Appendix A Table of notation table

Table A1 summarizes notations used in this paper.

Table A1 A summary of the notation used in this paper.

Notation	Description
$\mathcal{S}$	State space
$\mathcal{A}$	Action space
$\mathbf{s}$	A vector of state in $\mathcal{S}$, i.e., $\mathbf{s} = [s_1, s_2, \ldots, s_{\|f\|}]$
$\|\mathcal{S}\|$	The number of states in the state space.
$p(\mathbf{s}'\|\mathbf{s}, a)$	the dynamic transition from state $\mathbf{s} \in \mathcal{S}$ to the next state $\mathbf{s}'$ when performing action $a \in \mathcal{A}$ in state $\mathbf{s}$
$r(\mathbf{s}, a)$	A reward on state $\mathbf{s}$ and action a
γ	The discount factor
s_i	The i-th state variable.
s_i^t	The i-th state variable at time t.
$V_{\mathcal{S}}$	The vertex set on causal graph defined on the state variables
E	The causal edge set in the causal graph
$\mathcal{G}$	Causal graph that contains vertex $V_{\mathcal{S}}$ and edge set E
$\mathbf{Pa}_i^{\mathcal{G}}$	The parent set of s_i in graph $\mathcal{G}$.
a_i	The action (treatment) on state s_i.
$\mathbf{G}$	The adjacency matrix of the causal graph.
$\mathcal{C}^{Att}_{s_i \to s_j}$	The average treatment effect for the treated sample from s_i to s_j when s_i is treated.
$\hat{\mathcal{C}}^{Att}_{s_i \to s_j}$	The estimated ATT of $\mathcal{C}^{Att}_{s_i \to s_j}$.
$M_{\mathbf{s}}(\mathcal{G})$	The causal mask in the causal policy where $M_{\mathbf{s}}(\mathcal{G}) = \{m^{\mathcal{G}}_{\mathbf{s},a}\}_{a=1}^{\|\mathcal{A}\|}$
$m^{\mathcal{G}}_{\mathbf{s},a}$	The element of mask on action a in the state $\mathbf{s}$ on causal graph $\mathcal{G}$
$D_{TV}(\cdot,\cdot)$	Total variation distance.
$V_{\pi_{\mathcal{G}}}$	The value function on policy $pi_{\mathcal{G}}$
$\mathbf{h}_{pi_{\mathcal{G}}}$	State distribution of causal policy $\pi_{\mathcal{G}}$
$\mathbf{P}_{\pi_{\mathcal{G}}}(\mathbf{s}'\|\mathbf{s})$	The $\|\mathcal{S}\| \times \|\mathcal{S}\|$ state matrix and its entry in s', s where each present a probability from s to s' in policy $\pi_{\mathcal{G}}$
$M_{\pi_{\mathcal{G}}}$	The $\|\mathcal{S}\| \times \|\mathcal{A}\|\|\mathcal{S}\|$ transition matrix.
$R_{\max}$	The max reward.
$A \perp\!\!\!\perp_p B$	Denote the statistical independence constraint between variables A and B.
$A \perp\!\!\!\perp_p B \mid C$	Denote the statistical conditional independence constraint between variables A and B conditioned on C.

Appendix B Theoretical proofs

Appendix B.1 Causal discovery

In this section, we provide proof of the identifiability of causal order in the orientation step and the identifiability of causal structure after the pruning step. In identifying the causal order, we utilize the average treatment effect in treated (ATT) [57] which can be written as follows:

$$\mathcal{C}^{Att}_{s_i \to s_j} = \mathbb{E}[s_j(I_i = 1) - s_j(I_i = 0) | I_i = 1], \tag{B1}$$

where $s_j(a_i = 1)$ denotes the potential outcome of s_j if s_i were treated, $s_j(a_i = 0)$ denotes the potential outcome if s_i were not treated [53], and $\mathbb{E}$ denotes the expectation.

Theorem B1. Given a causal graph $\mathcal{G} = (V_{\mathcal{S}}, E)$, for each pair of states s_i, s_j with $i \neq j$, s_i is the ancestor of s_j if and only if $|\mathcal{C}^{Att}_{s_i \to s_j}| > 0$.

Proof. [Proof of Theorem B1.]
$\Longrightarrow$: If s_i is the ancestor of s_j, then the intervention of s_i will force manipulating the value of s_i by definition and thus result in the change of s_j compared with the s_j without intervention. That is, $s_j(a_i = 1) \neq s_j(I_i = 0)$ and therefore $|s_j(I_i = 1) - s_j(I_i = 0)| > 0$. By taking the average in population that is treated, we obtain $E[|s_j(I_i = 1) - s_j(a_i = 0)||I_i = 1] > 0$.
$\Longleftarrow$: Similarly, if $|\mathcal{C}^{Att}_{s_i \to s_j}| > 0$, we have $|s_j(I_i = 1) - s_j(I_i = 0)| > 0$ based on Eq. (B1). To show s_i is the ancestor of s_j, we prove by contradiction. Suppose s_i is not the ancestor of s_j, then the intervention of s_i will not change the value of s_j. That is, $s_j(I_i = 1) = s_j(I_i = 0)$ which creates the contradiction. Thus, s_i is the ancestor of s_j which finishes the proof.
The following theorem shows that the causal structure is identifiable given the correct causal order. The overall proof is built based on [41]. The main idea is that the causal structure can be identified given the correct causal order if we can identify the causal skeleton. To learn the causal skeleton, we can resort to identifying the (conditional) independence among the variables. Thus, in the following, we will show that under the causal Markov assumption, faithfulness assumption and the sufficiency assumption, the (conditional) independence of the variables can be identified by the proposed BIC score in our work due to its locally consistent property. We begin with the definition of the locally consistent scoring criterion.

