A Proofs

In this section, we provide proofs for the theoretical claims delineated in the paper. Throughout this paper, it is important to note that detailed parametrizations of the underlying SCM are not known to the agent. Instead, the agent has access to the expert's demonstrations, which are summarized as the observational distribution $P(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{Y})$.

We begin by revisiting the distribution of state visitation. Specifically, $\rho_\pi(s)$ can be calculated by:

$$\rho_\pi(s) = P(s) + \gamma \sum_{s', x} \mathcal{T}(s', x, s) \pi(x \mid s') \rho_\pi(s') \tag{23}$$

where $P(s)$ represents the initial state distribution, γ represents the discount factor, $\mathcal{T}$ represents the transition probabilities for the imitator. Subsequently, we are able to develop the occupancy measure for the policy π:

$$\rho_\pi(s, x) = \rho_\pi(s) \pi(x \mid s) \tag{24}$$

It is important to note that, although the format of the occupancy measure $\rho_\pi(s, x)$ shares a formal resemblance to the one presented in GAIL [19], $\rho_\pi(s, x)$ specifically represents an interventional distribution with policy $\text{do}(\pi)$. The identifiability of the transition $\mathcal{T}(s, x, s')$ directly impacts the identifiability of $P_\pi(s_t)$. If $P_\pi(s_t)$ is not identifiable, $\rho_\pi(s)$ and $\rho_\pi(s, x)$ are consequently not identifiable.

Theorem 1. *Given any positive observational distribution $P(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{Y}) > 0$, there exists an MDP model $\hat{M}$ compatible with the causal graph of Fig. 2b such that $P(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{Y}; \hat{M}) = P(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{Y})$ and for any policy π, any time step $t = 1, 2, \ldots$, any state $s \in \mathcal{S}$,*

$$V_\pi\left(s; \hat{M}\right) < \mathbb{E}\left[R_t \mid S_t = s; \hat{M}\right]. \tag{11}$$

Proof. Without loss of generality, the reward Y is normalized so that it has a range of $[0, 1]$ Based on the value function defined in Eq. (8), we first show how to expand it into a recursive version:

$$V_\pi(s_t) = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k Y_{t+k} \mid s_t\right] \tag{25}$$

$$= \mathbb{E}_\pi[Y_t \mid s_t] + \mathbb{E}_\pi\left[\sum_{k=1}^{\infty} \gamma^k Y_{t+k} \mid s_t\right] \tag{26}$$

$$= \mathbb{E}_\pi[Y_t \mid s_t] + \gamma \mathbb{E}_\pi\left[\sum_{k=1}^{\infty} \gamma^{k-1} Y_{t+k} \mid s_t\right] \tag{27}$$

$$= \mathbb{E}_\pi[Y_t \mid s_t] + \gamma \sum_{s_{t+1}} P_\pi(s_{t+1} \mid s_t) \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k Y_{t+1+k} \mid s_t, s_{t+1}\right] \tag{28}$$

$$= \mathbb{E}_\pi[Y_t \mid s_t] + \gamma \sum_{s_{t+1}} P_\pi(s_{t+1} \mid s_t) V_\pi(s_{t+1}), \tag{29}$$

where γ is the discount factor, $P_\pi(s_{t+1} \mid s_t)$ denotes the transition probability when executing policy π.

From the second last line to the last line is justified by the experimental markovian property, as discussed in Sec. 2, following the graph Fig. 2a. More details could be found in [54]. $\mathbb{E}_\pi[Y_t \mid s_t] = \mathbb{E}[Y_t \mid s_t, \text{do}(\pi)]$ denotes the expected reward received by the agent when executing policy π. Similarly, the transition probability

$$P_\pi(s_{t+1} \mid s_t) = \sum_{x_t} P_{x_t}(s_{t+1} \mid s_t) \pi(x_t \mid s_t), \tag{30}$$

and $P_{x_t}(s_{t+1} \mid s_t) = P(s_{t+1} \mid s_t, \text{do}(x_t)) = \mathcal{T}(s_t, x_t, s_{t+1})$. Generally speaking, when any unobserved confounder exists between S_{t+1} and X_t, the causal query $P_{x_t}(s_{t+1} \mid s_t)$ is not identifiable

14

[32, 43, 5, 51]. Building on the previous derivations, we arrive at the recursive formulation of the value function under policy π:

$$V_\pi(s_t) = \sum_{x_t} \mathcal{R}(s_t, x_t)\, \pi(x_t \mid s_t) + \gamma \sum_{s_{t+1}} \mathcal{T}(s_t, x_t, s_{t+1})\pi(x_t \mid s_t) V_\pi(s_{t+1}) \tag{31}$$

$$= \sum_{x_t} \pi(x_t \mid s_t) \left(\mathcal{R}(s_t, x_t) + \gamma \sum_{s_{t+1}} \mathcal{T}(s_t, x_t, s_{t+1}) V_\pi(s_{t+1}) \right). \tag{32}$$

Next, to establish the validity of the preceding claim, we proceed by applying the technique of mathematical induction. Let $|S|$ denote the number of distinct states for S.

Base case $t = T$. For the final timestep T, for each state index j where $\forall j,\ 1 \leq j \leq |S|$, the value function $V_\pi(s_{(T,j)})$ cane be defined as follows:

$$V_\pi(s_{(T,j)}) = \mathbb{E}_\pi\left[Y_T \mid S_T = s_{(T,j)} \right]$$
$$= \sum_{x_t} \mathbb{E}_{x_t}\left[Y_T \mid S_T = s_{(T,j)} \right] \pi(x_t \mid s_{(T,j)}) \tag{33}$$

where $s_{(T,j)}$ refers to the scenario where the state at the final timestep S_T is equal the specific state j.

