# Actively Identifying Causal Effects with Latent Variables Given Only Response Variable Observable

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Causal effect identification plays an important role in guiding decision-making, and some criteria have been proposed to achieve it with the prior knowledge of *causal graph*. In many real tasks, however, we are merely concerned with the causal effect of each decision variable on one specific target - *response variable*. In this paper, we consider identifying such effects with the existence of *latent confounders* when causal graph is *unknown*. Since some causal effects may be unidentifiable with only observational data, we additionally allow active interventions but restrict that only the response variable can be observed under interventions. By further exploiting interventional data, our approach can learn the structure information of the variables and thus identify the desired causal effects. Specifically, we provide the graphical characterization to efficiently estimate all possible causal effects in a *partially ancestral graph (PAG)* by generalized back-door criterion. The characterization guides learning a local structure with interventional data efficiently. The theoretical analysis and empirical studies validate the effectiveness of the proposed method.

## 1 Introduction

Identifying causal effects has attracted a lot of attention in artificial intelligence. In many decision problems, we usually have one specific decision target, i.e. *response variable*, and many decision variables. For example, a business person wants to maximize the profit. To achieve this, the person has many intervention choices such as lowering price, increasing advertisement investment, improving quality, and so on. A valid approach to guide the decisions is to estimate the causal effect of each variable on the response variable, which is called *target effect identification* in this paper.

To achieve it, an ideal method is to intervene on each decision variable with different attainable values and collect the corresponding interventional data of the response variable. However, when there are many decision variables and many attainable values for each variable, it will take enormous number of interventions, which is economically unfeasible. Considering that there is usually a large amount of observational data in reality, we adopt a more practicable approach that exploits mainly the observational data to achieve target effect identification in this paper.

Given a causal graph, there are many existing classical methods to identify causal effects with observational data (1; 2; 3). In practice, the prior knowledge of structure is usually not available, thus learning the graph at first is necessary. As shown by Verma and Pearl (4), we can only learn a Markov equivalence class with observational data without further assumptions, where there are possibly unknown causal relations resulting in unknown causal effects. To identify these causal effects, we *further* learn the causal relations in the Markov equivalence class by introducing a small amount of *active* interventions, i.e., we force some variable to a known value (5; 6; 7). Following the settings of Wang et al. (8); Triantafillou and Cooper (9), we restrict that *only response variable* is observable under the intervention. In real tasks, it is often hard to collect all variables under intervention. For

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

instance, patients are not fully checked after treatments and only their illness conditions (cured or not) are fed back to doctors (decision-makers). When we aim to achieve target effect identification in this setting, the method of Wang et al. (8) performs well if we assume that there are no latent confounders.

In our paper, we take a further step by allowing the existence of latent confounders. We argue that it is not a direct expansion compared to the method of Wang et al. (8) in techniques. Under causal sufficiency assumption, the adjacent edges of intervened variable $X$ lead to a definite interventional distribution of the response variable $Y$ by back-door criterion (10). Through this mapping we could learn the edges of $X$ by the interventional data of $Y$. When latent confounders exist, however, the mapping does not hold. Besides, *ancestral graph* is used to describe the relation between variables in this condition. In ancestral graphs, directed edges possibly contain latent confounders, which make back-door criterion fail to identify the causal effects. Hence, we also need to consider whether the variables in directed edges are influenced by latent confounders, in addition to learning the graph.

To overcome the difficulties above, we present a graphical characterization that allows us to efficiently find all possible causal effects of the intervened variable on the response variable in a partial ancestral graph. Guided by the characterization, we could learn a local structure and identify the presence of latent confounders behind some directed edges by each intervention, which lead to identifying some causal effects on response variable. After a few active interventions on different variables, our method **ACIC**, short for **AC**tive target effect **I**dentification with latent **C**onfounding, could achieve target effect identification. Due to space limit, we present related work and all the proofs in Appendix.

## 2 Preliminary

We guide readers to Appendix A.1 for the notions including *causal graph* ($G = (\mathbf{V}, \mathbf{E})$), *mixed graph*, *mark*, *arrowhead*, *tail*, *circle*($\circ$), *partial mixed graph*, *parent*, *child*, *spouse*, *possible directed path*, *(possible) ancestor*, $\text{PossAn}(X, G)$, $\text{An}(X, G)$, *(possible) descendant*, $\text{PossDe}(X, G)$, $\text{De}(X, G)$, *Adjacent*, $\text{Adj}(X, G)$, *almost directed cycle*, *collider path*, *minimal path*, *ancestral graph*, *m-separation (m-connecting or active)*, *maximum ancestral graph (MAG, $\mathcal{M}$)*, *discriminating path*, *Markov equivalent*, *Markov equivalence class (MEC)*, *partial ancestral graph (PAG, $\mathcal{P}$)*, *visible*, $\mathcal{M}_{\underline{X}}$, $\mathcal{P}_{\underline{X}}$, *causal effect* ($P(Y|do(X))$), *adjustment criterion*, and the common assumptions including *positivity* and *no selection bias*. We say an MAG $\mathcal{M}$ is *consistent to a PAG $\mathcal{P}$* if it belongs to the MEC represented by $\mathcal{P}$. *$\mathcal{M}$ is Markov equivalent to $\mathcal{P}$* if $\mathcal{M}$ is Markov equivalent to all the MAGs consistent to $\mathcal{P}$. $*$ is used as a wildcard that denotes any of the marks.

Zhang (11) proposed *visibility*, a graphical characterization to imply that there are no latent confounders behind a directed edge in an ancestral graph. In this paper, in addition to by graphical condition, such information can also be learned by interventional data. To describe the presence of latent confounders, we introduce a new notion. For a directed edge $V_i \to V_j$, we say the edge is *pure* if the two variables are not influenced by *latent confounders*, denoted by $V_i \xrightarrow{p} V_j$. We use a *purity matrix* to characterize the purity of each edge in an ancestral graph $G$, in which 1 is in $(i, j)$ entry *only if* the directed edge $V_i \to V_j$ in $G$ is pure, and 0 otherwise. Note that 0 *does not* imply that the edge is impure. It is also possible that we are unaware of whether it is pure or the edge is not directed. Visibility is sufficient but not necessary for purity, which is discussed in detail in Appendix A.2. Thus, given a PAG, we initialize the purity matrix by setting $(i, j)$ entry to 1 if $V_i \to V_j$ is visible and 0 otherwise. For a MAG $\mathcal{M}$, $\mathcal{M}_{\underset{\sim}{X}}$ denotes the graph by removing the directed edges out of $X$ in $\mathcal{M}$ that are labeled to be pure in purity matrix. For a PAG $\mathcal{P}$, see Appendix A.4 for the definition of $\mathcal{P}_{\underset{\sim}{X}}$, which is analogous to Def. 4.2 of Maathuis et al. (12) except for extending visibility to purity.