Definition B1 (Locally consistent scoring criterion)**.** Let D be a set of data consisting of m records that are iid samples from some distribution $p(\cdot)$. Let $\mathcal{G}$ be any DAG, and let $\mathcal{G}'$ be the DAG that results from adding the edge $X_i \to X_j$. A scoring criterion $S(\mathcal{G}, D)$ is locally consistent if in the limit as m grows large the following two properties hold:

1. If $X_j \not\perp\!\!\!\perp_p X_i \mid X_{\mathbf{Pa}_j^{\mathcal{G}}}$, then $S(\mathcal{G}', D) > S(\mathcal{G}, D)$.

2. If $X_j \perp\!\!\!\perp_p X_i \mid X_{\mathbf{Pa}_j^{\mathcal{G}}}$, then $S(\mathcal{G}', D) < S(\mathcal{G}, D)$.

Lemma B1 (Lemma 7 in [41])**.** The Bayesian scoring criterion (BIC) is locally consistent.

Note that, as pointed out by [41], the BIC, which can be rewritten as the ℓ_0-norm penalty as Eq. (6) in the main text, is locally consistent. This property allows us to correctly identify the independence relationship among states by using the locally consistent BIC score because we can always obtain a greater score if the searched graph consists of (conditional) independence in the data. Thus, we can always search a causal graph $\mathcal{G}$ with the highest score that is 'correct' in the sense that all (conditional) independence consists of the ground truth. This is concluded by the following theorem:

Theorem B2 (Identifiability)**.** Under the causal faithfulness and causal sufficiency assumptions, given the correct causal order and large enough data, the causal structure among states is identifiable from observational data.

Proof. [Proof of Theorem B2] Based on Lemma B1, Eq. (6) in the main text is locally consistent since it has the same form of the BIC score and we denote it using $S(\mathcal{G}, D)$. Then we can prune the redundant edge if $S(\mathcal{G}', D) > S(\mathcal{G}, D)$ where $\mathcal{G}'$ is the graph that removes one of the redundant edges. The reason is that for any pair of state s_i, s_j is redundant, there must exist a conditional set $\mathbf{Pa}^{\mathcal{G}}(s_j)$ such that $s_i \perp\!\!\!\perp s_j \mid Pa_{\mathcal{G}}(s_j)$. Then based on the second property in Definition B1, we have $S(\mathcal{G}', D) > S(\mathcal{G}, D)$ since $\mathcal{G}$ can be seen as the graph that adds a redundant edge from $\mathcal{G}'$. Moreover, since we have causal faithfulness and causal sufficiency assumptions, such independence will be faithful to the causal graph, and thus, by repeating the above step, we are able to obtain the correct causal structure.

Appendix B.2 Policy performance guarantee

In this section, we provide the policy performance guarantees step by step. We first recap the causal policy in the following definition:

Definition B2 (Causal policy)**.** Given a causal graph $\mathcal{G}$, we define the causal policy $\pi_{\mathcal{G}}(\cdot|\mathbf{s})$ under the causal graph $\mathcal{G}$ as follows:

$$\pi_{\mathcal{G}}(\cdot|\mathbf{s}) = M_{\mathbf{s}}(\mathcal{G}) \circ \pi(\cdot|\mathbf{s}), \tag{B2}$$

where $M_{\mathbf{s}}(\mathcal{G})$ is the causal mask vector at state $\mathbf{s}$ under the causal graph $\mathcal{G}$, and $\pi(\cdot|\mathbf{s})$ is the action probability distribution of the original policy output.

For example, the causal mask $M_{\mathbf{s}}(\mathcal{G}) = \{m^{\mathcal{G}}_{\mathbf{s},a}\}_{a=1}^{|\mathcal{A}|}$ constitute the vector of mask $m^{\mathcal{G}}_{\mathbf{s},a} \in \{0, 1\}$ of each action in $\mathcal{A}$ where $|\mathcal{A}|$ denotes the number of actions in the action space.

Outline of the proof of Theorem 3. Our goal is to show that under the causal policy, the value function under the correct causal graph will have greater value than the value function that has misspecified causal graph such that the differences of the value function can be bound by some constant $c > 0$:

$$V_{\pi_{\mathcal{G}^*}} - V_{\pi_{\mathcal{G}}} \leqslant c. \tag{B3}$$

To do so, one may first notice that the difference of the value function can be expressed and bounded by the total variation $D_{\mathrm{TV}}(\rho_{\pi_{\mathcal{G}}}, \rho_{\pi_{\mathcal{G}^*}})$:

$$|V_{\pi_{\mathcal{G}^*}} - V_{\pi_{\mathcal{G}}}| \leqslant \frac{2R_{\max}}{1-\gamma} D_{\mathrm{TV}}(\rho_{\pi_{\mathcal{G}}}, \rho_{\pi_{\mathcal{G}^*}}). \tag{B4}$$

Such a total variation can be further bound by the total variation of $D_{\mathrm{TV}}(\pi_{\mathcal{G}}(\cdot \mid \mathbf{s}), \pi_{\mathcal{G}^*}(\cdot \mid \mathbf{s}))$ (Lemma B3 and Lemma B4):

$$D_{\mathrm{TV}}(\rho_{\pi_{\mathcal{G}}}, \rho_{\pi_{\mathcal{G}^*}}) \leqslant \frac{1}{1-\gamma} \mathbb{E}_{\mathbf{s}\sim\mathbf{h}_{\pi_{\mathcal{G}^*}}} [D_{\mathrm{TV}}(\pi_{\mathcal{G}}(\cdot \mid \mathbf{s}), \pi_{\mathcal{G}^*}(\cdot \mid \mathbf{s}))]. \tag{B5}$$