In order to obtain the worst-case SCM $\hat{M}$, we need to minimize $V_\pi(s_T) - V(s_T)$ compatible with the observational distribution, by establishing its lower bound. To this end, we directly employ the natural bound [29], which has been discussed in Sec. 2.1:

$$\min_M \quad V_\pi(s_{(T,j)}; M) - V(s_{(T,j)}; M)$$
$$= \sum_{x_t} \mathbb{E}_{x_t}\left[Y_T \mid S_T = s_{(T,j)}; M \right] \pi(x_t \mid s_{(T,j)}) - V(s_{(T,j)}; M)$$
$$= \sum_{x_t} \mathbb{E}\left[Y_T \mid s_{(T,j)}, x_t \right] P(x_t \mid s_{(T,j)})\pi(X_T = x_t \mid s_{(T,j)}) - \sum_{x_t} \mathbb{E}\left[Y_T \mid s_{(T,j)}, x_t \right] P(x_t \mid s_{(T,j)})$$
$$< 0$$
$$\tag{34}$$

The last step is justified because $P(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{Y}) > 0$ and $0 \leq \pi(X_T = x_t \mid s_{(T,j)}) \leq 1$. Intuitive examples illustrating this conclusion are provided in Sec. 2.1 and Appendix F. Therefore, this confirms the validity of the inequality for the base case.

Specifically, in certain degenerate cases where there is only one possible action, the imitator has no choice but to follow that single option. Consequently, unobserved confounders are less likely to introduce significant effects in these scenarios. However, under such conditions, pursuing imitation learning is not meaningful, as there is no variability in choice for the imitator to learn from. Therefore, such cases are of limited relevance to the scope of this analysis.

Induction case. Suppose at $t + 1$, $V_\pi(s_{t+1}) < V(s_{t+1})$, we need to prove $V_\pi(s_t) < V(s_t)$.

$$V_\pi(s_t) = \mathbb{E}_\pi[Y_t \mid s_t] + \gamma \sum_j P_\pi(s_{(t+1,j)} \mid s_t) \underbrace{V_\pi(s_{(t+1,j)})}_{<V(s_{(t+1,j)})} \tag{35}$$

Without loss of generality, we assume that the state with the minimal value at $t + 1$ is denoted as $s_{(t+1,|S|)}$. Our approach is founded on the premise that in obtaining the worst-case SCM $\hat{M}$, it is strategic to allocate the lowest possible transition probabilities to the state with the highest value, while preferentially assigning higher probabilities to states demonstrating smaller values. Specifically, one starts with the estimate $P(S_{t+1} = s_{(t+1,1)}, x_t \mid S_t)$ for $P_{x_t}(S_{t+1} = s_{(t+1,1)} \mid S_t)$. Following this logic, we systematically allocate probability masses for indices $1 \leq j \leq |S| - 1$ as follows:

$$P_{x_t}(S_{t+1} = s_{(t+1,j)} \mid S_t) \leftarrow P(S_{t+1} = s_{(t+1,j)}, x_t \mid S_t)$$

15

In accordance with the established properties of probability distributions, it follows that:

$$\sum_{j=1}^{|S|} P_{x_t}(S_{t+1} = s_{(t+1,j)} \mid S_t) = 1.$$

Considering the state $s_{(t+1,|S|)}$, the corresponding probability can be assigned as:

$$P_{x_t}(S_{t+1} = s_{(t+1,|S|)} \mid S_t) = 1 - \sum_{j=1}^{|S|-1} P_{x_t}(S_{t+1} = s_{(t+1,j)} \mid S_t).$$

By substituting the assigned values, we are able to derive the following expression:

$$P_{x_t}(S_{t+1} = s_{(t+1,|S|)} \mid S_t) \leftarrow 1 - P(s_{(t+1,1)}, x_t \mid S_t) - P(s_{(t+1,2)}, x_t \mid S_t) \cdots - P(s_{(t+1,|S|-1)}, x_t \mid S_t),$$

where the right-hand side simplifies to:

$$\left(P(s_{(t+1,1)}, x_t \mid S_t) + P(s_{(t+1,2)}, x_t \mid S_t) \cdots + P(s_{(t+1,|S|-1)}, x_t \mid S_t)\right) = \left(\sum_{j=1}^{|S|-1} P\left(s_{(t+1,j)}, x_t \mid S_t\right)\right)$$

$$= P(x_t \mid S_t) - P(s_{(t+1,|S|)}, x_t \mid S_t).$$

It is established that the expression $0 \leq 1 - P(x_t \mid S_t) + P(s_{(t+1,|S|)}, x_t \mid S_t) \leq 1$ holds true. This inequality is supported by the following equation:

$$\sum_{j=1}^{|S|} P(s_{(t+1,j)}, x_t \mid S_t) = P(x_t \mid S_t).$$

To further analyze the expert policy, the associated value function $V(s_t)$ can be expanded as follows:

$$V(s_t) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k Y_{t+k} \mid S_t = s_t\right]$$

$$= \mathbb{E}[Y_t \mid s_t] + \gamma \sum_j P(s_{(t+1,j)} \mid s_t) V(s_{(t+1,j)}), \tag{36}$$

where γ is the discount factor, $P(s_{(t+1,j)} \mid s_t)$ denotes the observational transition probability. Notably, $P(s_{(t+1,j)} \mid s_t)$ and $P_\pi(s_{(t+1,j)} \mid s_t)$ are generally different, because they reflect two distinct probabilities: $P(s_{(t+1,j)} \mid s_t)$ corresponding to the observational distribution and the other, $P_\pi(s_{(t+1,j)} \mid s_t)$, representing the imitator's transition dynamics.

