Learning Shadow Variable Representation for Treatment Effect Estimation under Collider Bias

Baohong Li [1] **Haoxuan Li** [2] **Ruoxuan Xiong** [3] **Anpeng Wu** [1] **Fei Wu** [1] **Kun Kuang** [1]

Abstract

One of the significant challenges in treatment effect estimation is collider bias, a specific form of sample selection bias induced by the common causes of both the treatment and outcome. Identifying treatment effects under collider bias requires well-defined shadow variables in observational data, which are assumed to be related to the outcome and independent of the sample selection mechanism, conditional on the other observed variables. However, finding a valid shadow variable is not an easy task in real-world scenarios and requires domain-specific knowledge from experts. Therefore, in this paper, we propose a novel method that can automatically learn shadow-variable representations from observational data without prior knowledge. To ensure the learned representations satisfy the assumptions of the shadow variable, we introduce a tester to perform hypothesis testing in the representation learning process. We iteratively generate representations and test whether they satisfy the shadow-variable assumptions until they pass the test. With the help of the learned shadow-variable representations, we propose a novel treatment effect estimator to address collider bias. Experiments show that the proposed methods outperform existing treatment effect estimation methods under collider bias and prove their potential application value.

1. Introduction

Causal inference is a powerful statistical modeling tool for explanatory analysis (Wang et al., 2022; Zhang et al., 2023; 2024), and a central problem in causal inference is treatment effects estimation. The gold standard approach for treatment effect estimation is to conduct Randomized Controlled Trials (RCTs), but RCTs can be expensive (Kohavi & Longbotham, 2011) and sometimes infeasible (Bottou et al., 2013). Therefore, developing practical approaches to estimate treatment effects from observational data is crucial for causal inference.

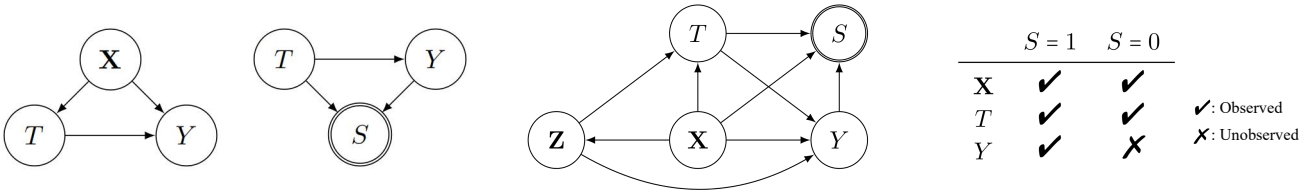
In observational studies, association does not imply causation, mainly due to the presence of spurious associations in the data. There are two primary sources of spurious associations: confounding bias and collider bias (Hernán & Robins, 2020). Most of the previous works focused on confounding bias that results from common causes of treatments and outcomes (Bang & Robins, 2005; Louizos et al., 2017; Shalit et al., 2017; Wager & Athey, 2018) while ignored collider bias which comes from non-random sample selection caused by both treatments and outcomes.

We use causal diagrams in Figure 1 to further illustrate the two biases, where $\mathbf{X}$ denotes the observed covariates, T denotes the treatment variable, Y denotes the outcome variable, and S denotes the sample selection indicator. Confounding bias results from common causes of treatment and outcome (Greenland, 2003; Guo et al., 2020). As shown in Figure 1(a), there are two sources of association between T and Y: the path $T \to Y$ that represents the treatment effect of T on Y, and the path $T \leftarrow \mathbf{X} \to Y$ between T and Y that includes the common cause $\mathbf{X}$, which introduces spurious associations into the observational data. Collider bias is a particular case of sample selection bias[1] that results from conditioning on a common effect of T and Y (Hernán & Robins, 2020). As shown in Figure 1(b), except for the path $T \to Y$, the other source of association between T and Y is from the open path $T \to S \leftarrow Y$. It links T and Y through their conditioned on common effect S, which introduces spurious associations. As shown in Figure 1(d), an analysis conditioned on S will cause collider bias, i.e., we can only observe the outcome of those selected units ($S = 1$), and the values of Y are missing for those unselected units ($S = 0$),

[1]College of Computer Science and Technology, Zhejiang University, Hangzhou, China [2]Center for Data Science, Peking University, Beijing, China [3]Department of Quantitative Theory & Methods, Emory University, Atlanta, USA. Correspondence to: Kun Kuang <kunkuang@zju.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

[1]Sample selection bias arises from non-random sample selection conditioned on S caused by certain variables in data, while collider bias is the particular case that T and Y both cause S.