Combining Eq. (B4) and Eq. (B5), we have

$$|V_{\pi_{\mathcal{G}^*}} - V_{\pi_{\mathcal{G}}}| \leqslant \frac{2R_{\max}}{(1-\gamma)^2} \mathbb{E}_{\mathbf{s}\sim\mathbf{h}_{\pi_{\mathcal{G}^*}}} [D_{\mathrm{TV}}(\pi_{\mathcal{G}}(\cdot \mid \mathbf{s}), \pi_{\mathcal{G}^*}(\cdot \mid \mathbf{s}))]. \tag{B6}$$

By this, we can delve into this bound by investigating the total variation of the causal policy. Based on the definition of the causal policy in Definition B2. One can deduce that the distance should be related to the difference of the causal mask, and it is true that as shown in Lemma B2:

$$D_{TV}(\pi_{\mathcal{G}^*}, \pi_{\mathcal{G}}) \leqslant \frac{1}{2}(\|M_{\mathbf{s}}(\mathcal{G}) - M_{\mathbf{s}}(\mathcal{G}^*)\|_1 + \|\mathbf{1}_{\left\{a: m^{\mathcal{G}^*}_{\mathbf{s},a}=1 \wedge m^{\mathcal{G}}_{\mathbf{s},a}=1\right\}}\|_1). \tag{B7}$$

Finally, by combining Eq. (B6) and Eq. (B7) and further due to the positive of the bound, we obtain the result in Theorem B3:

$$\begin{aligned} V_{\pi_{\mathcal{G}^*}} - V_{\pi_{\mathcal{G}}} \leqslant & \frac{R_{\max}}{(1-\gamma)^2}(\|M_{\mathbf{s}}(\mathcal{G}) - M_{\mathbf{s}}(\mathcal{G}^*)\|_1 \\ & + \|\mathbf{1}_{\{a: m^{\mathcal{G}^*}_{\mathbf{s},a}=1 \wedge m^{\mathcal{G}}_{\mathbf{s},a}=1\}}\|_1). \end{aligned} \tag{B8}$$

With the outline above, in the following, we provide the details proof of the Lemma B3, Lemma B4, Lemma B2, and Theorem B3, respectively.

Lemma B2. Let $\pi_{\mathcal{G}^*}(\cdot|\mathbf{s})$ be the policy under the true causal graph $\mathcal{G}^* = (V_{\mathcal{S}}, E^*)$. For any causal graph $\mathcal{G} = (V_{\mathcal{S}}, E)$, when the defined causal policy $\pi_{\mathcal{G}}(\cdot|\mathbf{s})$ converges, the following inequality holds:

$$D_{TV}(\pi_{\mathcal{G}^*}, \pi_{\mathcal{G}}) \leqslant \frac{1}{2}(\|M_{\mathbf{s}}(\mathcal{G}) - M_{\mathbf{s}}(\mathcal{G}^*)\|_1 + \|\mathbf{1}_{\{a: m^{\mathcal{G}^*}_{\mathbf{s},a}=1 \wedge m^{\mathcal{G}}_{\mathbf{s},a}=1\}}\|_1), \tag{B9}$$

where $\|M_{\mathbf{s}}(\mathcal{G}) - M_{\mathbf{s}}(\mathcal{G}^*)\|_1$ is the ℓ_1-norm of the masks measuring the differences of two policies, $\mathbf{1}$ is an indicator function and $\|\mathbf{1}_{\left\{a: m^{\mathcal{G}^*}_{\mathbf{s},a}=1 \wedge m^{\mathcal{G}}_{\mathbf{s},a}=1\right\}}\|_1$ measures the number of actions that are not masked on both policies.

Proof. [Proof of Lemma B2] Based on the definition of the total variation and the causal policy we have:

$$\begin{aligned} D_{TV}(\pi_{\mathcal{G}^*}, \pi_{\mathcal{G}}) &= \frac{1}{2}\|\pi_{\mathcal{G}^*}(\cdot|\mathbf{s}) - \pi_{\mathcal{G}}(\cdot|\mathbf{s})\|_1 \\ &= \frac{1}{2}\sum_a |\pi_{\mathcal{G}^*}(a|\mathbf{s}) - \pi_{\mathcal{G}}(a|\mathbf{s})| \\ &= \frac{1}{2}\sum_a |m^{\mathcal{G}^*}_{\mathbf{s},a}\pi^*(a|\mathbf{s}) - m^{\mathcal{G}}_{\mathbf{s},a}\pi(a|\mathbf{s})|. \end{aligned} \tag{B10}$$

Since the mask only takes value in $\{0, 1\}$, we can rearrange the summation by considering the different values of the mask on the two policies:

$$D_{TV}(\pi_{\mathcal{G}^*}, \pi_{\mathcal{G}}) = \frac{1}{2}\left(\sum_{a: m^{\mathcal{G}^*}_{\mathbf{s},a}=1 \wedge m^{\mathcal{G}}_{\mathbf{s},a}=0} |\pi^*(a|\mathbf{s})| + \sum_{a: m^{\mathcal{G}^*}_{\mathbf{s},a}=0 \wedge m^{\mathcal{G}}_{\mathbf{s},a}=1} |\pi(a|\mathbf{s})| + \sum_{a: m^{\mathcal{G}^*}_{\mathbf{s},a}=1 \wedge m^{\mathcal{G}}_{\mathbf{s},a}=1} |\pi^*(a|\mathbf{s}) - \pi(a|\mathbf{s})|\right), \tag{B11}$$

where the summation when $m^{\mathcal{G}^*}_{\mathbf{s},a} = 0 \wedge m^{\mathcal{G}}_{\mathbf{s},a} = 0$ is zero as policy on both side are masked out. Then, based on the fact that $0 \leqslant \pi(a|\mathbf{s}) \leqslant 1$ of the policy, we have the following inequality

$$D_{TV}(\pi_{\mathcal{G}^*}, \pi_{\mathcal{G}}) \leqslant \frac{1}{2}\left(\|M_{\mathbf{s}}(\mathcal{G}) - M_{\mathbf{s}}(\mathcal{G}^*)\|_1 + \|\mathbf{1}_{\left\{a: m^{\mathcal{G}^*}_{\mathbf{s},a}=1 \wedge m^{\mathcal{G}}_{\mathbf{s},a}=1\right\}}\|_1\right). \tag{B12}$$

Then we introduce the following Lemma B3, which bound the state distribution discrepancy based on the causal policy discrepancy.