In accordance with the established properties of probability distributions, it follows that:

$$\sum_{j=1}^{|S|} P(S_{t+1} = s_{(t+1,j)} \mid S_t) = 1.$$

Without loss of generality, suppose the policy is a deterministic policy. Actually, the following proof holds true regardless of the choice of x_t. Subsequently, we analyze the gap between $V_\pi(s_t)$ and $V(s_t)$ as follows:

$$V_\pi(s_t) - V(s_t)$$

$$= \left(\mathbb{E}_\pi[Y_t \mid s_t] + \gamma \sum_{j=1}^{|S|} P_\pi(s_{(t+1,j)} \mid s_t) V_\pi(s_{(t+1,j)})\right)$$

$$- \left(\mathbb{E}[Y_t \mid s_t] + \gamma \sum_{j=1}^{|S|} P(s_{(t+1,j)} \mid s_t) V(s_{(t+1,j)})\right)$$

$$= \left(\mathbb{E}_\pi[Y_t \mid s_t] + \gamma \sum_{j=1}^{|S|} \sum_{x_t} P_{x_t}(s_{(t+1,j)} \mid s_t) \pi(x_t \mid s_t) V_\pi(s_{(t+1,j)})\right) \tag{37}$$

$$- \left(\mathbb{E}[Y_t \mid s_t] + \gamma \sum_{j=1}^{|S|} P(s_{(t+1,j)} \mid s_t) V(s_{(t+1,j)})\right)$$

16

$$\min_{M} \quad V_\pi(s_t; M) - V(s_t; M)$$

$$= \mathbb{E}_\pi[Y_t \mid s_t; M] - \mathbb{E}[Y_t \mid s_t; M] + \gamma \sum_{j=1}^{|S|-1} \left(P(s_{(t+1,j)}, x_t \mid s_t) - P(s_{(t+1,j)} \mid s_t)\right) V(s_{(t+1,j)})$$

$$+ \gamma \left(1 - \left(\sum_{j=1}^{|S|-1} P\left(s_{(t+1,j)}, x_t \mid s_t\right)\right) - P(s_{(t+1,|S|)} \mid s_t)\right) V(s_{(t+1,|S|)})$$

$$= \underbrace{\mathbb{E}_\pi[Y_t \mid s_t; M] - \mathbb{E}[Y_t \mid s_t; M]}_{<0}$$

$$+ \gamma \sum_{j=1}^{|S|-1} \underbrace{\left(P(s_{(t+1,j)}, x_t \mid s_t) - P(s_{(t+1,j)} \mid s_t)\right)}_{<0} \underbrace{\left(V(s_{(t+1,j)}) - V(s_{(t+1,|S|)})\right)}_{>0}$$

$$< 0$$

$$(38)$$

where $\min_M \mathbb{E}_\pi[Y_t \mid s_t; M] - \mathbb{E}[Y_t \mid s_t; M] < 0$ follows a similar logic as previously introduced in the base case, and $V(s_{(t+1,j)}) - V(s_{(t+1,|S|)}) > 0$ is consistent with the ordering assumption, where the state $s_{(t+1,|S|)}$ represents the minimal value.

In some degenerated cases when $\mathbb{E}_\pi[Y_t \mid s_t] = 0$ and $\mathbb{E}[Y_t \mid s_t] = 0$, it might coincidentally follow that $V_\pi(s_t) = 0$, which is equal to $V(s_t) = 0$. Another instance of degeneracy occurs when the value function $V(s_{(t+1,j)})$ remains the same across all states. Such occurrences are extremely unlikely in practical scenarios, especially when $P(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{Y}) > 0$. $\qquad\square$

B Derivations for Causal GAIL Algorithms

Theorem 2. *Given an MDP M compatible with the causal graph of Fig. 3a, let $\mathscr{R}$ be a parametric family containing the conditional reward $\mathbb{E}[Y_t \mid s_t, x_t]$. Consider the following optimization program,*

$$\nu^* = \min_\pi \max_{\widetilde{\mathcal{R}} \in \mathscr{R}} \sum_{s,x} \widetilde{\mathcal{R}}(s, x) P(x \mid s) \left(\rho(s) - \pi(x \mid s)\rho_{\boldsymbol{\pi}}(s)\right) \qquad (14)$$

When the gap $\nu^* \leq 0$, the solution π^* is an imitating policy satisfying $\mathbb{E}_{\pi^*}[R_1] \geq \mathbb{E}[R_1]$.

Proof. Based on Eq. (13), we have:

$$\mathbb{E}_\pi[R_1] = \sum_{s,x} \mathcal{R}(s, x)\, \pi(x \mid s)\rho_\pi(s)$$

It follows from Eq. (10) that parametrization of $\mathcal{R}(s, x)$ can be bound from the observational distribution. The imitator's expected return could thus be lower bounded as

$$\mathbb{E}_\pi[R_1] \geq \sum_{s,x} \widetilde{\mathcal{R}}(s, x)\, P(x \mid s)\pi(x \mid s)\rho_\pi(s)$$

Note that the expert's expected return could be similarly decomposed as

$$\mathbb{E}[R_1] = \sum_{x,s} \widetilde{\mathcal{R}}(s, x) P(x \mid s)\rho(s),$$

where $\rho(s) = \sum_{t=0}^\infty \gamma^t P(S_t = s)$ is the expert's occupancy measure.

$$\nu^* = \min_\pi \max_M \quad \mathbb{E}[R_1; M] - \mathbb{E}_\pi[R_1; M] \qquad (39)$$

$$= \min_\pi \max_{\widetilde{\mathcal{R}}, \mathcal{R}} \quad \sum_{x,s} \widetilde{\mathcal{R}}(s, x) P(x \mid s)\rho(s) - \sum_{s,x} \mathcal{R}(s, x)\, \pi(x \mid s)\rho_\pi(s) \qquad (40)$$

$$= \min_\pi \max_{\widetilde{\mathcal{R}}} \quad \sum_{s,x} \widetilde{\mathcal{R}}(s, x) P(x \mid s) \left(\rho(s) - \pi(x \mid s)\rho_{\boldsymbol{\pi}}(s)\right), \qquad (41)$$

which is the ultimate target expression. $\qquad\square$

17

Next, we will show the derivation details for matching weighted occupancy measures between the imitator and the expert. Suppose $\psi^* = \max_{\mathcal{R}} a^\top \mathcal{R} - \psi(\mathcal{R})$ is a conjugate function of ψ. Following a similar logic in [19], we utilize a similar cost regularizer ψ_{GA}, leading to the formulation of Alg. 1. Basically, Alg. 1 minimizes Jensen-Shannon divergence between $P(x \mid s)\rho(s)$ and $P(x \mid s)\pi(x \mid s)\rho_\pi(s)$.

