbol{\tau}$). This extracts, from each trajectory, the second element of the $t = 1$ state and the fourth element of the $t = 3$ action. These N pairs are then supplied to Hoeffding's test, which returns a real number in the range $[-0.5, 1]$.

III. PROBLEM STATEMENT AND METHOD

We address the *causal confusion* problem in imitation learning and aim to mask spuriously correlated observations. To this end, we investigate the following problem statement:

How can we identify and eliminate spuriously correlated observations without relying on online expert queries or knowledge of the expert reward function?

Our approach addresses this problem in a theoretically-grounded way. Specifically, we make the following contributions:

1) We present an algorithm for identifying and masking causally confusing observations *without relying on reward function knowledge, expert queries, or causal graph knowledge*.
2) We prove that, under certain conditions, our procedure is *conservative*: if an observation causally affects the expert actions, it will not be masked.
3) We demonstrate the importance of *initial state interventions* by showing theoretically that the interventions reduce excess conservatism in the masking algorithm.

Section III-A presents and analyzes the assumptions underlying our method. Section III-B motivates and derives our method, which is then presented formally in Section III-C.

A. Assumptions

Our proposed method relies on the following assumptions to ensure the theoretical guarantees in Section IV.

Assumption 1. The system causal graph $\mathcal{G}_s$ is time invariant. Namely, consider two arbitrary time steps $t, t' \in \mathbb{N}$ with $t' \geq t$ and two arbitrary time-indexed variables X_t and $Y_{t'}$ in $\mathcal{G}_s$. Then if $X_t \to Y_{t'}$ is an edge in $\mathcal{G}_s$, then so is $X_{t+\Delta} \to Y_{t'+\Delta}$ for any $\Delta \in \mathbb{Z}$ such that $\min(t + \Delta, t' + \Delta) \geq 1$.

Time-invariance of the expert policy allows for causal inference via interventions on the initial state S_1. Otherwise we would require the ability to intervene at arbitrary time steps, which is unrealistic for most real-world systems.

Assumption 2. The expert policy attends only to observational information derived from the underlying state. Namely, if $O_t^o \to A_{t'}^a$ in $\mathcal{G}_s$ for $t, t' \in \mathbb{N}$ with $t' \geq t$, then there must exist an index $\mathfrak{s}$ such that $S_t^{\mathfrak{s}} \to O_t^o$.

Assumption 2 reflects the intuition that the expert policy itself must not be fooled by spurious information in the observation space. This is a natural assumption in the considered case where the dynamics of the underlying system depend only on S_t, not O_t.

Assumption 3. The expert policy reacts to observations within a *reaction horizon* $H \in \mathbb{N}$. Specifically, if $O_t^o \to A_{t_1}^a$ in $\mathcal{G}_s$ for some $t_1 > t$ and particular $t \in \mathbb{N}$, $o \in [d_{\mathcal{O}}]$, and $a \in [d_{\mathcal{A}}]$, then there exists a $t_2 \in [t \,..\, t + H - 1]$ such that $O_t^o \to A_{t_2}^a$.

Assumption 3 imposes a horizon within which the expert is assumed to react to a hypothetical intervention on a state or observation. For finite-length trajectories, H can be chosen to be the entire trajectory length, with the algorithm and theory still valid. As such, H introduces a hyperparameter that allows for more tractable computation under some assumptions on the expert. Our experiments show that H can be much smaller than the trajectory length for certain practical dynamic systems and experts.

Finally, we formalize a class of SCMs that behave nicely under interventions.

Assumption 4. The system SCM $\mathcal{M}_s = \langle \mathbf{V}, \mathbf{U}, \mathcal{F} \rangle$ is *interventionally absolutely continuous*, meaning that for any disjoint sets of nodes $\mathbf{X}$ and $\mathbf{Y}$, the interventional distribution $P\big(\mathbf{y} \mid \text{do}(\mathbf{X} = \mathbf{x})\big)$ is absolutely continuous with respect to the Lebesgue measure and has a bounded Radon-Nikodym derivative.

Assumption 4 stipulates that the probability distribution induced by our SCM on any set of non-intervened nodes is absolutely continuous with bounded density. This is a technical condition that allows us to assert that Hoeffding's test is consistent. We note that subsequent D-statistic variations allow for non-absolutely continuous joint distributions [26] — we leave the theoretical and practical implications of more sophisticated testing to future work.

B. Derivation

Our aim is to mask a particular observation O^o across all time steps if it has no causal effect on any expert action within the reaction horizon. As intervening on observations is impractical, this causality is challenging to deduce. We do, however, assume the ability to intervene on the system in one specific instance: setting the state variables S_1 at initialization. We manipulate S_1 to infer the possible existence of a true causal relationship.

We first motivate our approach from an arbitrary time step $t \geq 2$ before specializing on the initialization. Consider arbitrary observation and action indices $o \in [d_{\mathcal{O}}], a \in [d_{\mathcal{A}}]$ and time steps $t, t' \in \mathbb{N}$ with $t' \in [t \,..\, t+H-1]$. Assumption 2 states that a causal effect $O_t^o \to A_{t'}^a$ must arise from a larger causal path

$$S_t^{\mathfrak{s}} \to O_t^o \to A_{t'}^a \tag{2}$$

in $\mathcal{G}_s$, for some state variable index $\mathfrak{s} \in [d_{\mathcal{S}}]$. We now observe that by faithfulness of $\langle \mathcal{M}_s, \mathcal{G}_s \rangle$ it must be that $S_t^{\mathfrak{s}} \not\perp\!\!\!\perp O_t^o$ and $S_t^{\mathfrak{s}} \not\perp\!\!\!\perp A_{t'}^a$; i.e. the causal relationships in $\mathcal{G}_s$ imply probabilistic dependencies in the induced distribution from $\mathcal{M}_s$. Note that these are *statistical* statements which can be ascertained from the observational data. We define the

boolean variable ${}^{(t,t')}D^{o}_{\mathfrak{d},a}$ to check these independencies:

$$ {}^{(t,t')}D^{o}_{\mathfrak{d},a} \coloneqq (S^{\mathfrak{d}}_t \not\perp\!\!\!\perp O^{o}_t) \wedge (S^{\mathfrak{d}}_t \not\perp\!\!\!\perp A^{a}_{t'}), \tag{3} $$

and introduce the "potential cause" notation

$$ O^{o}_t \dashrightarrow A^{a}_{t'} \coloneqq \bigvee_{\mathfrak{d}=1}^{d_{\mathcal{S}}} \left({}^{(t,t')}D^{o}_{\mathfrak{d},a} \right). \tag{4} $$

The boolean-valued statement $O^{o}_t \dashrightarrow A^{a}_{t'}$ intuitively captures that, based on observational data, there may (but need not) exist a true causal edge $O^{o}_t \to A^{a}_{t'}$ generated by some $S^{\mathfrak{d}}_t$ as in (2). We denote by $O^{o}_t \not\dashrightarrow A^{a}_{t'}$ the logical negation of $O^{o}_t \dashrightarrow A^{a}_{t'}$. As we will elaborate in more detail shortly, if $O^{o}_t \not\dashrightarrow A^{a}_{t'}$ for all actions $a \in [d_{\mathcal{A}}]$ and t' in the reaction horizon, we want to "mask" the o^{th} observation as it has no causal effect on the expert action but could be spuriously correlated in a way that undermines the imitation learning policy performance.