To identify the causal effects with observational data in an ancestral graph, Maathuis et al. (12) proposed generalized back-door criterion (GBC) and presented the graphical condition about the identifiability of causal effects by GBC. GBC has the advantages such as free of distribution estimation and easy to calculate. Here we make slight modifications to the graphical condition for causal effect identifiability by GBC of (12). The reason is that they considered the setting with only observational data, where only the visible edges can be identified to be pure. When there is interventional data, the purity of some directed edges can be learned further. And such learned information could be utilized in identifying the causal effects by GBC. We show the modified result in Prop. 1. It can be directly proved by Theorem 4.1 of Maathuis et al. (12) with mere slight generalization from visible to pure, which is detailed in Appendix A.4. Before that, we introduce D-SEP$(X, Y, G)$ in Def. 1.

**Definition 1** (D-SEP$(X, Y, G)$(12))**.** Let $X$ and $Y$ be two distinct vertices in a mixed graph $G$. We say that $V \in$ D-SEP$(X, Y, G)$ if $V \neq X$, and there is a collider path between $X$ and $V$ in $G$, such that every vertex on this path (including $V$) is an ancestor of $X$ or $Y$ in $G$.

**Proposition 1.** *Let $G$ be an ancestral graph and $W$ be a purity matrix of $G$. Suppose $X \in$ An$(Y, G)$ and $Y$ are two distinct vertices in $G$. There exists a generalized back-door set*[1] *relative to $(X, Y)$ and $(G, W)$ if and only if* D-SEP$(X, Y, G_{\underset{\sim}{X}}) \cap$ PossDe$(X, G) = \emptyset$. *Moreover, if the set exists,* D-SEP$(X, Y, G_{\underset{\sim}{X}})$ *is such a set. Denote* D-SEP$(X, Y, G_{\underset{\sim}{X}})$ *by* $\mathbf{D}$*, the causal effect is*

$$P(Y|do(X = x)) = \int_{\mathbf{D}} P(\mathbf{D})P(Y|\mathbf{D}, X = x)\,\mathrm{d}\mathbf{D}. \tag{1}$$

There are two possible unidentifiable cases for $P(Y|do(X))$ by GBC in Prop. 1. One is that there are many possible causal effects due to the missing of exact structure information, but we do not know which is correct. In this case, we could address it by learning the structure information with interventional data. The other unidentifiable case is that the causal effect is unidentifiable by GBC *even if we know the MAG and purity of each edge*. We return "Fail" for such $X$ because GBC is not sufficient for identifying the causal effect in this case, and we say *GBC fails to identify* $P(Y|do(X))$.

# 3 The Proposed Approach

In this paper, we assume faithfulness, positivity, and no selection bias. Denote the decision variables by $X_1, \cdots, X_p$ and the response variable by $Y$. Given the observational data of these variables, our goal is to make target effect identification by generalized back-door criterion, i.e., for each variable $X_i, i = 1, 2, \cdots, p$, we identify $P(Y|do(X_i))$ by GBC if GBC does not fail to identify it. With the observational data, we can learn a PAG $\mathcal{P}$ that is consistent to the true MAG by FCI algorithm (13). And we initialize a purity matrix by setting $(i, j)$ entry to 1 if the directed edge between the two variables is visible and 0 otherwise. The causal effects of some variables on $Y$ are possibly unidentifiable in $\mathcal{P}$ by Prop. 1. To identify such effects, we further introduce *active* interventions and observe $Y$ under those. Beginning from $\mathcal{P}$, in each round our method selects one variable $X$ from $X_i, i = 1, \cdots, p$ to intervene and exploits the interventional data of $Y$ to learn the structure information, including revealing circles and updating the purity matrix. The process repeats until identifying all effects $P(Y|do(X_i)), i = 1, 2, \cdots, p$ (identifying failing of GBC is counted). Since it is recursive, we just present the method to learn the structure information by interventional data in one round. The criterion to select the intervention variable in each round is given at the end.

## 3.1 Two direct methods

A naive method to learn the structure by interventional data is enumerating each MAG $\mathcal{M}$ Markov equivalent to $\mathcal{P}$. For each $\mathcal{M}$, $P(Y|do(X))$ can be estimated with observational data by (1) and we judge whether the estimated causal effect is consistent with the interventional data. By such judgment, we could rule out the MAGs with inconsistent causal effects. Yet, this method is usually impractical since it takes a huge computation complexity mainly from two parts. One part is the exhaustive search in the space of MAGs. As is known to all, the space of MAGs is extremely large. For each searched MAG, we also need to judge whether it is Markov equivalent to $\mathcal{P}$ (14; 15). The other costly part is looking for D-SEP$(X, Y, \mathcal{M}_{\underset{\sim}{X}})$ according to Def. 1 to estimate causal effects by (1) for each MAG.

A cleverer approach instead of enumerating is to directly learn a local structure by interventional data, inspired by ACI proposed by Wang et al. (8). ACI tackles a similar task but assumes no latent confounders. The core of this method is to find an equivalent condition that *could be obtained based on a local structure* for causal effect. In this way, there is a bijection between the condition and $P(Y|do(X))$, by which we could learn a local structure with the interventional data of $Y$. Specifically, ACI is comprised of three steps: (a) propose Minimal Parental back-door admissible Set (MPS), a variable set that could be obtained by only the orientation of adjacent edges of $X$, as the equivalent condition for $P(Y|do(X))$; (b) find all possible MPSs in the partial graph to be learned; (c) identify which MPS is correct by interventional data and learn the adjacent edges of $X$ implied by the MPS.