First, we reformulate the equation into state-action occupancy measures:

$$\psi^* \left(P(x \mid s)\rho(s) - P(x \mid s)\pi(x \mid s)\rho_\pi(s) \right) = \psi^* \left(\rho(s,x) - P(x \mid s)\rho_\pi(s,x) \right) \tag{42}$$

Based on the definition of ψ^*, we have:

$$\psi^* \left(\rho(s,x) - P(x \mid s)\rho_\pi(s,x) \right) \tag{43}$$

$$= \max_{\mathcal{R}} \sum_{s,x} \left(\rho(s,x) - P(x \mid s)\rho_\pi(s,x) \right) \mathcal{R}(s,x) - \sum_{s,x} P(x \mid s)\rho_\pi(s,x)g_\phi(\mathcal{R}(s,x)) \tag{44}$$

$$= \sum_{s,x} \max_{\mathcal{R}} \rho(s,x)\mathcal{R} - P(x \mid s)\rho_\pi(s,x)\phi \left(-\phi^{-1}(-\mathcal{R}) \right) \tag{45}$$

$$= \sum_{s,x} \max_{\mathcal{R}'} \rho(s,x)(-\phi(\mathcal{R}')) - P(x \mid s)\rho_\pi(s,x)\phi \left(-\phi^{-1}(\phi(\mathcal{R}')) \right) \tag{46}$$

$$= \sum_{s,x} \max_{\mathcal{R}'} \rho(s,x)(-\phi(\mathcal{R}')) - P(x \mid s)\rho_\pi(s,x)\phi \left(-\mathcal{R}' \right) \tag{47}$$

where we make the change of variables $\mathcal{R} \to -\phi(\mathcal{R}')$. Suppose $D \in \mathcal{S} \times \mathcal{X} \mapsto (0,1)$ is a discriminator classifier (e.g, a neural network). Using the logistic loss $\phi(x) = \log \left(1 + e^{-x} \right)$, we can get:

$$\psi^* \left(\rho(s,x) - P(x \mid s)\rho_\pi(s,x) \right) \tag{48}$$

$$= \sum_{s,x} \max_{\mathcal{R}'} \rho(s,x) \log \left(\frac{1}{1 + e^{-\mathcal{R}'}} \right) + P(x \mid s)\rho_\pi(s,x) \log \left(1 - \frac{1}{1 + e^{-\mathcal{R}'}} \right) \tag{49}$$

$$= \max_{D \in (0,1)^{\mathcal{S} \times \mathcal{X}}} \mathbb{E}[\log(D(S,X))] + \mathbb{E}_\pi \left[P(x \mid s) \log(1 - D(S,X)) \right], \tag{50}$$

which is the ultimate target expression.

Theorem 3. *Given an MDP M compatible with the causal graph of Fig. 3b, let $\mathscr{R}$ be a parametric family containing the conditional reward $\mathbb{E}[Y_t \mid s_t, x_t]$, and $\mathscr{T}$ be a parametric family over conditional probabilities $P\left(s_{t+1} \mid s_t, x_t \right)$ defined in Eq. (20). Consider the following program,*

$$\nu^* = \min_\pi \max_{\mathcal{R} \in \mathscr{R}} \max_{\mathcal{T} \in \mathscr{T}} \sum_{s,x} \mathcal{R}(s,x) \left(P(x \mid s)\rho(s) - \pi(x \mid s)\rho_{\boldsymbol{\pi}}\left(s; \mathcal{T} \right) \right) \tag{21}$$

When the gap $\nu^* \leq 0$, the solution π^* is an imitating policy satisfying $\mathbb{E}_{\pi^*}\left[R_1 \right] \geq \mathbb{E}[R_1]$.

Proof. Based on Eq. (13), we have:

$$\mathbb{E}_\pi \left[R_1 \right] = \sum_{s,x} \mathcal{R}\left(s,x \right) \pi(x \mid s)\rho_\pi(s)$$

$$= \sum_{s,x} \mathcal{R}\left(s,x \right) \underbrace{\rho_\pi(s,x)}_{\text{Non-ID}}$$

The reward function $\mathcal{R}$ is identifiable and must be contained in the parametric space of the expert's nominal reward $\mathscr{R}$. In other words,

$$\mathcal{R}\left(s,x \right) = \widetilde{\mathcal{R}}\left(s,x \right). \tag{51}$$

18

The transition distribution $\mathcal{T}$ can be bounded from the demonstration data using Eq. (9). Therefore, we get:

$$\nu^* = \min_{\pi} \max_{M} \quad \mathbb{E}[R_1; M] - \mathbb{E}_\pi[R_1; M] \tag{52}$$

$$= \min_{\pi} \max_{\mathcal{T}, \widetilde{\mathcal{R}}, \mathcal{R}} \quad \sum_{s,x} \widetilde{\mathcal{R}}(s,x) P(x \mid s) \rho(s) - \mathcal{R}(s,x) \rho_\pi(s,x; \mathcal{T}) \tag{53}$$

$$= \min_{\pi} \max_{\mathcal{T}, \mathcal{R}} \quad \sum_{s,x} \mathcal{R}(s,x) \left(P(x \mid s) \rho(s) - \rho_\pi(s,x; \mathcal{T}) \right) \tag{54}$$

$$= \min_{\pi} \max_{\mathcal{T} \in \mathscr{T}, \mathcal{R} \in \mathscr{R}} \quad \sum_{s,x} \mathcal{R}(s,x) \left(P(x \mid s) \rho(s) - \pi(x \mid s) \rho_{\boldsymbol{\pi}}(s; \mathcal{T}) \right) \tag{55}$$