It is immediate from the above faithfulness argument that for $t \geq 2$, we have the implication

$$ O^{o}_t \to A^{a}_{t'} \implies O^{o}_t \dashrightarrow A^{a}_{t'}. \tag{5} $$

Note that (5) provides a *conservativeness* guarantee: if an observation causally influences an action, we will not mistakenly conclude from observational data that it does not, and hence incorrectly mask an observation that is actually used by the expert policy. However, this conservativeness is not apparent for $t = 1$ in the modified causal model $\langle \widetilde{\mathcal{M}}_s, \widetilde{\mathcal{G}}_s \rangle$, where we intervene to specify the initial state distribution, overriding the natural randomness resulting from W_1 and potentially breaking the faithfulness. As a simple counterexample, initializing S_1 to a constant vector would make $S^{\mathfrak{d}}_1$ independent of every other random variable in the causal graph, and therefore no potential causes could be discovered as (3) would always be false. Nonetheless, when a sufficiently sensible initialization distribution is used, we prove that the conservativeness result still holds under intervention on S_1 in Section IV.

The reverse implication to (5) does not hold. It is possible that spurious statistical relationships exist while a causal edge $O^{o}_t \to A^{a}_{t'}$ does not. Indeed, for $t \geq 2$, the abundance of chronologically antecedent variables virtually guarantees that all variables have share a common cause and hence a statistical dependence. The sole exception is the initial state S_1. By intervening on S_1, we eliminate the incoming edge from the only possible common ancestor W_1 in the causal graph (Figure 1). Therefore, we expect that this interventional ability should help eliminate potential causes $O^{o}_1 \dashrightarrow A^{a}_{t'}$ which do not exist in the true causal graph and reduce excessive conservativeness in the algorithm. We analyze this idea formally in Section IV.

The culmination of our efforts is described in Algorithm 1, which checks for potential causes, as defined in (4), at $t = 1$ using expert data ${}^{(1..N)}\boldsymbol{\tau}$ collected from the interventional system $\langle \widetilde{\mathcal{M}}_s, \widetilde{\mathcal{G}}_s \rangle$. Note that Algorithm 1 invokes the HOEFFDING routine to compute Hoeffding's D statistic for independence between two variables (see Section II-D). The test returns a real number in the range $[-0.5, 1]$, with a value greater than zero indicating dependence. Since perfect observational disentanglement is unrealistic, we introduce a threshold hyperparameter γ, which we set to 0.001 for the experiments.

Algorithm 1 is presented to maximize readability and can be implemented more efficiently. Namely, the Hoeffding tests between $S^{\mathfrak{d}}_t$ and $O^{o}_t, A^{a}_{t'}$ can be precomputed, yielding the runtime

$$ O\left(d_{\mathcal{S}}(d_{\mathcal{O}} + H d_{\mathcal{A}}) N \log N\right), $$

where $N \log N$ is the cost of evaluating Hoeffding's test for a specific pair of variables over N trajectories. In practice, Hoeffding's test executions are very fast—on the order of milliseconds for $N = 10^3$—and incur a negligible overhead compared with the training time of imitation learning.

Remark 1. The reader may have noticed that our approach bears a resemblance to *instrumental variable regression*, a statistical technique for estimating causal relationships that has also received some attention in the causal imitation learning literature [19]. We emphasize that $S^{\mathfrak{d}}_t$ does not constitute a valid instrumental variable in the causal path (2) as there may be many other paths between $S^{\mathfrak{d}}_t$ and $A^{a}_{t'}$ which are not mediated by O^{o}_t. Thus while the spirit of our approach is related to instrumental variable regression, we cannot use $S^{\mathfrak{d}}_t$ to precisely determine a causal relationship between O^{o}_t and $A^{a}_{t'}$ and only use $S^{\mathfrak{d}}_t$ to provide evidence of a potential cause.

C. Imitation Learning Workflow

Drawing on the masking approach developed in Section III-B, we summarize our overall deconfounded imitation learning workflow as the following four steps.

1) Collect random-policy trajectories to learn a disentangled observation representation using a β-VAE, denoted by $\psi_d \circ \psi_e : \mathcal{I} \to \mathcal{I}$, with an encoder $\psi_e : \mathcal{I} \to \mathcal{O}$ and decoder $\psi_d : \mathcal{O} \to \mathcal{I}$. For a well-trained β-VAE, $\psi_d \circ \psi_e$ approximates the identity.
2) Collect a sequence of N trajectories ${}^{(1..N)}\boldsymbol{\tau}$ from the expert policy, with the starting state distribution $\tilde{P}(s_1)$ over $\mathcal{S}$ having any density that is everywhere nonzero (e.g. uniform).
3) Execute Algorithm 1 on ${}^{(1..N)}\boldsymbol{\tau}$ to obtain the observation mask $\widetilde{m} \in \{0,1\}^{d_{\mathcal{O}}}$, where $\widetilde{m}_o = 1$ if the o^{th} observation is to be masked.
4) Train the final policy $g_\theta : \mathcal{I}^L \to \mathcal{A}$ on ${}^{(1..N)}\boldsymbol{\tau}$ using standard supervised learning; g_θ masks the disentangled observation space using $\widetilde{m}$ before executing a learnable policy network f_θ:

$$ g_\theta(\imath_t, \ldots, \imath_{t-L+1}) = f_\theta(\tilde{\psi}(\imath_t), \ldots, \tilde{\psi}(\imath_{t-L+1})), $$

where the masked β-VAE $\tilde{\psi} : \mathcal{I} \to \mathcal{I}$ has its weights fixed and is defined as

$$ \tilde{\psi}(\imath) = \psi_d(\neg\widetilde{m} \odot \psi_e(\imath)). $$

Note that this overall structure generally follows the seminal work of [11]. Our key contribution is Algorithm 1,

which provides a mask for the disentangled observations without relying on expert queries, the expert reward function, or specification of the causal graph. A visualization of Algorithm 1 is provided in Figure 2 for the CartPole system considered in the experiments. We show in Section IV that Algorithm 1 enjoys notable theoretical guarantees.

Algorithm 1 Masking algorithm

Hyperparameter $\gamma > 0$.

procedure MASK($^{(1..N)}\boldsymbol{\tau}$)

Initialize $\widetilde{m} \in \{0,1\}^{d_{\mathcal{O}}}$ to be an all-zero vector.

for $o = 1, \ldots, d_{\mathcal{O}}$ **do**

Mask the o^{th} observation according to

$$\widetilde{m}_o \leftarrow \left(O_1^o \not\dashrightarrow A_{t'}^a \ \ \forall a \in [d_{\mathcal{A}}], \ \ \forall t' \in [H]\right), \quad (6)$$

computing $O_1^o \not\dashrightarrow A_{t'}^a$ using CHECK.

return $\widetilde{m}$

procedure CHECK$\{O_t^o \dashrightarrow A_{t'}^a\}(^{(1..N)}\boldsymbol{\tau})$

for $s = 1, \ldots, d_{\mathcal{S}}$ **do**

$a \leftarrow$ HOEFFDING$(S_t^s \not\perp\!\!\!\perp O_t^o \text{ in } {}^{(1..N)}\boldsymbol{\tau}) > \gamma$