[1] Speaking strictly, the generalized back-door criterion here is slightly different from the original definition in Maathuis et al. (12) due to the introduction of purity. We prevent such discussion in the main paper from distracting readers. See Appendix A.3 and A.4 if it arises concerns.

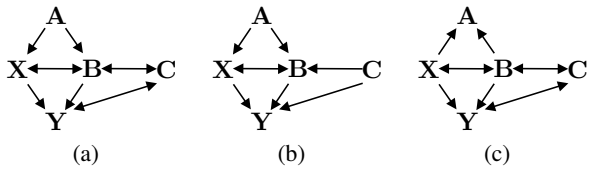


Figure 1: Suppose $X \to Y$ is pure. The adjacent edges of $X$ are the same in (a) and (b), but the causal effects are different due to the different marks at $C$. In addition, although D-SEP$(X, Y, \mathcal{M}_{\underset{\sim}{X}})$ are different in (a) and (c), the causal effects of $X$ on $Y$ are the same by (1), since $A \perp Y \mid X, B$.

However, ACI fails if there are latent confounders. In this situation the adjacent edges of $X$ is not sufficient for identifying causal effects. According to Prop. 1, even though there are the same adjacent edges of $X$ in two MAGs, the causal effects of $X$ on $Y$ are possibly different if they have different D-SEP$(X, Y, \mathcal{M}_{\underset{\sim}{X}})$. See Fig. 1(a) and Fig. 1(b) for example. The method of Malinsky and Spirtes (16) is possibly useful here, which estimates all possible causal effects in $\mathcal{P}$ by enumerating the structures in a relatively local region, but the two parts of computational costs above are still large.

Following the framework of ACI, we propose the three steps for our setting in the next three sections. Firstly, we propose MCS, as an *equivalent condition for* $P(Y|do(X))$ in ancestral graph by GBC in Sec. 3.2. Secondly, we present how to *find all MCSs in* $\mathcal{P}$ *by graphical characterization* in Sec. 3.3. It is the main challenge here in contrast to ACI. ACI achieves it by considering all possible orientations of adjacent edges of $X$ due to the fact that their proposed MPS merely depends on these edges. In our setting, however, it is unclear which edges MCS depends on. And obviously, enumerating all edges is not smart. Hence we propose a graphical characterization to guide it. Lastly, we give the method to *learn marks and purity matrix by interventional data* in Sec. 3.4 with all MCSs that are found before.

### 3.2 The equivalent condition for causal effects

By Prop. 1, D-SEP$(X, Y, \mathcal{M}_{\underset{\sim}{X}})$ can be used to identify $P(Y|do(X))$ when GBC does not fail to identify it. Thus, we construct the equivalent condition for $P(Y|do(X))$ based on D-SEP$(X, Y, \mathcal{M}_{\underset{\sim}{X}})$. Note that D-SEP$(X, Y, \mathcal{M}_{\underset{\sim}{X}})$ cannot be used directly since the causal effects are possibly equal in the MAGs with different D-SEP$(X, Y, \mathcal{M}_{\underset{\sim}{X}})$. See Fig. 1 for an example. The D-SEP$(X, Y, \mathcal{M}_{\underset{\sim}{X}})$ in (a) and (c) are $\{A, B\}$ and $\{B\}$, respectively, but $P(Y|do(X))$ are equal. The reason is $A \perp Y|X, B$ in the two MAGs. When calculating (1), the results are equal after integration on the generalized back-door set. To formalize this situation, we introduce *Minimal Conditional Set (MCS) regarding* D-SEP$(X, Y, \mathcal{M}_{\underset{\sim}{X}})$ *in* $\mathcal{M}$. The algorithm to find MCS and related properties are in Appendix B.

**Definition 2** (Minimal Conditional Set (MCS))**.** Let $\mathcal{M}$ be an MAG. $X$, $Y$, and $\mathbf{D}$ denote the intervened variable, response variable, and the set D-SEP$(X, Y, \mathcal{M}_{\underset{\sim}{X}})$ in $\mathcal{M}$, respectively. $\overline{\mathbf{D}}$ is a subset of $\mathbf{D}$. $\overline{\mathbf{D}}$ is a minimal conditional set regarding $\mathbf{D}$ in $\mathcal{M}$ if

(1) $(Y \perp \mathbf{D} \backslash \overline{\mathbf{D}} \mid \overline{\mathbf{D}} \cup X)_{\mathcal{M}}$,

(2) $(Y \not\perp \mathbf{D} \backslash \overline{\mathbf{D}}' \mid \overline{\mathbf{D}}' \cup X)_{\mathcal{M}}$, for any $\overline{\mathbf{D}}' \subset \overline{\mathbf{D}}$.

We call it MCS regarding D-SEP$(X, Y, \mathcal{M}_{\underset{\sim}{X}})$ or MCS for short. Note that MCS is essentially a set of variable, like the minimal parental back-door set of Wang et al. (8). For all MAGs $\mathcal{M}$ satisfying the conditions of Prop. 1 with the same $\overline{\mathbf{D}}$, where $\overline{\mathbf{D}}$ denotes the MCS regarding D-SEP$(X, Y, \mathcal{M}_{\underset{\sim}{X}})$ in respective $\mathcal{M}$, they have the same causal effect as

$$P(Y|do(X = x)) = \int_{\overline{\mathbf{D}}} P(\overline{\mathbf{D}}) P(Y|\overline{\mathbf{D}}, X = x) \, \mathrm{d}\overline{\mathbf{D}}. \tag{2}$$

Since structure inference by interventional data is involved, we make an additional assumption to match interventional data to causal structure, playing as the role of faithfulness given only observational data or of intervention-related assumptions(17; 8; 18; 9). See Asm. 1. The purpose is to avoid that different causal effects in general *happen to* be equal in some *special* distributions. Combining Asm. 1 with the discussions before, there is a bijection between MCS and $P(Y|do(X))$. In the next section, we present how to find all possible MCSs by our proposed graphical characterization.