which is the ultimate desired expression. $\qquad\square$

Consider again the expected return decomposition in Eq. (13). The reward function $\mathcal{R}$ is identifiable and must be contained in the parametric space of the expert's nominal reward $\mathscr{R}$. The transition distribution $\mathcal{T}$ can be bounded from the demonstration data using Eq. (9). One could thus obtain a lower bound over the imitator's performance by reasoning about the worst-case occupancy measure compatible with demonstrations. Formally, with the fixed reward function $\mathcal{R}$ and the fixed policy π, the imitator's return is bounded by

$$\mathbb{E}_\pi[R_1] \geq \min_{\mathcal{T}, \rho_\pi} \quad \sum_{s,x} \mathcal{R}(s,x) \pi(x \mid s) \rho_\pi(s) \tag{56}$$

$$\text{subject to:} \quad \rho_\pi(s) \geq 0, \quad \text{and} \quad \sum_s \rho_\pi(s) = \frac{1}{1-\gamma}$$

$$\rho_\pi(s) = P(s) + \gamma \sum_{s',x} \mathcal{T}(s',x,s) \pi(x \mid s') \rho_\pi(s')$$

$$\text{Obs. Constraints } \mathscr{T}: \quad \begin{cases} \sum_s \mathcal{T}(s',x,s) = 1, \quad \text{and } \mathcal{T}(s,x,s') \geq \widetilde{\mathcal{T}}(s,x,s') P(x \mid s) \\ \mathcal{T}(s,x,s') \leq \widetilde{\mathcal{T}}(s,x,s') P(x \mid s) + P(\neg x \mid s) \end{cases} \tag{57}$$

The above optimization problem is similar to the classic linear program for planning in MDPs [36]. The main difference is that the transition distribution $\mathcal{T}$ is no longer fixed but bounded in a convex space $\mathscr{T}$ specified from the observational data. Similar to the previous setting, we could solve an imitating policy by minimizing the performance gap between the imitator and the expert in the worst-case environment compatible with the observational data and prior knowledge.

Next, we will provide a heuristic algorithm to solve the optimization program presented in Eq. (19) and Eq. (20). Specifically, as discussed in Eq. (9), we are able to bound the transition distribution $\mathcal{T}$ by:

$$\mathcal{T}(s,x,s') \in \left[\widetilde{\mathcal{T}}(s,x,s') P(x \mid s), \widetilde{\mathcal{T}}(s,x,s') P(x \mid s) + P(\neg x \mid s) \right]. \tag{58}$$

The intuition for Alg. 3 is: in order to find the worst case, we need to put as less transition probability mass as possible to the state with maximal values, and allocate higher transition probabilities to states with smaller values. Without loss of generality, suppose $V_{x_t}(s_{(t+1,|S|)})$ is found to have the smallest relative value. For all other states $j \neq |S|$, we need to allocate as less transition probability mass as possible. Therefore, we take the lower bound:

$$P_{x_t}(S_{t+1} = s_{(t+1,j)} \mid s_t) := P(S_{t+1} = s_{(t+1,j)}, x_t \mid s_t) \tag{59}$$

$$:= P(S_{t+1} = s_{(t+1,j)} \mid s_t, x_t) P(x_t \mid s_t), \tag{60}$$

where $P_{x_t}(s_{t+1} \mid s_t) = P(s_{t+1} \mid s_t, \text{do}(x_t)) = \mathcal{T}(s_t, x_t, s_{t+1})$, and $P(s_{(t+1)} \mid s_t, x_t) = \widetilde{\mathcal{T}}(s_t, x_t, s_{t+1})$. For the state $s_{(t+1,|S|)}$, we have:

$$P_{x_t}(S_{t+1} = s_{(t+1,|S|)} \mid s_t) := 1 - \left(\sum_{j=1}^{|S|-1} P(S_{t+1} = s_{(t+1,j)}, x_t \mid s_t) \right). \tag{61}$$

19

Algorithm 3: Find Worst-Case Discounted Future Reward

1: **Input**: $P(s_{t+1}, x_t | s_t)$, the value function $V_{x_t}(s_t)$
2: **Output**: Probability mass assignments for non-ID transitions $\mathcal{T}$
3: Let $V_{x_t}(s_{(t+1,|S|)})$ is determined to have the minimal relative value
4: Set

$$P_{x_t}(S_{t+1} = s_{(t+1,j)} \mid s_t) := P(S_{t+1} = s_{(t+1,j)}, x_t \mid s_t), \text{where } j \neq |S|$$

$$P_{x_t}(S_{t+1} = s_{(t+1,|S|)} \mid s_t) := 1 - \left(\sum_{j=1}^{|S|-1} P(S_{t+1} = s_{(t+1,j)}, x_t \mid s_t) \right)$$

5: **return**

Following a similar logic in Alg. 1: we reformulate the equation into state-action occupancy measures:

$$\psi^* \left(P(x \mid s)\rho(s) - \pi(x \mid s)\rho_\pi(s; \mathcal{T}) \right) = \psi^* \left(\rho(s,x) - \rho_\pi(s,x; \mathcal{T}) \right) \tag{62}$$

Based on the definition of ψ^*, we have:

$$\psi^* \left(\rho(s,x) - \rho_\pi(s,x; \mathcal{T}) \right) \tag{63}$$

$$= \max_{\mathcal{T},\mathcal{R}} \sum_{s,x} \left(\rho(s,x) - \rho_\pi(s,x; \mathcal{T}) \right) \mathcal{R}(s,x) - \sum_{s,x} \rho_\pi(s,x; \mathcal{T}) g_\phi(\mathcal{R}(s,x)) \tag{64}$$

$$= \sum_{s,x} \max_{\mathcal{T},\mathcal{R}} \rho(s,x)\mathcal{R} - \rho_\pi(s,x; \mathcal{T})\phi \left(-\phi^{-1}(-\mathcal{R}) \right) \tag{65}$$

$$= \sum_{s,x} \max_{\mathcal{T},\mathcal{R}'} \rho(s,x)(-\phi(\mathcal{R}')) - \rho_\pi(s,x; \mathcal{T})\phi \left(-\phi^{-1}(\phi(\mathcal{R}')) \right) \tag{66}$$

$$= \sum_{s,x} \max_{\mathcal{T},\mathcal{R}'} \rho(s,x)(-\phi(\mathcal{R}')) - \rho_\pi(s,x; \mathcal{T})\phi \left(-\mathcal{R}' \right) \tag{67}$$