Lemma B3. Given a policy $\pi_{\mathcal{G}^*}(\cdot|\mathbf{s})$ under the true causal structure $\mathcal{G}^* = (V, E^*)$ and an policy $\pi_{\mathcal{G}}(\cdot|\mathbf{s})$ under the causal graph $\mathcal{G} = (V, E)$, we have that

$$D_{TV}(\mathbf{h}_{\pi_{\mathcal{G}}}(\mathbf{s}), \mathbf{h}_{\pi_{\mathcal{G}^*}}(\mathbf{s})) \leqslant \frac{1}{1-\gamma}\mathbb{E}_{\mathbf{s}\sim\mathbf{h}_{\pi_{\mathcal{G}^*}}}[D_{\mathrm{TV}}(\pi_{\mathcal{G}}(\cdot \mid \mathbf{s}), \pi_{\mathcal{G}^*}(\cdot \mid \mathbf{s}))]. \tag{B13}$$

Proof. [Proof of Lemma B3] The proof is inspired by [61], we show that the state distribution $\mathbf{h}_{\pi_{\mathcal{G}}}$ of causal policy $\pi_{\mathcal{G}}$ can be denoted as

$$\mathbf{h}_{\pi_{\mathcal{G}}} = (1-\gamma)(I - \gamma\mathbf{P}_{\pi_{\mathcal{G}}})^{-1}\mathbf{h}_0, \tag{B14}$$

where $\mathbf{P}_{\pi_{\mathcal{G}}}(\mathbf{s}'|\mathbf{s}) = \sum_{a\in\mathcal{A}} M^*(\mathbf{s}' \mid \mathbf{s}, a)\pi_{\mathcal{G}}(a \mid \mathbf{s})$, and $M^*(\mathbf{s}' \mid \mathbf{s}, a)$ is the dynamic model. Denote that $M_{\pi_{\mathcal{G}}} = (I - \gamma\mathbf{P}_{\pi_{\mathcal{G}}})^{-1}$, we then have

$$\begin{aligned} \mathbf{h}_{\pi_{\mathcal{G}}} - \mathbf{h}_{\pi_{\mathcal{G}^*}} &= (1-\gamma)\left[(I - \gamma\mathbf{P}_{\pi_{\mathcal{G}}})^{-1} - (I - \gamma\mathbf{P}_{\pi_{\mathcal{G}^*}})^{-1}\right]\mathbf{h}_0 \\ &= (1-\gamma)(M_{\pi_{\mathcal{G}}} - M_{\pi_{\mathcal{G}^*}})\mathbf{h}_0 \\ &= (1-\gamma)\gamma M_{\pi_{\mathcal{G}}}(\mathbf{P}_{\pi_{\mathcal{G}}} - \mathbf{P}_{\pi_{\mathcal{G}^*}})M_{\pi_{\mathcal{G}^*}}\mathbf{h}_0 \\ &= \gamma M_{\pi_{\mathcal{G}}}(\mathbf{P}_{\pi_{\mathcal{G}}} - \mathbf{P}_{\pi_{\mathcal{G}^*}})\mathbf{h}_{\pi_{\mathcal{G}^*}}. \end{aligned} \tag{B15}$$

Similarly to Lemma 4 in [61], we have

$$\begin{aligned} D_{TV}(\mathbf{h}_{\pi_{\mathcal{G}}}(\mathbf{s}), \mathbf{h}_{\pi_{\mathcal{G}^*}}(\mathbf{s})) &= \frac{\gamma}{2}\|M_{\pi_{\mathcal{G}}}(\mathbf{P}_{\pi_{\mathcal{G}}} - \mathbf{P}_{\pi_{\mathcal{G}^*}})\mathbf{h}_{\pi_{\mathcal{G}^*}}\|_1 \\ &\leqslant \frac{\gamma}{2}\|M_{\pi_{\mathcal{G}}}\|_1\|(\mathbf{P}_{\pi_{\mathcal{G}}} - \mathbf{P}_{\pi_{\mathcal{G}^*}})\mathbf{h}_{\pi_{\mathcal{G}^*}}\|_1. \end{aligned} \tag{B16}$$

Note that

$$\|M_{\pi_{\mathcal{G}}}\|_1 = \|\sum_{t=0}^{\infty}\gamma^t\mathbf{P}^t_{\pi_{\mathcal{G}}}\|_1 \leqslant \sum_{t=0}^{\infty}\gamma^t\|\mathbf{P}_{\pi_{\mathcal{G}}}\|_1^t \leqslant \sum_{t=0}^{\infty}\gamma^t = \frac{1}{1-\gamma}, \tag{B17}$$

and we also show that $\|(\mathbf{P}_{\pi_{\mathcal{G}}} - \mathbf{P}_{\pi_{\mathcal{G}^*}})\mathbf{h}_{\pi_{\mathcal{G}^*}}\|_1$ is bounded by