(a) Confounding bias. (b) Collider bias. (c) An illustrative case of shadow variables. (d) The data form of collider bias.

Figure 1. Different kinds of biases represented by causal diagrams.

leading to incorrect treatment effect estimation.

Previous studies show that treatment effects are unidentifiable under collider bias without further assumptions or prior knowledge. Fortunately, if some shadow variables are available in the observational data, identifying treatment effects is still possible from observational data (Miao & Tchetgen Tchetgen, 2016). As shown in Figure 1(c), shadow variables $\mathbf{Z}$ are assumed to be fully observed variables independent of the sample selection mechanism after conditioning on the outcome and other covariates, i.e., a valid shadow variable needs to simultaneously satisfy that $\mathbf{Z} \not\perp\!\!\!\perp Y \mid \mathbf{X}, T, S = 1$ and $\mathbf{Z} \perp\!\!\!\perp S \mid \mathbf{X}, T, Y$. For example, when studying the effect of students' mental health (T) on teachers' assessment (Y), collider bias occurs since teachers might not be willing to report their assessment of students with poor mental health. The teacher's response rate (S) may be related to their assessment of the student but is unlikely to be related to a separate parent's report after conditioning on the teacher's assessment and fully observed covariates; moreover, the parent's report ($\mathbf{Z}$) is likely highly correlated with the teacher's. In this case, the parental assessment can be considered a shadow variable (Ibrahim et al., 2001). With the help of shadow variables, treatment effects can be identified and estimated (d'Haultfoeuille, 2010; Wang et al., 2014; Miao & Tchetgen Tchetgen, 2016).

However, finding a well-defined shadow variable requires domain-specific knowledge from experts and needs to be investigated on a case-by-case basis (Li et al., 2023), which is also a hard task. Therefore, we propose a novel method named **ShadowCatcher** that automatically generates representations from the observed covariates satisfying the assumptions of shadow variables, which can serve the role of shadow variables when estimating treatment effects and thus achieve the goal of solving collider bias without introducing more prior knowledge. Specifically, we iteratively generate shadow-variable representations by conditional independence constraints and test whether the generated representations satisfy the assumptions until the generated representations can pass the hypothesis test. Furthermore, we also propose a novel **ShadowEstimator** to estimate treatment effects under collider bias by leveraging the generated shadow variables representations. We conduct experiments on synthetic and real-world datasets, including ablation studies, and the results demonstrate the effectiveness and potential application value of our proposed ShadowCatcher and ShadowEstimator.

The contributions in this paper are summarized as follows:

- We study a practical and challenging problem of treatment effect estimation from observational data under collider bias.
- We propose a novel ShadowCatcher that automatically generates representations serving the role of shadow variables from the observed covariates, addressing the challenge of finding valid shadow variables in real-world scenarios.
- We propose a novel ShadowEstimator to estimate treatment effects using the generated shadow-variable representations to address the collider bias in observational data.
- Extensive experiments show that our proposed methods can practically generate shadow-variable representations and address collider bias in treatment effect estimation.

2. Problem and Algorithm

2.1. Problem Formulation

Suppose that we have a random sample of n units from a super population $\mathcal{P}$ where each unit $i = 1, \cdots, n$ has a set of covariates $\mathbf{x}_i \in \mathcal{X}$, the treatment $t_i \in \mathcal{T}$, and the outcome $y_i \in \mathcal{Y}$. We use a binary selection indicator $s_i \in \{0, 1\}$ that indicates whether the unit i is selected into the sample. As shown in Figure 1(d), in the presence of collider bias, for a unit with $s_i = 1$, we can observe the values of $\mathbf{x}_i, t_i$, and y_i, while for a unit with $s_i = 0$, we can only observe the values of $\mathbf{x}_i$ and t_i and the value of y_i is missing.

In this paper, we focus on the case of binary treatment[2], i.e., $t_i \in \{0, 1\}$, where $t_i = 1$ denotes unit i is treated, and $t_i = 0$ denotes otherwise. Under the potential outcome framework (Imbens & Rubin, 2015), we define the potential

[2]To make the proposed ShadowCatcher and ShadowEstimator process more concise, here we consider the binary treatment setting, but our proposed methods can also be effectively applied to continuous treatment settings.

outcomes under treatment as $Y(1)$ and under control as $Y(0)$. With the observational data, our goal is to estimate the Conditional Average Treatment effect (CATE), which is defined as $\tau(\mathbf{x}) = \mathbb{E}[Y(1) - Y(0) \mid \mathbf{X} = \mathbf{x}]$. For a unit i, only the factual outcome $Y(t_i)$ is available. Therefore, to make CATE identifiable, we make the following widely used assumptions (Imbens & Rubin, 2015):