$b \leftarrow$ HOEFFDING$(S_t^s \not\perp\!\!\!\perp A_{t'}^a \text{ in } {}^{(1..N)}\boldsymbol{\tau}) > \gamma$

if $a \wedge b$ **then**

return True

return False

IV. THEORETICAL GUARANTEES

In this section, we delve into the theoretical properties of Algorithm 1. Theorem 1 demonstrates that if we intervene on the initial state S_1 and meet certain conditions in the infinite-trajectory regime, the algorithm remains *conservative*, ensuring that no observation that causally influences the expert is mistakenly masked. Additionally, Theorem 2 and Proposition 3 highlight the effectiveness of intervening on S_1 in mitigating overconservativeness in the masking algorithm. Specifically, Theorem 2 asserts that the correctly masked observations under the original causal model $\langle \mathcal{M}_s, \mathcal{G}_s \rangle$ will also be masked under the intervened causal model $\langle \widetilde{\mathcal{M}}_s, \widetilde{\mathcal{G}}_s \rangle$. Proposition 3 showcases a particular set of systems where the intervention only results in masks under $\langle \widetilde{\mathcal{M}}_s, \widetilde{\mathcal{G}}_s \rangle$, providing compelling evidence that the masking algorithm is more effective after intervening on S_1.

All subsequent theory relies on Assumptions 1-4, and for brevity we defer proofs and auxiliary lemmas to the appendix of the full technical report. We now introduce the main conservativeness theorem.

Theorem 1. *In the faithful system causal model $\langle \mathcal{M}_s, \mathcal{G}_s \rangle$, assume that the measure-valued function $w_1 \mapsto P(v \mid \mathrm{do}(\mathbf{Z} = \mathbf{z}), w_1)$ is continuous for any set of nodes $\mathbf{Z}$ and $V \notin \mathbf{Z}$.*

Let there exist a causal edge $O_t^o \to A_{t'}^a$ in $\mathcal{G}_s$ for some $t, t' \in \mathbb{N}$, $t' \geq t$, and indices $o \in [d_{\mathcal{O}}]$ and $a \in [d_{\mathcal{A}}]$. Then in the interventional causal model $\langle \widetilde{\mathcal{M}}_s, \widetilde{\mathcal{G}}_s \rangle$ where the initial state distribution $\tilde{P}(s_1)$ has everywhere-nonzero density on $\mathcal{S}$, O^o is almost surely not masked by Algorithm 1 for almost all uniform parameterizations of W_1 as the number of trajectories $N \to \infty$; i.e., (6) correctly evaluates to true.

Theorem 1 guarantees that Algorithm 1 maintains conservativeness by correctly preserving unmasked observations that causally impact expert actions. This outcome is consistent with the discussion in Section III-B, where we observed that the faithfulness of $\langle \mathcal{M}_s, \mathcal{G}_s \rangle$ ensures the correctness of the algorithm when we do not intervene on S_1 and allow the initial state to be naturally generated from W_1. Theorem 1 establishes that this property also holds in the interventional system $\langle \widetilde{\mathcal{M}}_s, \widetilde{\mathcal{G}}_s \rangle$, where we assign $S_1 \sim \tilde{P}(s_1)$.

We now theoretically demonstrate the benefits of intervening on $\tilde{P}(s_1)$. Specifically, we show that this intervention reduces the excess conservatism in the masking algorithm by removing income edges from W_1 in the causal graph, thereby eliminating a potential avenue of confounding.

Theorem 2. *Let m denote the potential-cause test evaluated by Algorithm 1 on the distribution induced by the non-interventional system $\langle \mathcal{M}_s, \mathcal{G}_s \rangle$, and let $\widetilde{m}$ be the original test on the interventional system $\langle \widetilde{\mathcal{M}}_s, \widetilde{\mathcal{G}}_s \rangle$ where $\tilde{P}(s_1)$ has everywhere-nonzero density on $\mathcal{S}$. Then if m_o correctly evaluates to true for a particular $o \in [d_{\mathcal{O}}]$, then $\widetilde{m}_o$ also evaluates to true almost surely as the number of trajectories $N \to \infty$.*

Theorem 2 assures us that intervening on $\tilde{P}(s_1)$ does not lead to more conservative masking than the original system. We now provide a specific class of SCMs for which the intervention strictly improves the mask.

Proposition 3. *Let $\widetilde{m}$ and m be as in Theorem 2, and consider a particular observation index $o \in [d_{\mathcal{O}}]$ such that the only incoming edge to O_1^o is $W_1 \to O_1^o$. Then if in $\mathcal{G}_s$ there exists the fork $S_1^s \leftarrow W_1 \to O_1^o$ for some $s \in [d_{\mathcal{S}}]$ and a directed path from S_1^s to some A_t^a, with $t \in [H]$, $a \in [d_{\mathcal{A}}]$, $\widetilde{m}_o$ correctly masks the o^{th} observation almost surely as the number of trajectories $N \to \infty$ while m_o does not.*

V. EXPERIMENTS

We evaluate our approach on two custom simulated environments: CartPole and Reacher. Each of these environments contains a nuisance feature which is likely to induce causal confusion. Our masking approach can successfully eliminate these spuriously correlated features. Precise experimental details are deferred to Appendix II of the full technical report.

A. Environments

Both considered environments are modified to include a nuisance feature corresponding to the previous action taken by the expert (analogous to the brake light example). For each environment, the expert is a standard constrained finite-time optimal control policy which minimizes cumulative trajectory loss. This expert reward function is not provided to the imitation learning agent.

CartPole. This environment consists of a standard planar cart-pole system with a continuous scalar horizontal force applied to the cart. A quadratic cost is imposed for deviations

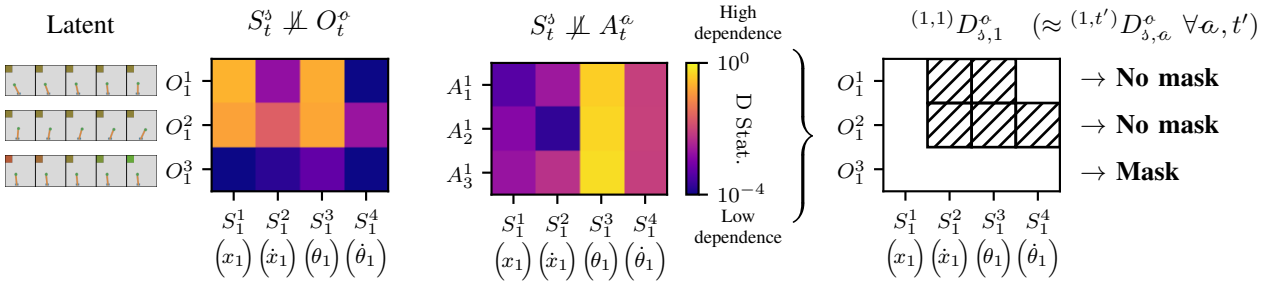


Fig. 2: Masking algorithm visualization for the CartPole environment with reaction horizon $H = 3$. Latent space interpolation of the β-VAE reveals that O^1 and O^2 capture some combined positional/angular information, while O^3 captures the disentangled confounder (color of the confounding square). This last observation shares virtually no dependence (Hoeffding's D statistic less than $\gamma = 10^{-3}$) with any state variable due to interventions on S_1 (note the log scale). This means that $^{(1,t')}D^{o}_{\mathfrak{z},a}$ is false (no cross hatches) for $o = 3$ and all $\mathfrak{z} \in [d_{\mathcal{S}}]$, regardless of a and t'; i.e. $O_1^3 \not\dashrightarrow A_{t'}^{a}$ for all $a \in [d_{\mathcal{A}}]$ and $t' \in [H]$, and we can mask the confounder O^3.

from the vertical target state. The spuriously correlated feature is a colored square in the upper-left corner of each image, which interpolates between green and red depending on the most recently executed action.