$$\begin{aligned}\|(\mathbf{P}_{\pi_{\mathcal{G}}} - \mathbf{P}_{\pi_{\mathcal{G}^*}})\mathbf{h}_{\pi_{\mathcal{G}^*}}\|_1 &\leqslant \sum_{\mathbf{s},\mathbf{s}'} |\mathbf{P}_{\pi_{\mathcal{G}}}(\mathbf{s}'|\mathbf{s}) - \mathbf{P}_{\pi_{\mathcal{G}^*}}(\mathbf{s}'|\mathbf{s})|\mathbf{h}_{\pi_{\mathcal{G}^*}}(\mathbf{s}) \\ &= \sum_{\mathbf{s},\mathbf{s}} \left| \sum_{a\in\mathcal{A}} M^*(\mathbf{s} \mid \mathbf{s}, a)(\pi_{\mathcal{G}}(a \mid \mathbf{s}) - \pi_{\mathcal{G}^*}(a \mid \mathbf{s})) \right| \mathbf{h}_{\pi_{\mathcal{G}^*}}(\mathbf{s}) \\ &\leqslant \sum_{(\mathbf{s},a),\mathbf{s}} M^*(\mathbf{s} \mid \mathbf{s}, a)|\pi_{\mathcal{G}}(a \mid \mathbf{s}) - \pi_{\mathcal{G}^*}(a \mid \mathbf{s})|\mathbf{h}_{\pi_{\mathcal{G}^*}}(\mathbf{s}) \\ &= \sum_{s} \mathbf{h}_{\pi_{\mathcal{G}^*}}(\mathbf{s}) \sum_{a\in\mathcal{A}} |\pi_{\mathcal{G}}(a \mid \mathbf{s}) - \pi_{\mathcal{G}^*}(a \mid \mathbf{s})| \\ &= 2\mathbb{E}_{\mathbf{s}\sim\mathbf{h}_{\pi_{\mathcal{G}^*}}}[D_{\mathrm{TV}}(\pi_{\mathcal{G}}(\cdot \mid \mathbf{s}), \pi_{\mathcal{G}^*}(\cdot \mid \mathbf{s}))].\end{aligned} \tag{B18}$$

Thus, we have

$$\begin{aligned}D_{TV}(\mathbf{h}_{\pi_{\mathcal{G}}}(\mathbf{s}), \mathbf{h}_{\pi_{\mathcal{G}^*}}(\mathbf{s})) &\leqslant \frac{\gamma}{2}\|M_{\pi_{\mathcal{G}}}\|_1\|(\mathbf{P}_{\pi_{\mathcal{G}}} - \mathbf{P}_{\pi_{\mathcal{G}^*}})\mathbf{h}_{\pi_{\mathcal{G}^*}}\|_1 \\ &\leqslant \frac{1}{1-\gamma}\mathbb{E}_{\mathbf{s}\sim\mathbf{h}_{\pi_{\mathcal{G}^*}}}[D_{\mathrm{TV}}(\pi_{\mathcal{G}}(\cdot \mid \mathbf{s}), \pi_{\mathcal{G}^*}(\cdot \mid \mathbf{s}))].\end{aligned} \tag{B19}$$

Next, we further bound the state-action distribution discrepancy based on the causal policy discrepancy.

Lemma B4. Given a policy $\pi_{\mathcal{G}^*}(\cdot|\mathbf{s})$ under the true causal structure $\mathcal{G}^* = (V, E^*)$ and an policy $\pi_{\mathcal{G}}(\cdot|\mathbf{s})$ under the causal graph $\mathcal{G} = (V, E)$, we have that

$$D_{\mathrm{TV}}(\rho_{\pi_{\mathcal{G}}}, \rho_{\pi_{\mathcal{G}^*}}) \leqslant \frac{1}{1-\gamma}\mathbb{E}_{\mathbf{s}\sim\mathbf{h}_{\pi_{\mathcal{G}^*}}}[D_{\mathrm{TV}}(\pi_{\mathcal{G}}(\cdot \mid \mathbf{s}), \pi_{\mathcal{G}^*}(\cdot \mid \mathbf{s}))]. \tag{B20}$$

Proof. [Proof of Lemma B4] Note that for any policy $\pi_{\mathcal{G}}$ under any causal graph $\mathcal{G}$, the state-action distribution $\rho_{\pi_{\mathcal{G}}}(\mathbf{s}, a) = \pi_{\mathcal{G}}(a \mid \mathbf{s})\mathbf{h}_{\pi_{\mathcal{G}}}(\mathbf{s})$, we have

$$\begin{aligned}D_{\mathrm{TV}}(\rho_{\pi_{\mathcal{G}}}, \rho_{\pi_{\mathcal{G}^*}}) &= \frac{1}{2}\sum_{(\mathbf{s},a)} |[\pi_{\mathcal{G}^*}(a \mid \mathbf{s}) - \pi_{\mathcal{G}}(a \mid \mathbf{s})]\mathbf{h}_{\pi_{\mathcal{G}}}(\mathbf{s}) + [\mathbf{h}_{\pi_{\mathcal{G}^*}}(\mathbf{s}) - \mathbf{h}_{\pi_{\mathcal{G}}}(\mathbf{s})]\pi_{\mathcal{G}}(a \mid \mathbf{s})| \\ &\leqslant \frac{1}{2}\sum_{(\mathbf{s},a)} |\pi_{\mathcal{G}^*}(a \mid \mathbf{s}) - \pi_{\mathcal{G}}(a \mid \mathbf{s})|\mathbf{h}_{\pi_{\mathcal{G}}}(\mathbf{s}) + \frac{1}{2}\sum_{(\mathbf{s},a)} \pi_{\mathcal{G}}(a \mid \mathbf{s})|\mathbf{h}_{\pi_{\mathcal{G}^*}}(\mathbf{s}) - \mathbf{h}_{\pi_{\mathcal{G}}}(\mathbf{s})| \\ &= \mathbb{E}_{\mathbf{s}\sim\mathbf{h}_{\pi_{\mathcal{G}^*}}}[D_{\mathrm{TV}}(\pi_{\mathcal{G}}(\cdot \mid \mathbf{s}), \pi_{\mathcal{G}^*}(\cdot \mid \mathbf{s}))] + D_{TV}(\mathbf{h}_{\pi_{\mathcal{G}}}(\mathbf{s}), \mathbf{h}_{\pi_{\mathcal{G}^*}}(\mathbf{s})) \\ &\leqslant \frac{1}{1-\gamma}\mathbb{E}_{\mathbf{s}\sim\mathbf{h}_{\pi_{\mathcal{G}^*}}}[D_{\mathrm{TV}}(\pi_{\mathcal{G}}(\cdot \mid \mathbf{s}), \pi_{\mathcal{G}^*}(\cdot \mid \mathbf{s}))],\end{aligned} \tag{B21}$$

where the last inequality follows Lemma B3.