Suppose $D \in \mathcal{S} \times \mathcal{X} \mapsto (0,1)$ is a discriminator classifier (e.g, a neural network). Using the logistic loss $\phi(x) = \log \left(1 + e^{-x} \right)$, we can get:

$$\psi^* \left(\rho(s,x) - \rho_\pi(s,x; \mathcal{T}) \right) \tag{68}$$

$$= \sum_{s,x} \max_{\mathcal{T},\mathcal{R}'} \rho(s,x) \log \left(\frac{1}{1 + e^{-\mathcal{R}'}} \right) + \rho_\pi(s,x; \mathcal{T}) \log \left(1 - \frac{1}{1 + e^{-\mathcal{R}'}} \right) \tag{69}$$

$$= \max_{\mathcal{T},D} \mathbb{E}[\log(D(S,X))] + \mathbb{E}_\pi \left[\log(1 - D(S,X)); \mathcal{T} \right]. \tag{70}$$

Therefore, we are able to obtain the ultimate target expression:

$$\nu^* = \min_\pi \max_{\mathcal{T} \in \mathscr{T}, D \in (0,1)^{\mathcal{S} \times \mathcal{X}}} \mathbb{E}[\log(D(S,X))] + \mathbb{E}_\pi \left[\log(1 - D(S,X)); \mathcal{T} \right]. \tag{71}$$

C Finding the Worst-Case Transition Distribution

In this section, we provide a practical algorithm, Alg. 3, designed to solve the optimization problem formulated in Eq. (19) and Eq. (20). The underlying rationale of Alg. 3 is to search for the worst-case scenario by allocating the minimal transition probability mass to the state with the highest value while assigning greater transition probabilities to states with lower values. The resulting solution should still adhere to a set of predefined observational constraints to ensure feasibility. This approach ensures that the most "adversarial" outcome is prioritized during the optimization process.

To further clarify the approach above, consider the following numerical example. Suppose there are only two states. The value function $V_{x_t}(s_{t+1})$ takes on two values: $V_{x_t}(s_{(t+1,1)}) = 0.8$ and $V_{x_t}(s_{(t+1,2)}) = 0.2$. Because $V_{x_t}(s_{(t+1,1)}) > V_{x_t}(s_{(t+1,2)})$, the algorithm seeks the worst-case discounted future reward by allocating $P_{x_t}(s_{(t+1,1)} \mid s_t) \leftarrow P(s_{(t+1,1)}, x_t \mid s_t)$ and $P_{x_t}(s_{(t+1,2)} \mid s_t) \leftarrow 1 - P(s_{(t+1,1)}, x_t \mid s_t)^3$. As such, we are able to collect trajectories from the imitator, even though $P_\pi(s_{t+1} \mid s_t)$ is not identifiable.

---

[3]In this case, although $P_{x_t}(s_{(t+1,2)} \mid s_t)$ has a lower bound of $P(s_{(t+1,2)}, x_t \mid s_t)$, it cannot be set exactly equal to this value. It is crucial to maintain the condition $P_{x_t}(s_{(t+1,1)} \mid s_t) + P_{x_t}(s_{(t+1,2)} \mid s_t) = 1$.

20

D   More Details for the Experiments

All experiments were conducted using Intel Cascade Lake processors, with 30 vCPUs and 120 GB memory on a system running Ubuntu 18.04. Upon acceptance of this manuscript, we intend to make the source code available in the camera-ready version of the paper.

MDP$_{\text{obs}}$ Previously, 1000 random discrete causal models are sampled and all the performance gaps are less than 0. In other words, when both the reward and the transition are confounded, all imitators fail to match expert performance.

Specifically, let's take a look at one example instance of those randomly sampled SCM instances. Its detailed parameterization is provided as follows:

$$
\begin{aligned}
P(s_0) &= 0.5, & P(s_1) &= 0.5 \\
P(x_0, y_0, s_0' \mid s_0) &= 0.1888, & P(x_0, y_0, s_1' \mid s_0) &= 0.2099, \\
P(x_0, y_1, s_0' \mid s_0) &= 0.0294, & P(x_0, y_1, s_1' \mid s_0) &= 0.2116, \\
P(x_1, y_0, s_0' \mid s_0) &= 0.1465, & P(x_1, y_0, s_1' \mid s_0) &= 0.0226, \\
P(x_1, y_1, s_0' \mid s_0) &= 0.0645, & P(x_1, y_1, s_1' \mid s_0) &= 0.1267, \\
P(x_0, y_0, s_0' \mid s_1) &= 0.1762, & P(x_0, y_0, s_1' \mid s_1) &= 0.1775, \\
P(x_0, y_1, s_0' \mid s_1) &= 0.0290, & P(x_0, y_1, s_1' \mid s_1) &= 0.1786, \\
P(x_1, y_0, s_0' \mid s_1) &= 0.1761, & P(x_1, y_0, s_1' \mid s_1) &= 0.0893, \\
P(x_1, y_1, s_0' \mid s_1) &= 0.1472, & P(x_1, y_1, s_1' \mid s_1) &= 0.0261,
\end{aligned}
\tag{72}
$$

where s' denotes the next state; $P(x_0, y_0, s_0' \mid s_0)$ is the abbreviation format for $P(X_t = x_0, Y_t = y_0, S_{t+1} = s_0' \mid S_t = s_0)$.