**Assumption 1** (Generalized interventional-faithfulness)**.** *For two Markov equivalent MAGs $\mathcal{M}^1$ and $\mathcal{M}^2$ with the same observational distribution, if there are different minimal conditional sets regarding* D-SEP$(X, Y, \mathcal{M}^1_{\underset{\sim}{X}})$ *and* D-SEP$(X, Y, \mathcal{M}^2_{\underset{\sim}{X}})$ *in the respective graphs, then $P\big(Y|do(X = x)\big)$ are different in the two MAGs.*

### 3.3 Finding all MCSs in $\mathcal{P}$ by graphical characterization

At the beginning, we provide an outline of finding possible MCSs regarding D-SEP$(X, Y, \mathcal{M}_{\underset{\sim}{X}})$ in all $\mathcal{M}$ consistent to $\mathcal{P}$ when GBC dose not fail to identify $P(Y|do(X))$ in $\mathcal{M}$. The justification for the restriction on $\mathcal{M}$ is that if $\mathcal{M}$ is the true MAG and GBC fails to identify $P(Y|do(X))$ in $\mathcal{M}$, there is no MCS matched to the interventional data. We thus cannot learn a MCS by the intervention. In this case it is unnecessary to find MCS in such $\mathcal{M}$. Hence, we consider the MAG $\mathcal{M}$ that satisfies D-SEP$(X, Y, \mathcal{M}_{\underset{\sim}{X}}) \cap$ De$(X, \mathcal{M}) = \emptyset$, according to Prop. 1. There are three main parts to achieve it. The first part is to enumerate all possible mark combinations of $X$. The true MAG is with one mark combination among them. As mentioned earlier, the marks of $X$ are not sufficient for identifying $P(Y|do(X))$. In another word, MCSs could be different in the MAGs with the same marks at $X$. Hence, in the second part we propose the graphical condition to indicate when there are different MCSs in the MAGs with the same marks at $X$. And this graphical condition can be obtained in a partial graph. Based on that, we present the algorithm to find all possible MCSs in $\mathcal{P}$ in the third part.

We first enumerate all possible mark combinations of $X$. To denote the PAGs with deterministic marks at $X$, we introduce *local MAG of $X$* in Def. 3. Then based on local MAG, we present a sufficient and necessary condition for $V \in$ D-SEP$(X, Y, \mathcal{M}_{\underset{\sim}{X}})$ in Thm. 2, which plays a role in cushion for the main result involving MCS regarding D-SEP$(X, Y, \mathcal{M}_{\underset{\sim}{X}})$ later.

**Definition 3.** Given a PAG $\mathcal{P}$ and an intervention variable $X$, an ancestral graph $M$ is a *local MAG of $X$ based on $\mathcal{P}$* if $(1)$ $M$ is with deterministic marks (arrowheads or tails) at $X$; $(2)$ $M$ is obtained from $\mathcal{P}$ by merely marking all the circles at $X$ without generating new unshielded colliders. We call *local MAG* for short and denote it by $M$, which is different from calligraphic $\mathcal{M}$ that denotes MAG.

**Theorem 2.** *Let $M$ be a local MAG and $\mathcal{M}$ be an MAG consistent to $M$. Suppose $V$ ($V \neq X, Y$) is a variable in $\mathcal{M}$. If* D-SEP$(X, Y, \mathcal{M}_{\underset{\sim}{X}}) \cap$ De$(X, \mathcal{M}) = \emptyset$*, then $V \in$* D-SEP$(X, Y, \mathcal{M}_{\underset{\sim}{X}})$ *holds if and only if there is at least one collider path from $X$ to $V$ starting by an arrowhead at $X$ in $M$ such that each variable except for $X$ on the path is an ancestor of $X$ or $Y$ in $\mathcal{M}$.*

The advantage of considering local MAG of $X$ rather than PAG with circles at $X$ is that based on local MAG we could obtain a set of variables, PD-SEP$(X, Y, M)$[2], as defined in Def. 4. It implies all the variables possibly belonging to D-SEP$(X, Y, \mathcal{M}_{\underset{\sim}{X}})$, i.e. D-SEP$(X, Y, \mathcal{M}_{\underset{\sim}{X}}) \subseteq$ PD-SEP$(X, Y, M)$, where $\mathcal{M}$ is any MAG consistent to $M$ such that D-SEP$(X, Y, \mathcal{M}_{\underset{\sim}{X}}) \cap$ De$(X, \mathcal{M}) = \emptyset$. More importantly, this set has a good property that if PD-SEP$(X, Y, M) \backslash$D-SEP$(X, Y, \mathcal{M}_{\underset{\sim}{X}}) \neq \emptyset$, there must be a variable $V \in$ PD-SEP$(X, Y, M)$ that is not an ancestor of $X$ or $Y$ in $\mathcal{M}$, as shown by the combination of Thm. 2 and Def. 4. This property plays an important role for the main result later.

**Definition 4** (PD-SEP$(X, Y, M)$)**.** Let $M$ be a local MAG of $X$. Variable $V \in$ PD-SEP$(X, Y, M)$ if and only if there is at least one collider path between $X$ and $V$ in $M$, and each variable except for $X$ on the path is a possible ancestor of $X$ or $Y$ in $M$ and not a descendant of $X$ in $\mathcal{M}$.

*Remark.* In the literature, there is a well-defined notion Possible-D-SEP$(X, Y)$ (13; 19), which is introduced with the similar intention that indicates which variables possibly belong to D-SEP$(X, Y, \mathcal{M})$. Since ours is pretty different in both required conditions and the restriction on ancestral relations, we use a new name PD-SEP$(X, Y, M)$ to distinguish them.

Combined with the previous discussions, we present our main result in Thm. 3. Perhaps surprisingly, it implies that although there are possibly different D-SEP$(X, Y, \mathcal{M}_{\underset{\sim}{X}})$ in distinct MAG $\mathcal{M}$ consistent to $M$, the MCSs regarding them are the same as long as there are no *critical variables for* $(X, Y)$ in $M$. We define critical variables for $(X, Y)$ in Def. 5, and show an example in Fig. 2, where $F_t$ is a critical variable and the circles colored by red are critical marks.

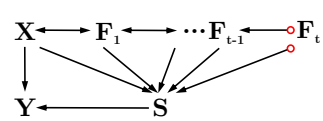

Figure 2: An illustration for critical variable $F_t$ and critical marks (red).