Reacher. We consider a top-down version of a two-dimensional two-joint Reacher environment [27]. The environment penalizes squared distance of the end effector to a black target dot. The target location is included in the state vector, thus satisfying Assumption 2. Two torques, one per joint, are specified as the control inputs; the nuisance feature is a red dot in the upper-left corner whose horizontal position and vertical position encode the two control inputs from the previous time step. This "joystick" introduces a different kind of nuisance feature than in the CartPole environment.

B. Imitation learning policies

We compare the performance of our masked policy against that of vanilla behavior cloning. The baseline behavior cloning policy is denoted by BCVANILLA, and our masked policy is denoted by MASKED. For reference, we also measure the performance of the behavior cloning policy with the confounding signals manually removed by superimposing a white square on the upper-left corner, denoted BCMANUAL. We emphasize that BCMANUAL requires human judgement to manually eliminate spurious confounders; we show that we can replicate this performance in a principled and automated way.

C. Discussion

Figure 3 displays our experimental results. Across both environments, the MASKED policy performs comparably to the manually masked baseline BCMANUAL. The BCVANILLA policy incurs a dramatically higher loss than both other policies on CartPole. This is because BCVANILLA incorrectly attends to its own actions at the previous time step due to the introduction of the nuisance feature, leaving it unable to consistently stabilize the inverted pendulum. It is worth noting that MASKED comes close to replicating the manually deconfounded baseline's performance without requiring expert queries, access to the expert reward function, or pre-specified information on the causal graph in the deconfounding procedure. Note, however, that there is a small gap between the performance of our method and manual masking, visible for the Reacher environment. This is likely attributable to imperfect disentanglement in the β-VAE.

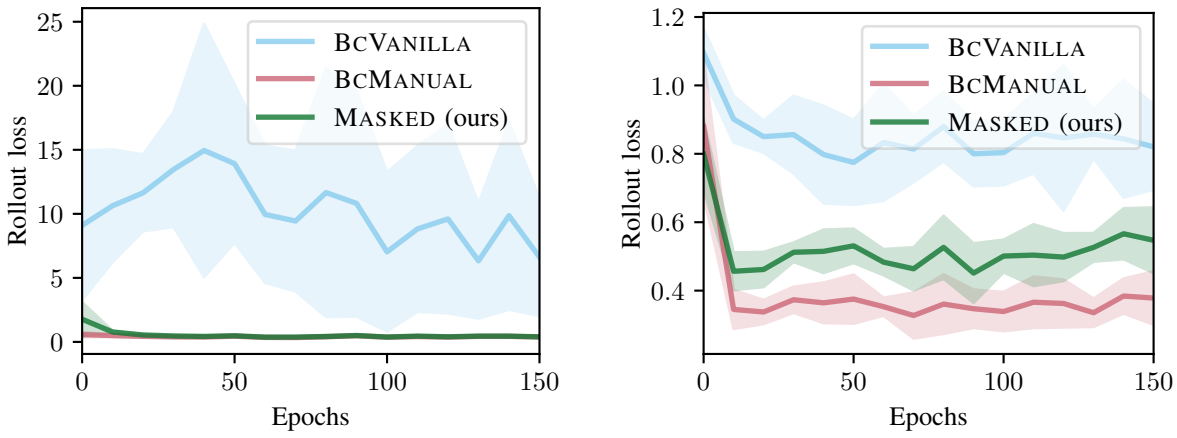
Figure 2 provides a visualization of our masking procedure and the produced mask for the CartPole environment. While we use a latent space size of three (the precise number of independent factors of variation) for visualization purposes, our masking procedure is fully functional for larger choices of the latent size. For Reacher, although there are 6 factors of variation in each image, a larger latent space of 12 yielded superior disentanglement and reconstruction performance.

D. Limitations

The most significant limitation of our work, besides the explicitly stated assumptions, is the requirement that confounding factors are observable and can be neatly disentangled. While this holds for the environments considered in this work, more complex environments may introduce entanglement between causally confusing features and important features to which the expert policy actually attends. We introduce the Hoeffding threshold hyperparameter γ to mitigate this concern; however, investigating more principled methods for handling incomplete disentanglement would be an exciting area of future work.

VI. CONCLUSION

This work introduces a novel method to address the causal confusion problem in imitation learning. The proposed method leverages the typical imitation learning ability to intervene in the initial system state. Unlike previous works, our method masks causally confusing observations without relying on online expert queries, knowledge of the expert reward function, or specification of the causal graph. Our theoretical results establish that our masking algorithm is *conservative*, with excess conservatism strictly reduced by interventions on the initial state. We illustrate the effectiveness



(a) CartPole

(b) Reacher

Fig. 3: Evaluation rollout loss on CartPole (a) and Reacher (b) across training epochs. Solid lines denote mean performance over 5 runs while shaded areas indicate standard deviation. Our MASKED policy approaches the performance of the manually-deconfounded BCMANUAL baseline, while BCVANILLA struggles due to causally confusing features.

of our method with experiments on the CartPole and Reacher environments.

REFERENCES

[1] Sylvain Calinon and Aude Billard. "Incremental learning of gestures by imitation in a humanoid robot". In: *ACM/IEEE International Conference on Human-Robot Interaction*. 2007.

[2] Sanjay Krishnan et al. "SWIRL: A sequential windowed inverse reinforcement learning algorithm for robot tasks with delayed rewards". In: *The International Journal of Robotics Research* 38 (2018), pp. 126–145.

[3] Tianyu Wang, Vikas Dhiman, and Nikolay A. Atanasov. "Inverse reinforcement learning for autonomous navigation via differentiable semantic mapping and planning". In: *arXiv preprint arXiv:2101.00186* (2021).

[4] Alex Kuefler et al. "Imitating driver behavior with generative adversarial networks". In: *IEEE Intelligent Vehicles Symposium*. 2017.

[5] Ahmed Hussein et al. "Deep imitation learning for 3D navigation tasks". In: *Neural computing and applications* 29 (2018), pp. 389–404.

[6] Zhenyu Shou et al. "Optimal passenger-seeking policies on E-hailing platforms using Markov decision process and imitation learning". In: *Transportation Research Part C: Emerging Technologies* 111 (2020), pp. 91–113.

[7] Mariusz Bojarski et al. "End to End Learning for Self-Driving Cars". In: *arXiv preprint arXiv:1604.07316* (2016).

[8] He Yin et al. "Imitation learning with stability and safety guarantees". In: *IEEE Control Systems Letters* 6 (2021), pp. 409–414.

[9] Samuel Pfrommer et al. "Safe reinforcement learning with chance-constrained model predictive control". In: *Learning for Dynamics and Control Conference*. PMLR. 2022, pp. 291–303.