The expert is able to observe the state S_t, the unobserved variable U_t, and the reward Y_t. However, the imitator, lacking access to both U_t or the reward $\mathbb{E}_\pi[Y_t]$, makes decisions solely on S_t. In other words, all methods utilize the same policy scope $\pi(x \mid s)$. As shown in Fig. 4a, imitators consistently failed to match expert performance. Prevalent negative performance gaps indicate that most of imitators were significantly worse than experts; only in rare cases did the performance gaps near -0.5, supporting our theoretical insights presented in Thm. 1. Furthermore, as depicted in Fig. 5a, CAIL does not achieve expert-level performance, specifically, $\mathbb{E}_\pi[R_t] - \mathbb{E}[R_t] = -1.9019$. However, although CAIL performs worse than the expert, CAIL still consistently outperforms BC and GAIL by effectively learning from the constructed worst-case MDP instances.
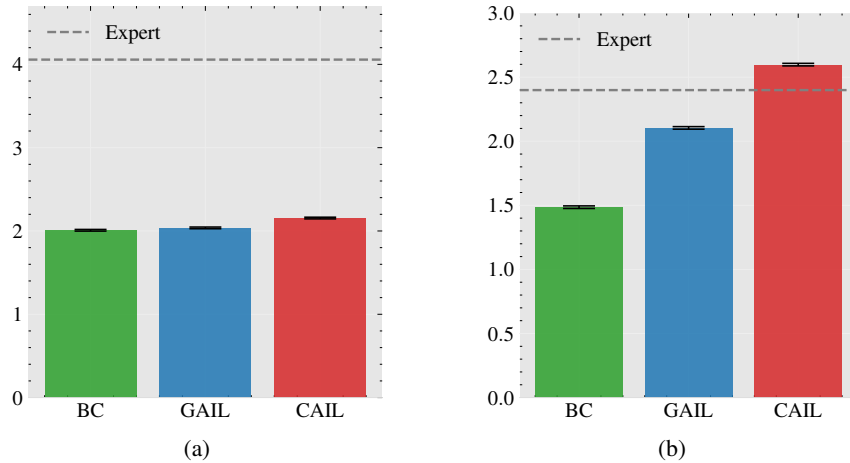
Figure 5: Simulation results for experiments that are not included in the main manuscript.

MDP$_{\text{obs-Y}}$: Additional Experiment. Consider an SCM instance compatible with Fig. 3a including binary observed variables $S_t, X_t, Y_t \in \{0, 1\}$. S_t represents the state at each time step. X_t denotes the action. The unobserved variable U_t represents some information accessible to the expert but

21

inaccessible to the imitator. Additionally, the imitator lacks access to the reward $\mathbb{E}_\pi[Y_t]$. Its detailed parameterization is provided as follows:

$$
\begin{aligned}
P(s_0) &= 0.5, & P(s_1) &= 0.5 \\
P(x_0, y_0, s_0' \mid s_0) &= 0.1775, & P(x_0, y_0, s_1' \mid s_0) &= 0.2029, \\
P(x_0, y_1, s_0' \mid s_0) &= 0.0001, & P(x_0, y_1, s_1' \mid s_0) &= 0.0001, \\
P(x_1, y_0, s_0' \mid s_0) &= 0.0993, & P(x_1, y_0, s_1' \mid s_0) &= 0.0199, \\
P(x_1, y_1, s_0' \mid s_0) &= 0.2001, & P(x_1, y_1, s_1' \mid s_0) &= 0.3001, \\
P(x_0, y_0, s_0' \mid s_1) &= 0.2859, & P(x_0, y_0, s_1' \mid s_1) &= 0.1359, \\
P(x_0, y_1, s_0' \mid s_1) &= 0.0001, & P(x_0, y_1, s_1' \mid s_1) &= 0.0001, \\
P(x_1, y_0, s_0' \mid s_1) &= 0.2969, & P(x_1, y_0, s_1' \mid s_1) &= 0.2809, \\
P(x_1, y_1, s_0' \mid s_1) &= 0.0001, & P(x_1, y_1, s_1' \mid s_1) &= 0.0001,
\end{aligned}
\tag{73}
$$

where s' denotes the next state; $P(x_0, y_0, s_0' \mid s_0)$ is the abbreviation format for $P(X_t = x_0, Y_t = y_0, S_{t+1} = s_0' \mid S_t = s_0)$. As depicted in Fig. 5b, CAIL performs the best among all strategies. Both BC and GAIL fail to match expert performance. Such result shows the effectiveness of Alg. 1.

E Broader Impacts

This paper investigates the theoretical framework of causal imitation learning from confounded demonstrations. Our framework is versatile, applicable to various real-world domains such as autonomous driving, robotics, industrial automation, and medical decisions modeling. One of the positive impacts of this study is the exploration of the risks associated with training IRL algorithms when demonstrations are generally contaminated by unobserved confounders. We theoretically prove that when both the transition distribution $\mathcal{T}$ and reward function $\mathcal{R}$ are not identifiable, there is no policy π learnable from confounded demonstrations that is guaranteed to perform at least as the expert in all possible scenarios. Such theoretical findings have been validated through extensive randomly generated causal models. When either the reward function or the transition distribution is confounded, we augment the GAIL framework by utilizing partial identification techniques, so that the imitator is optimized within the worst-case scenarios. Specfically, the worst-case reward function in Alg. 1 and the worst-case occupancy measure in Alg. 2. By mitigating the risks associated with unobserved confounders in expert demonstrations, our framework supports the development of more transparent and accountable AI systems. This transparency is crucial in high-stakes areas such as healthcare and transportation, where decision-making errors can have significant repercussions. More broadly, our framework significantly enhances the reliability and safety of autonomous systems in various fields, which prioritize safety and robustness during their decision-making processes. They are increasingly important because black-box AI systems, – whose internal workings remain opaque – become more and more prevalent, and our understandings of their potential implications remain limited.

F Impossibility Result in Two-Stage MDPs

In this extension of the MAB model introduced in Sec. 1, we explore a two-stage framework (see Fig. 2b). Our previous discussions demonstrated that in MAB settings affected by unobserved confounders, the expert consistently outperforms the imitator; that is, i.e., $\mathbb{E}_x[Y] < \mathbb{E}[Y]$.