Based on all the above Lemma B4, we finally give the policy performance guarantee of our proposed framework. Specifically, we bound the policy value gap (i.e., the difference between the value of learned causal policy and the optimal policy) based on the state-action distribution discrepancy.

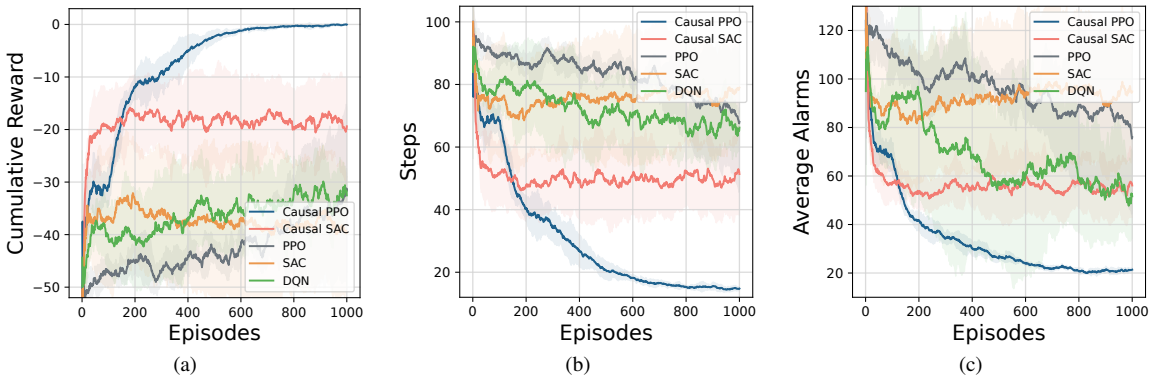
Theorem B3. Given a causal policy $\pi_{\mathcal{G}^*}(\cdot|\mathbf{s})$ under the true causal graph $\mathcal{G}^* = (V_{\mathcal{S}}, E^*)$ and a policy $\pi_{\mathcal{G}}(\cdot|\mathbf{s})$ under the causal graph $\mathcal{G} = (V_{\mathcal{S}}, E)$, recalling $R_{\max}$ is the upper bound of the reward function, we have the performance difference of $\pi_{\mathcal{G}^*}(\cdot|\mathbf{s})$ and $\pi_{\mathcal{G}}(\cdot|\mathbf{s})$ be bounded as below,

$$\begin{aligned}V_{\pi_{\mathcal{G}^*}} - V_{\pi_{\mathcal{G}}} \leqslant &\frac{R_{\max}}{(1-\gamma)^2}(\|M_{\mathbf{s}}(\mathcal{G}) - M_{\mathbf{s}}(\mathcal{G}^*)\|_1 \\ &+ \|\mathbf{1}_{\{a:m^{\mathcal{G}^*}_{\mathbf{s},a}=1 \wedge m^{\mathcal{G}}_{\mathbf{s},a}=1\}}\|_1).\end{aligned} \tag{B22}$$

Proof. [Proof of theorem B3]

Note that for any policy $\pi_{\mathcal{G}}$ under any causal graph $\mathcal{G}$, its policy value can be reformulated as $V_{\pi_{\mathcal{G}}} = \frac{1}{1-\gamma}\mathbb{E}_{(\mathbf{s},a)\sim\rho_{\pi_{\mathcal{G}}}}[r, a]$. Based on this, we have

$$\begin{aligned}|V_{\pi_{\mathcal{G}^*}} - V_{\pi_{\mathcal{G}}}| &= \left|\frac{1}{1-\gamma}\mathbb{E}_{(\mathbf{s},a)\sim\rho_{\pi_{\mathcal{G}}}}[r, a] - \frac{1}{1-\gamma}\mathbb{E}_{(\mathbf{s},a)\sim\rho_{\pi_{\mathcal{G}^*}}}[r, a]\right| \\ &\leqslant \frac{1}{1-\gamma}\sum_{(\mathbf{s},a)\in\mathcal{S}\times\mathcal{A}} |(\rho_{\pi_{\mathcal{G}}}(\mathbf{s}, a) - \rho_{\pi_{\mathcal{G}^*}}(\mathbf{s}, a))r(\mathbf{s}, a)| \\ &\leqslant \frac{2R_{\max}}{1-\gamma}D_{\mathrm{TV}}(\rho_{\pi_{\mathcal{G}}}, \rho_{\pi_{\mathcal{G}^*}}).\end{aligned} \tag{B23}$$



(a) (b) (c)

Figure C1 (a)-(c)Cumulative rewards, intervention steps, and average number of alarms per episode for Causal PPO based on random initialization structures at different K in the topology-free environment.