[2]Note that PD-SEP$(X, Y, M)$ can be obtained by $M$, since all the descendants of $X$ in $\mathcal{M}$ consistent to $M$ are knowable in $M$, which is detailed by Lemma 18 in Appendix.

**Algorithm 1** Find all possible MCSs in $\mathcal{P}$

**input:** Intervention variable $X$, PAG $\mathcal{P}$

1: $L \leftarrow \emptyset$ // It is to record each local MAG $M^j$ and corresponding MCS$^j$
2: **for** each local MAG $M^j$ of $X$ based on $\mathcal{P}$ **do**
3: **if** there is critical variable set $\mathbf{C}^j$ for $(X, Y)$ in $M^j$ **then**
4: CRITICAL($M^j, \mathbf{C}^j, \mathbf{S}^j$) // $\mathbf{S}^j$ ($\mathbf{S}^{j_k}$ below) is the set $\mathbf{S}$ defined in Def. 5 in $M^j$ ($M^{j_k}$)
5: **else**
6: $L = L \cup (M^j, \text{MCS}^j)$
7: **end if**
8: **end for**
9: **function** CRITICAL($M^j, \mathbf{C}^j, \mathbf{S}^j$)
10: **for** each element $\mathbf{C}^{j_k}$ in the power set of $\mathbf{C}^j$ **do**
11: Obtain a new local MAG $M^{j_k}$ by orienting $F \to S$ for $\forall F \in \mathbf{C}^{j_k}, \forall S \in \mathbf{S}^j$ and marking the critical marks of $F \in \mathbf{C}^j \backslash \mathbf{C}^{j_k}$ as arrowheads
12: **if** there is critical variable set $\mathbf{C}^j$ for $(X, Y)$ in $M^{j_k}$ **then**
13: CRITICAL($M^{j_k}, \mathbf{C}^{j_k}, \mathbf{S}^{j_k}$)
14: **else**
15: $L = L \cup (M^{j_k}, \text{MCS}^{j_k})$
16: **end if**
17: **end for**
18: **end function**

**output:** $L$

**Definition 5** (Critical variable for $(X, Y)$)**.** In a local MAG $M = (\mathbf{V}, \mathbf{E})$ with a path $X \leftrightarrow F_1 \leftrightarrow \cdots \leftrightarrow F_{t-1} \leftrightarrow F_t$ or $X \leftrightarrow F_1 \leftrightarrow \cdots \leftrightarrow F_{t-1} \leftarrow\!\circ F_t$, $t \geq 1$, $F_t$ is called a *critical variable* for $(X, Y)$ if there is a non-empty variable set $\mathbf{S}$ relative to $F_t$ defined as follows: $S \in \mathbf{S}$ if and only if in $M$ (1) $S \in \mathbf{V} \backslash \{X\}$, (2) $S$ is a child of $X, F_1, \cdots, F_{t-1}$, (3) $S$ is not a child of $F_t$, (4) $S$ is at one minimal possible directed path from $F_t$ to $Y$, and no variable in the path except for the endpoints is in PD-SEP$(X, Y, M)$. Each circle at $F_t$ in the edge with $F_{t-1}$ or $\mathbf{S}$ is called a *critical mark* of $F_t$.

**Theorem 3.** *Let $M$ be a local MAG of $X$ based on a PAG $\mathcal{P}$. Then condition (1) below is sufficient for condition (2):*

*(1) there exists an MAG $\mathcal{M}$ consistent to $M$ satisfying (a) D-SEP$(X, Y, \mathcal{M}_{\underset{\sim}{X}}) \cap \text{De}(X, \mathcal{M}) = \emptyset$, (b) $X \in \text{An}(Y, \mathcal{M})$, such that $\mathbf{A}_{\mathcal{M}} \neq \mathbf{A}'_{\mathcal{M}}$, where $\mathbf{A}_{\mathcal{M}}$ denotes the MCS regarding D-SEP$(X, Y, \mathcal{M}_{\underset{\sim}{X}})$ in $\mathcal{M}$ and $\mathbf{A}'_{\mathcal{M}}$ denotes the MCS regarding PD-SEP$(X, Y, M)$ in $\mathcal{M}$;*

*(2) there exists at least one critical variable for $(X, Y)$ in $M$.*

*Remark.* It is noteworthy that all the MAGs consistent to $M$ are Markov equivalent. Hence the MCSs regarding PD-SEP$(X, Y, M)$ in these MAGs are the same, namely, $\mathbf{A}'_{\mathcal{M}}$ is invariant for different $\mathcal{M}$. If there is no critical variable as condition (2), then for any MAG $\mathcal{M}$ consistent to $M$, $\mathbf{A}_{\mathcal{M}} = \mathbf{A}'_{\mathcal{M}}$ holds. In such case we conclude $\mathbf{A}_{\mathcal{M}}$ is invariant for different $\mathcal{M}$ consistent to $M$, i.e., different D-SEP$(X, Y, \mathcal{M}_{\underset{\sim}{X}})$ in distinct $\mathcal{M}$ consistent to $M$ share the same MCS regarding them.

Def. 5 and Thm. 3 form a graphical characterization to obtain all possible MCSs in MAGs consistent to a local MAG $M$, based on which we propose Alg. 1 to find all MCSs in $\mathcal{P}$. Given a PAG $\mathcal{P}$, it is easy to find all local MAGs $M$ based on $\mathcal{P}$ by marking the circles at $X$ without generating new unshielded colliders (Line 2). For each $M$, if there is no critical variable in $M$, the MCS regarding D-SEP$(X, Y, \mathcal{M}_{\underset{\sim}{X}})$ is unique, and it equals to the MCS regarding PD-SEP$(X, Y, M)$ (Line 6). When there is a non-empty critical variable set $\mathbf{C}$, the MCSs in the MAGs consistent to $M$ could be different. The reason is that the variable in $\mathbf{C}$ could be the ancestor of $X$ or $Y$ in some MAGs but the non-ancestor in other MAGs, due to the unidentified critical marks. Conditioning on the non-ancestor variables above makes some generalized back-door paths from $X$ to $Y$ active, thus these variables cannot appear in MCS. Hence, we further discuss the critical marks. For any subset $\mathbf{C}_k$ of the critical variable set $\mathbf{C}$, there could possibly be some MAG $\mathcal{M}$ consistent to $M$ where $\mathbf{C}_k \subseteq$ D-SEP$(X, Y, \mathcal{M}_{\underset{\sim}{X}})$ and $(\mathbf{C} \backslash \mathbf{C}_k) \cap$ D-SEP$(X, Y, \mathcal{M}_{\underset{\sim}{X}}) = \emptyset$. We use a new local MAG to represent the mentioned MAGs (Line 10,11). And in the new local MAG, we further consider whether