We now extend our analysis to the two-stage MDPs. Specifically, the agent first observes the state S_1, selects an action X_1, and subsequently, it receives a reward Y_1. The process then progresses to the second stage, where the agent transitions to state S_2. It chooses an action X_2, and then it receives a further reward Y_2. A pivotal distinction between this scenario and prior examples lies in the transition probability $P_{\pi_1}(S_2 \mid S_1)$. Therefore, we investigate their cumulative reward:

$$
\mathbb{E}_{\pi_1, \pi_2}[Y_1 + Y_2] \qquad \text{and} \qquad \mathbb{E}[Y_1 + Y_2]. \tag{74}
$$

As a motivating example, we assume that all variables are binary. Our analysis begins by comparing the performance at the final stage, specifically, $\mathbb{E}_{\pi_1, \pi_2}[Y_2]$.

Suppose $f(S_2) = \mathbb{E}[Y_2 \mid S_2, X_2]P(X_2 \mid S_2)$. Without loss of generality, we assume an ordering in the functional values associated with different states: $f(S_2 = 0) > f(S_2 = 1)$. To address the

22

non-identifiability issue caused by the transition distribution $P_{\pi_1}(S_2 \mid S_1)$, as discussed in Eq. (9), we formulate the worst-case SCM by allocating $f(S_2 = 0)$ with probability mass $P(S_2 = 0, X_1 \mid S_1)$. In other words, we assign the lower bound $P(S_2 = 0, x_1 \mid Z_1)$ to the non-identifiable query $P_{x_1}(S_2 = 0 \mid Z_1)$. As such, we are able to rewrite the expert's rewards as follows:

$$\mathbb{E}[Y_2] = f(S_2 = 0) * P(S_2 = 0, X_1 = 0|Z_1)P(Z_1) \tag{75}$$
$$+ f(S_2 = 0) * P(S_2 = 0, X_1 = 1|Z_1)P(Z_1) \tag{76}$$
$$+ f(S_2 = 1) * P(S_2 = 1, X_1 = 0|Z_1)P(Z_1) \tag{77}$$
$$+ f(S_2 = 1) * P(S_2 = 1, X_1 = 1|Z_1)P(Z_1) \tag{78}$$

and the imitator's reward can be written as

$$\mathbb{E}_{\pi_1,\pi_2}[Y_2] = \pi_1(X_1 = 0|Z_1) \cdot A + \pi_1(X_1 = 1|Z_1) \cdot B \tag{79}$$
$$A = f(S_2 = 0) * P(S_2 = 0, X_1 = 0|Z_1)P(Z_1) \tag{80}$$
$$+ f(S_2 = 1) * (1 - P(S_2 = 0, X_1 = 0|Z_1))P(Z_1) \tag{81}$$
$$B = f(S_2 = 0) * P(S_2 = 0, X_1 = 1|Z_1)P(Z_1) \tag{82}$$
$$+ f(S_2 = 1) * (1 - P(S_2 = 0, X_1 = 1|Z_1))P(Z_1) \tag{83}$$

where is $\mathbb{E}_{\pi_1,\pi_2}[Y_2]$ a convex combination of the quantities A and B. Therefore, $\mathbb{E}_{\pi_1,\pi_2}[Y_2] \leq \max\{A, B\}$. Given that $f(S_2 = 0) > f(S_2 = 1)$, we are able to establish that $A < \mathbb{E}[Y_2]$ and $B < \mathbb{E}[Y_2]$. Therefore, $\mathbb{E}_{\pi_1,\pi_2}[Y_2] < \mathbb{E}[Y_2]$. Using a similar rationale introduced in Sec. 1, we get $\mathbb{E}_{\pi_1}[Y_1] < \mathbb{E}[Y_1]$. Consequently,

$$\mathbb{E}_{\pi_1,\pi_2}[Y_1 + Y_2] < \mathbb{E}[Y_1 + Y_2]. \tag{84}$$

In other words, the imitator is unable to learn a policy that can obtain the expert's performance in the worst-case 2-stage MDP compatible with the observational distribution $P(X_1, X_2, S_1, S_2, Y_1, Y_2)$.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The major claims made in the abstract and introduction has accurately reflect the paper's contributions and scope. Specifically, we summarized our contributions in Sec. 1, e.g., the theoretical findings and the propose innovative algorithms. Additionally, Table 1 provides a brief summary of this paper's main contributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the requirements and limitations of the work in Sec. 1 and Sec. 2. To address infinite horizon decision-making challenges, we utilize the Markov Property, as outlined in Def. 1. However, our study generalizes standard imitation methods by focusing on scenarios where causal consistency does not universally hold true.

23

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The assumptions and problem settings can be found in Sec. 2. Due to space constraints, all detailed proofs are provided in Appendices A and B.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All the important details to reproduce the major experimental results in this paper can be found Sec. 4 and Appendix D. Proposed algorithms are provided in Alg. 1 and Alg. 2.

   Guidelines:

24

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: If the paper is accepted, we intend to make the source code available in the camera-ready version of the paper. During the meantime, all the important details to reproduce the major experimental results in this paper can be found in Sec. 4 and Appendix D.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

25

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All the important training and test details in this paper can be found Sec. 4 and Appendix D.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Error bars and other appropriate information about the statistical significance the experiments could be found in Sec. 4 and Appendix D.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Detailed information on the computer resources can be found in Sec. 4 and Appendix D.

26

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: This paper introduces novel causal imitation learning algorithms that adapt to confounded expert demonstrations within MDPs by using partial identification techniques. The research conducted in this conform with paper NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Societal impacts are discussed in Sec. 1 and Appendix E. Our framework can be applied to various fields in reality, including autonomous driving, robotics, industrial automation, medical decisions modeling and so on. One of positive impacts of this work is that we discuss the potential risk of training IRL algorithms when demonstrations are contaminated by unobserved confounders, and how to utilize partial identification techniques to make the imitator robust.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

27

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer:[Yes]

Justification: Please check Sec. 4 and Appendix D.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Please check Sec. 4 and Appendix D.

Guidelines:

- The answer NA means that the paper does not release new assets.

28

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing with human subjects.

    Guidelines:
    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing with human subjects.

    Guidelines:
    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

29