[10] Todd Hester et al. "Deep Q-learning From Demonstrations". In: *AAAI Conference on Artificial Intelligence*. 2017.

[11] Pim De Haan, Dinesh Jayaraman, and Sergey Levine. "Causal confusion in imitation learning". In: *Advances in Neural Information Processing Systems*. 2019.

[12] Jean Kaddour et al. "Causal machine learning: A survey and open problems". In: *arXiv preprint arXiv:2206.15475* (2022).

[13] Pedro A Ortega et al. "Shaking the foundations: delusions in sequence models for interaction and control". In: *arXiv preprint arXiv:2110.10819* (2021).

[14] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. "A reduction of imitation learning and structured prediction to no-regret online learning". In: *International Conference on Artificial Intelligence and Statistics*. 2011.

[15] Jongjin Park et al. "Object-aware regularization for addressing causal confusion in imitation learning". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 3029–3042.

[16] Junzhe Zhang, Daniel Kumor, and Elias Bareinboim. "Causal imitation learning with unobserved confounders". In: *Advances in Neural Information Processing Systems*. 2020.

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

[18] Jalal Etesami and Philipp Geiger. "Causal transfer for imitation learning and decision making under sensor-shift". In: *AAAI Conference on Artificial Intelligence*. 2020.

[19] Gokul Swamy et al. "Causal Imitation Learning under Temporally Correlated Noise". In: *International Conference on Machine Learning*. 2022.

[20] Risto Vuorio et al. "Deconfounded Imitation Learning". In: *arXiv preprint arXiv:2211.02667* (2022).

[21] Judea Pearl. *Causality. Models, Reasoning, and Inference*. 2nd ed. Cambridge University Press, 2009.

[22] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2000.

[23] David J Sheskin. *Handbook of parametric and nonparametric statistical procedures*. crc Press, 2020.

[24] Wassily Hoeffding. "A Non-Parametric Test of Independence". In: *The Annals of Mathematical Statistics* 19.4 (1948), pp. 546–557.

[25] Chaim Even-Zohar. "independence: Fast rank tests". In: *arXiv preprint arXiv:2010.09712* (2020).

[26] Julius R Blum, Jack Kiefer, and Murray Rosenblatt. *Distribution free tests of independence based on the sample distribution function*. Sandia Corporation, 1961.

[27] Greg Brockman et al. "OpenAI Gym". In: *arXiv preprint arXiv:1606.01540* (2016).

[28] Andrew G. Barto, Richard S. Sutton, and Charles W. Anderson. "Neuronlike adaptive elements that can solve difficult learning control problems". In: *IEEE Transactions on Systems, Man, and Cybernetics* SMC-13.5 (1983), pp. 834–846.

[29] Christopher P Burgess et al. "Understanding disentangling in β-VAE". In: *arXiv preprint arXiv:1804.03599* (2018).

[30] A.K Subramanian. *PyTorch-VAE*. `https://github.com/AntixK/PyTorch-VAE`. 2020.

[31] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[32] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

I Proofs for Section IV

Lemma 4. *Consider an SCM $\mathcal{M}$ with a faithful causal graph $\mathcal{G}$ that contains a directed path from X to Y. Then provided a set $\mathbf{Z}$ contains all ancestors of X but none of its descendents, then for any assignment $\mathbf{z}$ to $\mathbf{Z}$ there exist values x, x' such that*

$$\Big\| P\big(y \mid \mathrm{do}(x), \mathbf{z}\big) - P\big(y \mid \mathrm{do}(x'), \mathbf{z}\big) \Big\|_1 > 0,$$

viewed as induced distributions over Y.

Proof: As $\mathbf{Z}$ contains no descendents of X, it cannot block the directed path between X and Y and hence the Causal Markov Condition does not declare X and Y independent. Faithfulness stipulates that X and Y are therefore dependent given $\mathbf{z}$, so there exists x, x' such that

$$\Big\| P\big(y \mid x, \mathbf{z}\big) - P\big(y \mid x', \mathbf{z}\big) \Big\|_1 > 0.$$

The second rule of do calculus states that we can exchange observation and intervention if X and Y are independent given $\mathbf{z}$ in the causal graph $\mathcal{G}_{\underline{X}}$ obtained by removing outgoing edges from X. If we remove outgoing edges from X, the only remaining paths between X and Y must contain an edge $X \leftarrow Z$ for some variable Z. This makes Z an ancestor of X, and therefore Z is included in $\mathbf{Z}$, and both paths of the form $X \leftarrow Z \leftarrow J$ and $X \leftarrow Z \rightarrow J$ are blocked by $\mathbf{Z}$. This means that X and Y are d-separated by $\mathbf{Z}$ in $\mathcal{G}_{\underline{X}}$, and we can apply the second do-calculus rule to conclude that

$$P\big(y \mid \mathrm{do}(x), \mathbf{z}\big) = P\big(y \mid x, \mathbf{z}\big),$$
$$P\big(y \mid \mathrm{do}(x'), \mathbf{z}\big) = P\big(y \mid x', \mathbf{z}\big),$$

and hence

$$\Big\| P\big(y \mid \mathrm{do}(x), \mathbf{z}\big) - P\big(y \mid \mathrm{do}(x'), \mathbf{z}\big) \Big\|_1 > 0.$$

■

Lemma 5. *Consider a set $E \subseteq \mathbb{R}$ where for each $x \in E$, there exists a ball $B(x, \epsilon_x)$ which contains no point in E. Then E has measure zero with respect to the standard Lebesgue measure on $\mathbb{R}$.*

Proof: As E is a subset of $\mathbb{R}$, it is Lindelöf, and the cover of E by the collection of balls $\{B(x, \epsilon_x) \mid x \in E\}$ has a finite subcover. Enumerate this subcover as I_i; we then have

$$\lambda(E) = \lambda(E \cap (\cup_i I_i)) \leq \sum_i \lambda(E \cap I_i) = 0,$$

as each $E \cap I_i$ contains only a singleton. ■

Theorem 1. *In the faithful system causal model $\langle \mathcal{M}_s, \mathcal{G}_s \rangle$, assume that the measure-valued function $w_1 \mapsto P(v \mid \mathrm{do}(\mathbf{X} = \mathbf{x}), w_1)$ is continuous for any set of nodes $\mathbf{X}$ and $V \notin \mathbf{X}$.*

Let there exist a causal edge $O_t^{\mathfrak{o}} \rightarrow A_{t'}^{\mathfrak{a}}$ in $\mathcal{G}_s$ for some $t, t' \in \mathbb{N}$, $t' \geq t$, and indices $\mathfrak{o} \in [d_{\mathcal{O}}]$ and $\mathfrak{a} \in [d_{\mathcal{A}}]$. Then in the interventional causal model $\langle \widetilde{\mathcal{M}}_s, \widetilde{\mathcal{G}}_s \rangle$ where the initial state distribution $\tilde{P}(s_1)$ has everywhere-nonzero density on $\mathcal{S}$, $O^{\mathfrak{o}}$ is almost surely not masked by Algorithm 1 for almost all uniform parameterizations of W_1 as the number of trajectories $N \rightarrow \infty$; i.e., (6) *correctly evaluates to true.*

Proof: By Assumptions 1 and 3, we can WLOG consider $t = 1$ with $t' \in [H]$. If $O_1^{\mathfrak{o}} \rightarrow A_{t'}^{\mathfrak{a}}$, by Assumption 2 there exists an edge $S_1^{\mathfrak{s}} \rightarrow O_1^{\mathfrak{o}}$ for some $\mathfrak{s}$. We now want to show that in the SCM $\widetilde{\mathcal{M}}_s$ where we intervene distributionally on S_1, we have that $S_1^{\mathfrak{s}} \not\perp\!\!\!\perp O_1^{\mathfrak{o}}$ and $S_1^{\mathfrak{s}} \not\perp\!\!\!\perp A_{t'}^{\mathfrak{a}}$. The arguments are similar, so we will just state the proof for the former.