Table C1 Results of causal structure learning of topology-free environment

Methods	F1 score	Precision	Recall	Accuracy	SHD
Random Initiation	0.006 ± 0.006	0.025 ± 0.025	0.003 ± 0.003	0.669 ± 0.983	169.0 ± 5.362
Causal PPO (Random)	**0.755** ± 0.023	**0.814** ± 0.024	**0.705** ± 0.025	**0.993** ± 0.001	**68.50** ± 6.225
Causal SAC (Random)	0.595 ± 0.027	0.558 ± 0.057	0.643 ± 0.017	0.987 ± 0.002	132.0 ± 15.859

Combining Lemma B4 and Lemma B2, we have

$$\begin{aligned} V_{\pi_{\mathcal{G}^*}} - V_{\pi_{\mathcal{G}}} &\leqslant \frac{2R_{\max}}{1-\gamma} D_{\mathrm{TV}}(\rho_{\pi_{\mathcal{G}}}, \rho_{\pi_{\mathcal{G}^*}}) \\ &\leqslant \frac{2R_{\max}}{(1-\gamma)^2} \mathbb{E}_{\mathbf{s}\sim d_{\pi_{\mathcal{G}^*}}} [D_{\mathrm{TV}}(\pi_{\mathcal{G}}(\cdot \mid \mathbf{s}), \pi_{\mathcal{G}^*}(\cdot \mid \mathbf{s}))] \\ &\leqslant \frac{R_{\max}}{(1-\gamma)^2} \left(\| M_{\mathbf{s}}(\mathcal{G}) - M_{\mathbf{s}}(\mathcal{G}^*) \|_1 + \| \mathbf{1}_{\left\{a: m^{\mathcal{G}^*}_{\mathbf{s},a}=1 \wedge m^{\mathcal{G}}_{\mathbf{s},a}=1\right\}} \|_1 \right), \end{aligned} \tag{B24}$$

which completes the proof.

Appendix C Additional experiment of topology-free environment

Considering that topology-free fault alarm scenarios also exist in real O&M environments, we constructed another topology-free alarm environment with 100-dimensional alarm types based on real alarm data. The specific experimental configurations are shown in the Table E1. We also conducted comparative experiments in this environment. In policy learning, we used the model-free algorithms PPO [25], SAC [28], and DQN [22] as baseline, and applied our method to PPO and SAC, resulting in Causal PPO and Causal SAC. To better demonstrate the advantages of our method in causal structure learning, we use random graphs as the initial structures for the causal learning process.

As shown in Figure C1, our methods outperform the baseline algorithms in terms of cumulative rewards, number of interactions, and average number of alarms per episode metrics. In terms of structure learning, discovering causality among 100-dimensional causal alarm nodes is challenging. However, as shown in Table C1, compared to the randomized initial graph, our approaches can gradually learn a basic causal structure, which helps improve the convergence performance of the policy. This also demonstrates the applicability of our algorithm in multiple scenarios.

Appendix D Additional experiment on cart-pole environment

To evaluate the performance of our approach on classic control tasks, we included the *cart-pole* environment from the OpenAI Gym toolkit. The cart-pole environment is a well-known benchmark in reinforcement learning, where the goal is to balance a pole on a moving cart by applying forces to the cart. The state space consists of the cart's position, velocity, pole angle, and pole angular velocity, while the action space is discrete, allowing the agent to push the cart either left or right.

In the cart-pole environment, there is a clear causal relationship between the pole's angle and the cart's acceleration: when the pole tilts to the right, continuing to apply force in that direction exacerbates the tilt, whereas applying force to the left helps restore balance. Leveraging this causal structure, we introduce a causal action masking mechanism that softly masks actions aligned with the tilt direction at extreme angles, thereby reducing ineffective exploration and expediting policy convergence. Specifically, since the goal Y of cart-pole environment is to control the angle of pole, the causal mask

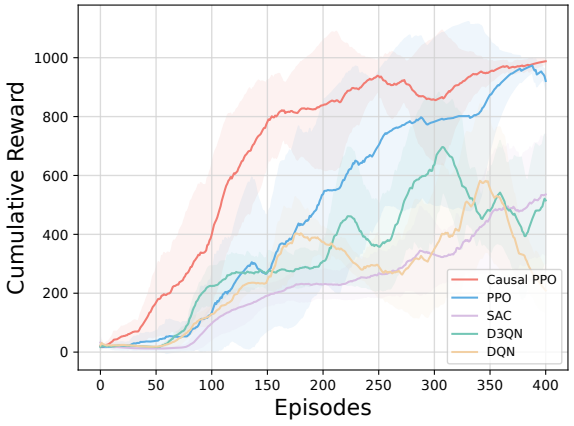


Figure D1 Cumulative rewards in the cart-pole environment.

Table E1 Environment configurations used in experiments.

Environment	Parameters	Value
Topology environment	Max step size	100
	State dimension	1800
	Action dimension	900
	Action type	Discrete
	time range	50
	max hop	2
	α range	[0.0001, 0.0013]
	μ range	[0.0005, 0.0008]
	root cause num	50
Topology-free environment	Max step size	100
	State dimension	200
	Action dimension	100
	Action type	Discrete
	time range	100
	max hop	1
	α range	[0.00015, 0.0025]
	μ range	[0.0005, 0.0008]
	root cause num	20

is learned by setting it proportionally to the effect of the action $m^{\mathcal{G}}_{\mathbf{s},i} \propto |s_{\text{angle}}(I_i = 1)|$ such that the action will more likely be masked if it increases the angle.

The experimental results (shown in Figure D1.) indicate that the proposed Causal PPO significantly outperforms other baselines in terms of cumulative rewards, and demonstrates faster convergence and higher stability during training, which fully proves that explicitly embedding causal inference in the action space is of key significance for efficient reinforcement learning of samples.

Appendix E Hyper-parameters

We first list all important hyper-parameters in the implementation for the specific Topology environment and Topology-free environment in Table E1. We also provide the ground truth of the causal structure that is used in our FaultAlarmRL environment in Table E2. The hyper-parameters for FaultAlarmRL environment are also provided in Table E3.