critical variable exists (Line 12). We present Lemma 4 to guarantee the soundness of Alg. 1, i.e. no matter what the true MAG $\mathcal{M}$ is, the MCS regarding D-SEP$(X, Y, \mathcal{M}_{\underset{\sim}{X}})$ in $\mathcal{M}$ could be returned.

**Lemma 4.** *Let $\mathcal{P}$ be a PAG of MAG $\mathcal{M}$. If $X \in \text{An}(Y, \mathcal{M})$ and* D-SEP$(X, Y, \mathcal{M}_{\underset{\sim}{X}}) \cap \text{De}(X, \mathcal{M}) = \emptyset$*, then the MCS regarding* D-SEP$(X, Y, \mathcal{M}_{\underset{\sim}{X}})$ *in $\mathcal{M}$ and the corresponding local MAG are contained in the output of Algorithm 1.*

### 3.4 Learning marks and purity matrix by interventional data

In this section we propose how to learn marks and purity matrix by the interventional data of $Y$. After intervening on $X$, there are three possible situations implied by the interventional data of $Y$.

(1) $X$ has no causal effect on $Y$, i.e., $P(Y|do(X)) = P(Y)$,

(2) $X$ has causal effect on $Y$ and it can be identified by GBC, i.e., there exists a generalized back-door set $\mathbf{D}$ such that $P(Y|do(X)) = \int P(\mathbf{D})P(Y|X, \mathbf{D})\,\mathrm{d}\mathbf{D}$,

(3) $X$ has causal effect on $Y$ but GBC fails to identify it.

For situation $(1)$, it is trivial to learn the marks by the interventional data according to Prop. 5.

**Proposition 5.** *If $P(Y|do(X)) = P(Y)$, the marks at $X$ are arrowheads in all the minimal possible directed paths from $X$ to $Y$ in a partially ancestral graph.*

If $P(Y|do(X))$ is not equal to $P(Y)$, it implies that the interventional data accords with situation $(2)$ or situation $(3)$ but is uncertain to us. Hence, we need to find all possible MCSs based on $\mathcal{P}$, then judge whether one of the estimated causal effects with the observational data by (2) is consistent to the interventional data. If so, we can see that the causal effect of $X$ on $Y$ can be identified by GBC in the true MAG. Specifically, we first obtain a group of pairs $(M^j, \text{MCS}^j)$ by Alg. 1, where $M^j$ is a local MAG and MCS$^j$ is the corresponding MCS. Let $\hat{P}_{M^j}\big(Y|do(X = x)\big)$ and $\hat{P}\big(Y|do(X = x)\big)$ denote the estimated causal effect in each $M^j$ by MCS$^j$ and that under real intervention respectively. *Disc*$(P, Q)$ is the distribution discrepancy between $P$ and $Q$. If there are $m$ intervention samples $\big((do(X = x_1), Y_1\big), \cdots, \big(do(X = x_s), Y_s\big), \cdots, \big(do(X = x_m), Y_m)\big)$, we take MCS$^\star$ by

$$\text{MCS}^\star = \underset{\text{MCS}^j}{\arg\min} \sum_{s=1}^{m} \textit{Disc}\Big(\hat{P}_{\text{MCS}^j}\big(Y|do(X = x_s)\big), \hat{P}\big(Y|do(X = x_s)\big)\Big). \tag{3}$$

Any distance metric can be used here. Since the attention is not on calculation, we take the expectation difference as the metric for convenience, i.e., MCS$^\star = \arg\min_{\text{MCS}^j} \sum_{s=1}^{m} |\hat{\mathbb{E}}_{\text{MCS}^j}\big(Y|do(X = x_s)\big) - Y_s|$. If the distance between the estimated causal effect based on MCS$^\star$ and interventional data is larger than a given threshold, we consider that GBC fails to identify such causal effect, we thus mark such $X$ as "FAIL" and orient some edges by Prop. 6. Otherwise, if there is only one local MAG $M^*$ corresponding to MCS$^\star$, we orient $\mathcal{P}$ by $M^*$. If there are more than one local MAG with MCS$^\star$, we orient $\mathcal{P}$ by the common marks of the local MAGs with MCS$^\star$. Besides, when GBC does not fail to identify the causal effect, all the edges out of $X$ in the minimal directed paths from $X$ to $Y$ in the ancestral graph are learned to be pure. Thus we could update the purity matrix accordingly.

**Proposition 6.** *In situation $(3)$, let $\mathbf{T}$ denote all variables adjacent to $X$ in the minimal possible directed paths from $X$ to $Y$. For $T \in \mathbf{T}$, if for $\forall V \in \mathbf{T} \backslash T$, it holds either $T \notin \text{Adj}(V, \mathcal{P})$ or there is a variable $S \notin \text{Adj}(V, \mathcal{P})$ such that there is a collider path $X \circ\!\!\rightarrow T \leftrightarrow \cdots \leftarrow\!\!* S$ and every vertex except $S$ on the path is a parent of $V$, then $X \rightarrow T$.*

After learning with interventional data, we can further update structure information according to the property of MAG. Zhang (20) proposed ten complete rules when there is only observational data. However, how to completely orient with the learned knowledge by interventional data is still an open problem. We add an additional rule referring to the known results for CPDAG (21). But regretfully, whether the eleven rules are complete is unknown. With the updated graph, some directed edges are newly identified to be visible, thus to be pure. Hence we also update the purity matrix accordingly.

Rule 11: If $a \circ\!\!\rightarrow b \rightarrow c$, $a, b, c \in \text{Adj}(d)$, $a \notin \text{Adj}(c)$, and $a, d, c$ do not form an unshielded collider, then $d \rightarrow c$.