We want to show that $S_1^{\mathfrak{s}}$ and $O_1^{\mathfrak{o}}$ are not independent in $\widetilde{\mathcal{M}}_s$. Note that in the modified structural assignment for $S_1^{\mathfrak{s}}$ in $\widetilde{\mathcal{M}}_s$, $S_1^{\mathfrak{s}}$ is distributed with everywhere-nonzero density on $\mathcal{S}$. Therefore checking the desired independence is equivalent to showing

$$\Big\| P\big(o_1^{\mathfrak{o}} \mid \mathrm{do}(S_1^{\mathfrak{s}} = \alpha)\big) - P\big(o_1^{\mathfrak{o}} \mid \mathrm{do}(S_1^{\mathfrak{s}} = \alpha')\big) \Big\|_1 > 0 \tag{7}$$

as distributions over $o_1^{\mathfrak{o}}$ for some $\alpha, \alpha' \in \mathbb{R}$ with $\alpha \neq \alpha'$. Here, the do statement captures our ability to intervene on the initial state, decoupling any potential correlational influence from W_1.

By Lemma 4, we have that for any particular value w_1 of W_1,

$$\Big\| P\big(o_1^{\mathfrak{o}} \mid \mathrm{do}(S_1^{\mathfrak{s}} = \alpha), w_1\big) - P\big(o_1^{\mathfrak{o}} \mid \mathrm{do}(S_1^{\mathfrak{s}} = \alpha'), w_1\big) \Big\|_1 > 0,$$

for some α, α'. This is equivalent to

$$\|h(\alpha, \alpha', w_1)\|_1 \not\equiv \mathbf{0} \quad \forall w_1, \tag{8}$$

where we define

$$h(\alpha, \alpha', w_1) := P\big(o_1^o \mid \mathrm{do}(S_1^\diamondsuit = \alpha), w_1\big) - P\big(o_1^o \mid \mathrm{do}(S_1^\diamondsuit = \alpha'), w_1\big),$$

and $\mathbf{0}$ denotes an identically zero function over α, α'. Note that $h(\alpha, \alpha', w_1)$ specifies a signed measure over o_1^o. Now observe that

$$P\big(o_1^o \mid \mathrm{do}(S_1^\diamondsuit = \alpha)\big) = \int P\big(o_1^o \mid \mathrm{do}(S_1^\diamondsuit = \alpha), w_1\big) p(w_1) d\mu(w_1),$$

where μ is a probability measure on the unobserved variable w_1 which we will instantiate shortly, and $p(w_1)$ denotes the probability density of W_1, i.e. the Radon-Nikodym derivative of the measure $P(w_1)$. Note that the result of this integral is still a signed measure over o_1^o. So we have that showing our desired inequality (7) is equivalent to showing

$$\Big\| \int h(\alpha, \alpha', w_1) p(w_1) d\mu(w_1) \Big\|_1 \not\equiv \mathbf{0}$$

as a function of α, α' for "almost all" measures μ—as there is no natural measure on the space of measures, we formalize this assuming a uniform distribution on w_1 below. Note that the outer norm computes the L_1 norm of a signed measure over o_1^o. For notational convenience, we will now define the concatenation $z = [\alpha, \alpha']$, with $z \in \mathbb{R}^2$. Note that we defined $W_1 \sim \mathcal{U}(a, b)$, for real parameters $a < b$. We can now concretely refine μ in the above statement, using our new z-notation, to showing that

$$g_a^b(z) := \Big\| \int_a^b h(z, w_1) dw_1 \Big\|_1 \not\equiv \mathbf{0} \tag{9}$$

as a function of z for almost every (a, b); i.e., the subset of (a, b) parameter space where (9) is violated is measure zero with respect to the standard Lebesgue measure in $\mathbb{R}^2$. Note that we drop the $p(w_1)$ factor since for the uniform distribution this is a constant which factors out.

Note that this can be analyzed by considering sections were we fix a and consider the set of b where (9) is violated; if this set has measure zero, then the overall set of cartesian pairs (a, b) which violate (9) has measure zero.

Correspondingly, fix any a, and consider a particular $\bar{b}$ where $g_a^{\bar{b}}(z) \equiv \mathbf{0}$ as a function over z. We expand the L_1 norm in (9) as

$$g_a^b(z) = \int \left| \frac{d}{d\lambda} \Big(\int_a^b h(z, w_1) dw_1 \Big) \right| d\lambda, \tag{10}$$

using the interventional absolute continuity assumption to invoke the Radon-Nikodym derivative on our signed measure over o_1^o with respect to the standard Lebesgue measure λ. Note that since $g_a^{\bar{b}}(z) \equiv \mathbf{0}$, we have that

$$\frac{d}{d\lambda} \Big(\int_a^b h(z, w_1) dw_1 \Big) = 0 \tag{11}$$

almost everywhere as a density function over $\mathbb{R}$. We now differentiate both sides of (10) with respect to b at $\bar{b}$. Due to the absolute value in (10), we must take care to differentiate from above and below and show both these cases are nonzero. As they follow similarly, we show the case for above:

$$\begin{aligned}
\frac{d}{db}\Big|_{\bar{b}^+} g_a^b(z) &= \frac{d}{db}\Big|_{\bar{b}^+} \int \left| \frac{d}{d\lambda} \Big(\int_a^b h(z, w_1) dw_1 \Big) \right| d\lambda && (12)\\
&= \int \frac{d}{db}\Big|_{\bar{b}^+} \left| \frac{d}{d\lambda} \Big(\int_a^b h(z, w_1) dw_1 \Big) \right| d\lambda && (13)\\
&= \int \left| \frac{d}{d\lambda} \Big(\frac{d}{db}\Big|_{\bar{b}^+} \int_a^b h(z, w_1) dw_1 \Big) \right| d\lambda && (14)\\
&= \int \left| \frac{d}{d\lambda} h(z, \bar{b}) \right| d\lambda && (15)\\
&= \Big\| h(z, \bar{b}) \Big\|_1 && (16)\\
&\not\equiv \mathbf{0}, \qquad \text{(as a function over } z\text{)} && (17)
\end{aligned}$$

where (13) follows from boundedness of the Radon-Nikodym derivative of $h(z, \bar{b})$, (14) follows from absolute value properties and (11), (15) follows from distributional continuity, and (17) follows from (8).