Table E2 Ground truth

Cause	Effect	Cause	Effect
MW_RDI	LTI	MW_BER_SD	LTI
MW_RDI	CLK_NO_TRACE_MODE	MW_BER_SD	S1_SYN_CHANGE
MW_RDI	S1_SYN_CHANGE	MW_BER_SD	PLA_MEMBER_DOWN
MW_RDI	LAG_MEMBER_DOWN	MW_BER_SD	MW_RDI
MW_RDI	PLA_MEMBER_DOWN	MW_BER_SD	MW_LOF
MW_RDI	ETH_LOS	MW_BER_SD	ETH_LINK_DOWN
MW_RDI	ETH_LINK_DOWN	MW_BER_SD	NE_COMMU_BREAK
MW_RDI	NE_COMMU_BREAK	MW_BER_SD	R_LOF
MW_RDI	R_LOF	R_LOF	LTI
TU_AIS	LTI	R_LOF	S1_SYN_CHANGE
TU_AIS	CLK_NO_TRACE_MODE	R_LOF	LAG_MEMBER_DOWN
TU_AIS	S1_SYN_CHANGE	R_LOF	PLA_MEMBER_DOWN
RADIO_RSL_LOW	LTI	R_LOF	ETH_LINK_DOWN
RADIO_RSL_LOW	S1_SYN_CHANGE	R_LOF	NE_COMMU_BREAK
RADIO_RSL_LOW	LAG_MEMBER_DOWN	LTI	CLK_NO_TRACE_MODE
RADIO_RSL_LOW	PLA_MEMBER_DOWN	HARD_BAD	LTI
RADIO_RSL_LOW	MW_RDI	HARD_BAD	CLK_NO_TRACE_MODE
RADIO_RSL_LOW	MW_LOF	HARD_BAD	S1_SYN_CHANGE
RADIO_RSL_LOW	MW_BER_SD	HARD_BAD	BD_STATUS
RADIO_RSL_LOW	ETH_LINK_DOWN	HARD_BAD	POWER_ALM
RADIO_RSL_LOW	NE_COMMU_BREAK	HARD_BAD	LAG_MEMBER_DOWN
RADIO_RSL_LOW	R_LOF	HARD_BAD	PLA_MEMBER_DOWN
BD_STATUS	S1_SYN_CHANGE	HARD_BAD	ETH_LOS
BD_STATUS	LAG_MEMBER_DOWN	HARD_BAD	MW_RDI
BD_STATUS	PLA_MEMBER_DOWN	HARD_BAD	MW_LOF
BD_STATUS	ETH_LOS	HARD_BAD	ETH_LINK_DOWN
BD_STATUS	MW_RDI	HARD_BAD	NE_COMMU_BREAK
BD_STATUS	MW_LOF	HARD_BAD	R_LOF
BD_STATUS	ETH_LINK_DOWN	HARD_BAD	NE_NOT_LOGIN
BD_STATUS	RADIO_RSL_LOW	HARD_BAD	RADIO_RSL_LOW
BD_STATUS	TU_AIS	HARD_BAD	TU_AIS
NE_COMMU_BREAK	LTI	ETH_LOS	LTI
NE_COMMU_BREAK	CLK_NO_TRACE_MODE	ETH_LOS	CLK_NO_TRACE_MODE
NE_COMMU_BREAK	S1_SYN_CHANGE	ETH_LOS	S1_SYN_CHANGE
NE_COMMU_BREAK	LAG_MEMBER_DOWN	ETH_LOS	LAG_MEMBER_DOWN
NE_COMMU_BREAK	PLA_MEMBER_DOWN	ETH_LOS	PLA_MEMBER_DOWN
NE_COMMU_BREAK	ETH_LOS	ETH_LOS	ETH_LINK_DOWN
NE_COMMU_BREAK	ETH_LINK_DOWN	MW_LOF	LTI
NE_COMMU_BREAK	NE_NOT_LOGIN	MW_LOF	CLK_NO_TRACE_MODE
ETH_LINK_DOWN	LTI	MW_LOF	S1_SYN_CHANGE
ETH_LINK_DOWN	CLK_NO_TRACE_MODE	MW_LOF	LAG_MEMBER_DOWN
ETH_LINK_DOWN	S1_SYN_CHANGE	MW_LOF	PLA_MEMBER_DOWN
S1_SYN_CHANGE	LTI	MW_LOF	ETH_LOS
POWER_ALM	BD_STATUS	MW_LOF	MW_RDI
POWER_ALM	ETH_LOS	MW_LOF	ETH_LINK_DOWN
POWER_ALM	MW_RDI	MW_LOF	NE_COMMU_BREAK
POWER_ALM	MW_LOF	MW_LOF	R_LOF

Table E3 Hyper-parameters of methods used in experiments.

Models	Parameters	Value
Causal DQN & Causal D3QN	Learning rate	0.0003
	Size of buffer $mathcalB$	100000
	Epoch per max iteration	100
	Batch size	64
	Reward discount γ	0.99
	MLP hiddens	128
	MLP layers	2
	Update timestep	5
	Random sample timestep	512
	ϵ-greedy ratio	0.1
	ϵ-causal ratio η	0.2
Causal PPO	Actor learning rate	0.0003
	Critic learning rate	0.0003
	Epoch per max iteration	100
	Batch size	64
	Reward discount γ	0.99
	MLP hiddens	128
	MLP layers	2
	Clip	0.2
	K epochs	50
	Update timestep	256
	Random sample timestep	512
	ϵ-greedy ratio	0.1
	ϵ-causal ratio η	0.3
DQN & D3QN	Learning rate	0.0003
	Size of buffer $mathcalB$	100000
	Epoch per max iteration	100
	Batch size	64
	Reward discount γ	0.99
	MLP hiddens	128
	MLP layers	2
	Update timestep	5
	Random sample timestep	512
	ϵ-greedy ratio	0.1
PPO	Actor learning rate	0.0003
	Critic learning rate	0.0003
	Epoch per max iteration	100
	Batch size	64
	Reward discount γ	0.99
	MLP hiddens	128
	MLP layers	2
	Clip	0.2
	K epochs	50
	Update timestep	512
	Random sample timestep	512