**Proposition 7.** *Rule 11 is sound.*

**Algorithm 2** ACIC (ACtive target effect Identification with latent Confounding)

**input:** PAG $\mathcal{P}$ by FCI algorithm;

1: Initialize $\mathcal{I}_1 = \{V | V \in \mathbf{V}(\mathcal{P}), V \in \text{PossAn}(Y, \mathcal{P})$ and $P(Y|do(V))$ is unidentifiable by Prop. 1 in $\mathcal{P}\}$ // Record the variables whose causal effects on $Y$ are unidentifiable by Prop. 1

2: Initialize $\mathcal{I}_2 = \emptyset$ // Record the variables whose causal effects are failed to be identified by GBC

3: **while** $\mathcal{I}_1 \backslash \mathcal{I}_2 \neq \emptyset$ **do**

4: Select the variable $X$ with the maximum number of circles from $\mathcal{I}_1 \backslash \mathcal{I}_2$ to intervene

5: Learn structure information according to Section 3.3 and Section 3.4

6: Update $\mathcal{I}_1$ and $\mathcal{I}_2$ by Prop. 1 with the learned knowledge including structure and purity

7: **end while**

**output:** The estimated causal effect of each variable on $Y$.

Combining all the parts above, we present the whole process in Alg. 2, while a detailed version is in Appendix F. In a graph obtained by FCI or learned after each intervention, we judge which variables are possible ancestors of $Y$ by Prop 8. Among them if there are variables with unidentified causal effects on $Y$ (Line 1) and we are not sure whether GBC fails to identify the effects (Line 2), we select one variable from them to intervene (Line 4) and learn the structure information by interventional data (Line 5). Since intervention is expensive in reality, we hope achieving target effect identification with fewer intervention times. Hence, among the variables mentioned above, we greedily intervene on the one with the maximum number of circles since we could possibly learn all the marks at $X$. To help the readers understand the procedure of the method, we give a running example in Appendix G.

**Proposition 8.** *In a PAG $\mathcal{P}$, if there is no minimal possible directed path from $X$ to $Y$, then $X$ cannot be ancestor of $Y$ in any MAG consistent to $\mathcal{P}$. And it holds that $X \notin \text{PossAn}(Y, \mathcal{P})$.*

# 4 Theoretical Results

In this section, we first prove the identifiability of causal effects. Then we provide an analysis about the computation complexity of estimating all possible causal effects of each variable $X$ on $Y$.

**Theorem 9.** *Given a PAG $\mathcal{P}$ and the observational distribution of the variables, the causal effect of $X$ on $Y$ is identifiable by GBC with the interventional data of $Y$ under intervention on $X$ if GBC does not fail to identify the causal effect $P(Y|do(X))$.*

Thm. 9 implies causal effect identifiability of each $X$ on $Y$ by interventions. In practice, when we intervene on $X$, in addition to identifying $P(Y|do(X))$, our method could learn some marks and the purity of some edges, which lead to the causal effect identification of *other variables* on $Y$. Hence we usually make target effect identification by far fewer intervention times than the variable number.

Estimating all possible causal effects takes the main computational cost. The process of estimating possible effects in a PAG comprises finding all MCSs in Sec. 3.3 and using (2). We analyze the complexity of Sec. 3.3. Since the complexity is strongly related to the graph, it is hard to analyze for a general graph due to the randomness of the skeleton. We consider a special case - the complete graph, which is often considered as the most difficult graph to learn, because it has the most edges and we can learn no marks by only observational data as a result of no conditional independent relationship between the variables. To ensure that the interventional data accords with situation (2) or (3) when finding all MCSs in the PAG is necessary, we set $Y$ as the descendant of all of $X_1, \cdots, X_p$. Due to space limit, we only present a brief analysis here, while a detailed version is provided in Appendix E.

**Proposition 10.** *Let $\mathcal{M}$ be a complete MAG with $p+1$ variables $X_1, \cdots, X_p, Y$, where the causal order of the variables except $Y$ is completely random and $Y$ is at the last. Denote the graph obtained by FCI with observational data by $\mathcal{P}$ and intervention variable by $X_i$. And let $M$ be a local MAG of $X_i$ with $p-1-k$ tails and $k$ arrowheads at $X_i$. The computational complexity of finding all possible causal effects $P(Y|do(X_i))$ in all the MAGs consistent to $M$ is $O(2^k)$. Further, the computational complexity of finding all causal effects $P(Y|do(X_i))$ in all the MAGs consistent to $\mathcal{P}$ is $O(3^p)$.*

Let $\mathbf{S}$ denote the set of variable that has an edge with an arrowhead at $X_i$. Since the graph is complete, for any subset $\mathbf{S}_1$ of $\mathbf{S}$, we could construct an MAG $\mathcal{M}$ based on $M$ such that MCS in $\mathcal{M}$ is $\mathbf{S}_1$, which is detailed in Appendix E.2. Hence there are $2^k$ causal effects of $X_i$ on $Y$ in the MAGs consistent to $M$. Our method thus achieves the minimum complexity in finding possible causal effects in $M$.

However, when estimating causal effects of $X_i$ on $Y$ in $\mathcal{P}$, the complexity of our method is $O(3^p)$, while there are $2^{p-1}$ causal effects in $\mathcal{P}$. The gap is from two aspects. One is the search for critical variables. The other is that the causal effects in MAGs from different local MAGs are possibly equal, while these causal effects are considered separately based on each local MAG. The complexity of first part is less than that of the latter part. While if we do not consider the graphical characterization based on critical variable and adopt the local method of Malinsky and Spirtes (16) in estimating causal effects that enumerates MAGs in a subregion, the complexity is exponentially larger than ours, as shown in Table 1.

| | local MAG $M$ | PAG $\mathcal{P}$ |
|---|---|---|
| A lower bound | $O(2^k)$ | $O(2^p)$ |
| Algorithm 1 | $O(2^k)$ | $O(3^p)$ |
| Malinsky and Spirtes (16) | None | $O(3^{p^2})$ |

Table 1: Complexity of estimating all effects.