Proceeding similarly, we can show that both

$$\left.\frac{d}{db}\right|_{\bar{b}^+} g_a^b(z) \not\equiv \mathbf{0} \quad \text{and} \quad \left.\frac{d}{db}\right|_{\bar{b}^-} g_a^b(z) \not\equiv \mathbf{0}.$$

It is then immediate that there exists a ball $B(\bar{b}, \epsilon_{\bar{b}})$ such that $g_a^b(z) \not\equiv \mathbf{0}$ for all $b \in B(\bar{b}, \epsilon_{\bar{b}}) \setminus \bar{b}$. Applying Lemma 5 concludes that for a fixed a, the set of b for which (9) is violated is measure zero, and hence by Fubini for almost every uniform measure $\mathcal{U}(a, b)$ on w_1, we have that (7) holds for some α, α'. Therefore $S_1^{\mathfrak{s}} \not\perp\!\!\!\perp O_1^{\mathfrak{o}}$ in the interventional distribution on $S_1^{\mathfrak{s}}$.

A similar argument shows that $S_1^{\mathfrak{s}} \not\perp\!\!\!\perp A_{t'}^{\mathfrak{a}}$. By absolute continuity of the induced interventional distributions, we now have that Hoeffding's independence test is consistent, and hence the dependences are detected with probability 1 as $N \to \infty$. Therefore $O_1^{\mathfrak{o}} \dashrightarrow A_{t'}^{\mathfrak{a}}$, and $\widetilde{m}_{\mathfrak{o}}$ evaluates to false (6) as $N \to \infty$. ■

Theorem 2. *Let m denote the potential-cause test evaluated by Algorithm 1 on the distribution induced by the non-interventional system $\langle \mathcal{M}_s, \mathcal{G}_s \rangle$, and let $\widetilde{m}$ be the original test on the interventional system $\langle \widetilde{\mathcal{M}}_s, \widetilde{\mathcal{G}}_s \rangle$ where $\tilde{P}(s_1)$ has everywhere-nonzero density on $\mathcal{S}$. Then if $m_{\mathfrak{o}}$ correctly evaluates to true for a particular $\mathfrak{o} \in [d_{\mathcal{O}}]$, then $\widetilde{m}_{\mathfrak{o}}$ also evaluates to true almost surely as the number of trajectories $N \to \infty$.*

Proof: If $m_{\mathfrak{o}}$ evaluates to true, then for any $\mathfrak{s} \in [d_{\mathcal{S}}]$, $\mathfrak{a} \in [d_{\mathcal{A}}]$, and $t' \in [H]$, we have that either $S_1^{\mathfrak{s}} \perp\!\!\!\perp_{\mathcal{M}_s} O_1^{\mathfrak{o}}$ or $S_1^{\mathfrak{s}} \perp\!\!\!\perp_{\mathcal{M}_s} A_{\mathfrak{a}}^{t'}$, where $\perp\!\!\!\perp_{\mathcal{M}_s}$ denotes independence in the distribution induced by the non-interventional SCM $\mathcal{M}_s$. It suffices to show that both these independencies hold in the distribution induced by $\widetilde{\mathcal{M}}_s$. As both arguments follow similarly, we consider showing that $S_1^{\mathfrak{s}} \perp\!\!\!\perp_{\widetilde{\mathcal{M}}_s} O_1^{\mathfrak{o}}$.

As we are given $S_1^{\mathfrak{s}} \perp\!\!\!\perp_{\mathcal{M}_s} O_1^{\mathfrak{o}}$, it is immediate by faithfulness that there exists no collider-free path from $S_1^{\mathfrak{s}}$ to $O_1^{\mathfrak{o}}$ in $\mathcal{G}_s$. Since $\widetilde{\mathcal{G}}_s$ is simply $\mathcal{G}_s$ with the incoming edges to S_1 removed, it holds that there is no collider-free path between $S_1^{\mathfrak{s}}$ and $O_1^{\mathfrak{o}}$ in $\widetilde{\mathcal{G}}_s$. Therefore $S_1^{\mathfrak{s}} \perp\!\!\!\perp_{\widetilde{\mathcal{M}}_s} O_1^{\mathfrak{o}}$, and as $N \to \infty$ this is correctly detected with probability 1 by the consistency of Hoeffding's test. ■

Proposition 3. *Let $\widetilde{m}$ and m be as in Theorem 2, and consider a particular observation index $\mathfrak{o} \in [d_{\mathcal{O}}]$ such that the only incoming edge to $O_1^{\mathfrak{o}}$ is $W_1 \to O_1^{\mathfrak{o}}$. Then if in $\mathcal{G}_s$ there exists the fork $S_1^{\mathfrak{s}} \leftarrow W_1 \to O_1^{\mathfrak{o}}$ for some $\mathfrak{s} \in [d_{\mathcal{S}}]$ and a directed path from $S_1^{\mathfrak{s}}$ to some $A_t^{\mathfrak{a}}$, with $t \in [H], \mathfrak{a} \in [d_{\mathcal{A}}]$, $\widetilde{m}_{\mathfrak{o}}$ correctly masks the $\mathfrak{o}^{th}$ observation almost surely as the number of trajectories $N \to \infty$ while $m_{\mathfrak{o}}$ does not.*

Proof: We first show that m does not mask $\mathfrak{o}$ and take all causal and probabilistic statements to refer to the unintervened causal model $\langle \mathcal{M}_s, \mathcal{G}_s \rangle$. By faithfulness, the fork $S_1^{\mathfrak{s}} \leftarrow W_1 \to O_1^{\mathfrak{o}}$ in $\mathcal{G}_s$ produces a statistical dependence $S_1^{\mathfrak{s}} \not\perp\!\!\!\perp_{\mathcal{M}_s} O_1^{\mathfrak{o}}$ in the probability distribution induced by $\mathcal{M}_s$. Similarly, the directed path from $S_1^{\mathfrak{s}}$ to $A_t^{\mathfrak{a}}$ yields $S_1^{\mathfrak{s}} \not\perp\!\!\!\perp_{\mathcal{M}_s} A_t^{\mathfrak{a}}$. By consistency of Hoeffding's test, as $N \to \infty$ we get that ${}^{(1,t)}D_{\mathfrak{s},\mathfrak{a}}^{\mathfrak{o}}$ evaluates to true almost surely (3) and thus $O_1^{\mathfrak{o}} \dashrightarrow A_t^{\mathfrak{a}}$ by (4). Therefore $m_{\mathfrak{o}}$ is not masked (6).