## 5 Experiments

In this section, we apply our method on synthetic dataset to validate the effectiveness and efficiency of the proposed method. The code is developed based on R package "pcalg" (22).

We generate 100 random causal graphs and evaluate the number of correctly identified causal effects. In each graph, there are $p = 15$ variables and an edge occurs between two variables with probability 0.3. We randomly take 3 variables as latent confounders. And the last observed variable in the causal order is set to the response variable. We generate linear Gaussian data[3] according to the causal graph. Based on these information, we know not only the true MAG, but also whether each edge is pure. Hence we have the ground truth causal effect of each variable on the response variable by GBC. Beginning from the PAG obtained by the true MAG, we record the number of correctly identified causal effects under different intervention times with different methods.

Our method is compared to three baseline methods, *ACIC-random*, *Do* and *ACI* (8). In ACIC-random, we randomly select the intervention variable whose causal effect on $Y$ has not been identified. For Do, it identifies causal relations by judging whether the distribution of observed variable take a change under intervention, which idea is applied widely in many active causal discovery methods (5; 6). Since in these methods all the variables could be observed under intervention, we allow Do to select which variable $X_j$ to observe under intervention instead of observing $Y$. The results are shown in Figure 3. The superiority of ACIC to ACIC-random implies the greedy strategy to intervene makes sense in saving intervention times. The phenomenon that ACIC is more efficient than Do verifies that exploiting causal effect elaborately could take us more message about the structure. There are many effects wrongly identified by ACI, which indicates the risk of ignoring latent confounders.

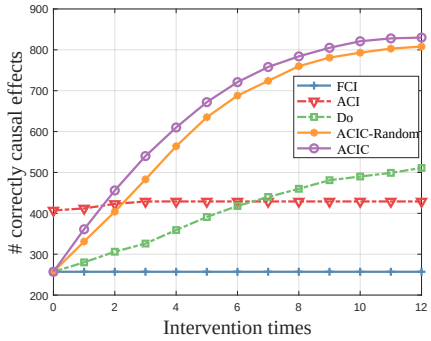


Figure 3: The number of correctly identified causal effects by different methods.

## 6 Conclusion

In this paper, we propose a method to actively identify the causal effect of each variable on the response variable with the existence of latent confounders. We present the graphical characterization to find all possible causal effects in a PAG. Guided by the graphical characterization, we learn structure information with interventional data of the response variable. Our method can make target effect identification effectively and efficiently in both saving intervention times and reducing computation complexity. And we verify them theoretically and empirically.

There are two main aspects to improve the method. One is on how to completely update a PAG with background knowledge. The other is on how to reduce the computational complexity further. We look forward to the future work that could address these problems.

[3]When variables are continuous, the positivity assumption tends to be violated. Here we ignore such risks since the identifiability has been proven and we just want to test whether the proposed method could accurately learn the structure and identify the generalized back-door set.

# References


[1] Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Proceedings of the 18th National Conference on Artificial Intelligence and 14th Conference on Innovative Applications of Artificial Intelligence*, pages 567–573, 2002.

[2] Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings of the 21st National Conference on Artificial Intelligence and the 18th Innovative Applications of Artificial Intelligence Conference*, pages 1219–1226, 2006.

[3] Yonghan Jung, Jin Tian, and Elias Bareinboim. Learning causal effects via weighted empirical risk minimization. In *Advances in Neural Information Processing Systems*, 2020.

[4] Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, pages 255–270, 1990.

[5] Yang-Bo He and Zhi Geng. Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research*, 9:2523–2547, 2008.

[6] Alain Hauser and Peter Bühlmann. Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, 55(4):926–939, 2014.

[7] Juan Gamella and Christina Heinze-Deml. Active invariant causal prediction: Experiment selection through stability. In *Advances in Neural Information Processing Systems*, 2020.

[8] Tian-Zuo Wang, Xi-Zhu Wu, Sheng-Jun Huang, and Zhi-Hua Zhou. Cost-effectively identifying causal effects when only response variable is observable. In *Proceedings of the 37th International Conference on Machine Learning*, pages 10060–10069, 2020.

[9] Sofia Triantafillou and Gregory Cooper. Learning adjustment sets from observational and limited experimental data. *arXiv preprint arXiv:2005.08749*, 2020.

[10] Marloes H Maathuis, Markus Kalisch, Peter Bühlmann, et al. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164, 2009.

[11] Jiji Zhang. Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 9:1437–1474, 2008.

[12] Marloes H Maathuis, Diego Colombo, et al. A generalized back-door criterion. *The Annals of Statistics*, 43(3):1060–1088, 2015.

[13] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT Press, 2000.

[14] R Ayesha Ali, Thomas S Richardson, Peter Spirtes, et al. Markov equivalence for ancestral graphs. *The Annals of Statistics*, 37(5B):2808–2837, 2009.

[15] Zhongyi Hu and Robin Evans. Faster algorithms for markov equivalence. In *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 739–748, 2020.

[16] Daniel Malinsky and Peter Spirtes. Estimating causal effects with ancestral graph markov models. In *Conference on Probabilistic Graphical Models*, pages 299–309, 2016.

[17] Biwei Huang, Kun Zhang, Jiji Zhang, Joseph D. Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21:89:1–89:53, 2020.

[18] Amin Jaber, Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Causal discovery from soft interventions with unknown targets: Characterization and learning. In *Advances in Neural Information Processing Systems*, 2020.

[19] Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pages 294–321, 2012.

[20] Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008.

[21] Christopher Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the 11st Annual Conference on Uncertainty in Artificial Intelligence*, pages 403–410, 1995.

[22] Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H. Maathuis, and Peter Bühlmann. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26, 2012.

# Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See the conclusion.
   (c) Did you discuss any potential negative societal impacts of your work? [No] The techniques in this paper seem that have less societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] The assumptions are shown at the beginning of "the proposed Approach" and Assumption 1.
   (b) Did you include complete proofs of all theoretical results? [Yes] See the proofs in Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] We will release the code if it is accepted.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See the Section "Experiments".
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Our method is not sensitive to the random seed.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See the Section "Experiments"
   (b) Did you mention the license of the assets? [No] Although we did not mention it, the code we use is allowed for the academic use by the authors.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No] We will include the code if the paper is accepted.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]