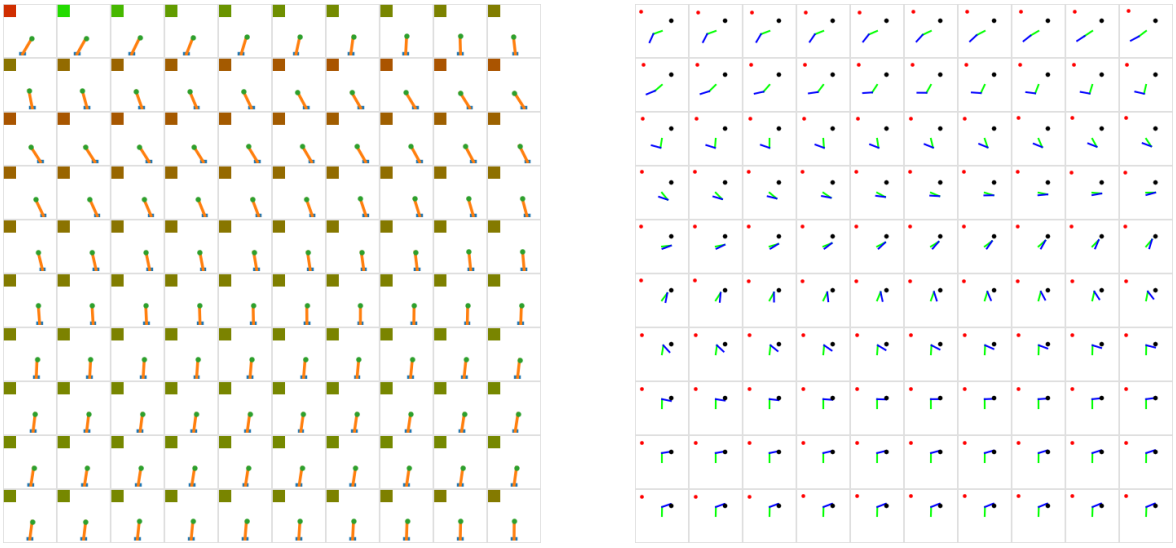
We now show that $\widetilde{m}$ does mask $\mathfrak{o}$ and take all causal and probabilistic statements to refer to the *intervened* causal model $\langle \widetilde{\mathcal{M}}_s, \widetilde{\mathcal{G}}_s \rangle$. Since W_1 only has outgoing edges, and the edge from $W_1 \to S_1^{\mathfrak{s}'}$ is removed in $\widetilde{\mathcal{G}}_s$ for every $\mathfrak{s}' \in [d_{\mathcal{S}}]$, there exists no path from $S_1^{\mathfrak{s}'}$ to $O_1^{\mathfrak{o}}$ in $\widetilde{\mathcal{G}}_s$, and therefore $S_1^{\mathfrak{s}'} \perp\!\!\!\perp_{\widetilde{\mathcal{M}}_s} O_1^{\mathfrak{o}}$ in the probability distribution induced by $\widetilde{\mathcal{M}}_s$. As $N \to \infty$ this independence is detected by Hoeffding's test, and since $\mathfrak{s}'$ was arbitrary ${}^{(1,t')}D_{\mathfrak{s}',\mathfrak{a}'}^{\mathfrak{o}}$ is false for every $\mathfrak{s}' \in [d_{\mathcal{S}}]$, $\mathfrak{a}' \in [d_{\mathcal{A}}]$, and $t' \in [H]$. Therefore $O_1^{\mathfrak{o}} \not\dashrightarrow A_{t'}^{\mathfrak{a}'}$ for any $\mathfrak{a}' \in [d_{\mathcal{A}}], t' \in [H]$, and (6) evaluates to true. Therefore $\widetilde{m}_{\mathfrak{o}}$ is masked. ■

II Experiments

We include here essential environment, architecture, and hyperparameter details.

A. Environments

We consider two environments: CartPole and Reacher. Both systems are rendered to 64×64 RGB images, pictured in Figure 4.



(a) CartPole (b) Reacher

Fig. 4: Illustrative rollouts of the expert policy on CartPole (a) and Reacher (b). The Reacher run is truncated to 100 time steps for visualization purposes. Note the nuisance feature in the upper left hand corner of the images.

1) CartPole: The CartPole environment [28] describes a nonlinear dynamic system consisting of four states: the cart position x, the cart velocity $\dot{x}$, the pole angle θ, and the pole angular velocity $\dot{\theta}$. The state vector at time t is $S_t := (x_t, \dot{x}_t, \theta_t, \dot{\theta}_t)$. The agent action is a continuous horizontal force acting on the cart, bounded symmetrically in the range $[-25, 25]$. The length of the pole is 1 meter, with the masses of the cart and the pole set to 1 and 0.1 kilograms, respectively. We specify the gravitational acceleration constant as $g = 9.8\text{m/s}^2$. The system is then discretized with $\Delta t = 0.05$ for 100 time steps using the forward Euler method, with the standard CartPole dynamics equations adapted from OpenAI Gym [27]. We minimize cumulative stepwise quadratic form loss to the upright target state $S_{\text{target}} = (0, 0, 0, 0)$, with an additional quadratic control cost.

To each frame, we add a 15×15 square nuisance feature at the top-left corner of each image. The color of the square interpolates linearly between green and red, depending on the action (cart force) from the previous time step. At the initial time step $t = 1$ there is no previous action, and thus we use a random number drawn from $\text{Unif}(-25, 25)$ to generate the square.

We generate $5,000$ random-policy trajectories for training the β-VAE and $1,000$ expert trajectories for imitation learning. Random-policy trajectories are terminated when the states become out-of-bound, and are thus generally significantly shorter than the expert trajectories.

2) Reacher: Reacher is also implemented based on the classic OpenAI Gym environment [27]. The system contains six states: target position x^*, y^*; joint one angle and velocity $\theta_1, \dot{\theta}_1$; and joint two angle and velocity $\theta_2, \dot{\theta}_2$. The target positions x^*, y^* is fixed over the course of one trajectory to a random point in the reachable area. Both links have mass 1 kilogram and length 0.5 meters. Agents specify torques at both joints, bounded in the range $[-2, 2]$. The objective penalizes squared distance of the end effector from (x^*, y^*)—visualized as a black dot—at each time step, along with a quadratic control cost. We simulate with a time step $\Delta t = 0.05$ seconds for 200 time steps.

For Reacher, we demonstrate that our method can eliminate a visually different type of confounding than the colored square in the CartPole experiment. The considered confounder is a red dot in the upper-left corner of the image that moves translationally according to the agent action in the previous time step. Specifically, the horizontal position is linearly interpolated according to the first joint torque, and the vertical position is linearly interpolated according to the second joint torque. Similarly to CartPole, we choose a random previous action for the first time step.

We generate $2,000$ random-policy trajectories for training the β-VAE and $2,000$ expert trajectories for imitation learning.

B. VAE training

We train a standard β-VAE [29] as implemented by [30]. We choose a latent space dimension of 3 for CartPole and 12 for Reacher, although we note that larger choices for the latent space dimension yield similar results. We train for 150 epochs at a learning rate of 0.0005 on Reacher and 100 epochs at a learning rate of 0.005 on CartPole. Both use an exponential learning rate scheduler with decay factor 0.95. Our batch size is 256 for Reacher and 64 for CartPole. Finally, we choose the disentaglement factor $\beta = 100$ for CartPole and $\beta = 1000$ for Reacher.

C. Behavior cloning training

For Reacher, we use a standard pre-activation ResNet-18 [31]; for the simpler CartPole environment, a simple ConvNet with 3×3 convolutions and channel sizes $[32, 64, 128, 256, 512]$ suffices (this architecture resembles the VAE encoder). Both architectures are trained with the Adam optimizer [32] at an initial learning rate of 0.001 and exponential learning rate decay with factor 0.96. We use a batch size of 256 and evaluate the performance of the agent with 25 validation rollouts every 10 epochs. As a single image of the environment cannot convey higher-order state information such as velocity, we input the previous two images into our policies—i.e., $L = 2$ in (1). Thus, we make the necessary architectural change to the underlying models of setting the number of input channels to 6. Since the first time step does not have an associated previous image, we use a blank image as a surrogate.