- **Stable Unit Treatment Value Assumption.** The distribution of the potential outcome of one unit is independent of the treatment assignment of another unit.
- **Overlap Assumption.** A unit has a nonzero probability of being treated and being selected, $0 < \mathbb{P}(T = 1 \mid \mathbf{X} = \mathbf{x}) < 1$ and $0 < \mathbb{P}(S = 1 \mid \mathbf{X} = \mathbf{x}) < 1$.
- **Unconfoundedness Assumption.** The treatments are independent of the potential outcomes given the covariates, i.e., $Y(1), Y(0) \perp\!\!\!\perp T \mid \mathbf{X}$.

Based on these assumptions, CATE can be estimated as

$$\tau(\mathbf{x}) \overset{(1)}{=} \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}, T = 1] - \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}, T = 0].$$

However, because the values of Y are missing in $S = 0$ units, we can only estimate the CATE of $S = 1$ samples, which differs from the true CATE of the entire data because $\mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}, T = t, S = 1] \neq \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}, T = t]$. It leads to a biased estimation using the observed samples because conditioning on S opens a non-causal path $T \rightarrow S \leftarrow Y$. Therefore, it is necessary to develop approaches to solve collider bias for treatment effect estimation. Fortunately, studies show that treatment effects can be identifiable under collider bias if some shadow variables are available in the observational data (d'Haultfoeuille, 2010; Miao & Tchetgen Tchetgen, 2016; Miao et al., 2024).

2.2. Preliminaries of the Shadow Variable

Valid shadow variables $\mathbf{Z}$ are supposed to be fully observed covariates, i.e., the values of $\mathbf{Z}$ are observable in both $S = 0$ and $S = 1$ data and satisfy the following assumption:

Assumption 2.1. (d'Haultfoeuille, 2010) A valid shadow variable should satisfy the conditional dependence assumption $\mathbf{Z} \not\perp\!\!\!\perp Y \mid \mathbf{X}, T, S = 1$ and conditional independence assumption $\mathbf{Z} \perp\!\!\!\perp S \mid \mathbf{X}, T, Y$.

As shown in Figure 1(c), Assumption 2.1 indicates that the shadow variable does not affect the sample selection mechanism after conditioning on the outcome and other observed covariates, and it is associated with the outcome given the covariates. This assumption is widely used in the literature of collider bias (d'Haultfoeuille, 2010; Wang et al., 2014; Miao & Tchetgen Tchetgen, 2016; Zhao & Shao, 2016; Li et al., 2023; Miao et al., 2024), and an illustrative example can be found in Section 1.

Throughout the paper, we use $f(\cdot)$ to denote the data distribution function. The key problem of collider bias is that the outcome values are missing in $S = 0$ data, which results in $f(Y \mid \mathbf{X}, \mathbf{Z}, T, S = 0)$ not available from the observed data. We can use the odds ratio function $\mathrm{OR}(\mathbf{X}, \mathbf{Z}, T, Y)$ to encode the deviation between the distribution of $S = 1$ data and that of $S = 0$ data. Under Assumption 2.1 (Miao & Tchetgen Tchetgen, 2016), it equals

$$\mathrm{OR}(\mathbf{X}, T, Y) \overset{(2)}{=} \frac{f(S = 0 \mid \mathbf{X}, T, Y) \cdot f(S = 1 \mid \mathbf{X}, T, Y = 0)}{f(S = 0 \mid \mathbf{X}, T, Y = 0) \cdot f(S = 1 \mid \mathbf{X}, T, Y)},$$

and the proof can be found in Appendix D.1. In Eq. (2), $Y = 0$ is used as a reference value, and $\mathrm{OR}(\mathbf{X}, T, Y = 0) = 1$. Note that it can be replaced by any other value within the support of Y. The odds ratio function measures the degree to which the $S = 0$ data differs from the $S = 1$ data and thus can be used to recover the unknown $f(Y \mid \mathbf{X}, \mathbf{Z}, T, S = 0)$ from the known $f(Y \mid \mathbf{X}, \mathbf{Z}, T, S = 1)$ through the following proposition:

Proposition 2.2. *(Miao & Tchetgen Tchetgen, 2016; Miao et al., 2024) Under Assumption 2.1, we have*

$$\frac{\mathrm{OR}(\mathbf{X}, T, Y)}{\mathbb{E}[\mathrm{OR}(\mathbf{X}, T, Y) \mid \mathbf{X}, \mathbf{Z}, T, S = 1]} \overset{(3)}{=} \frac{f(Y \mid \mathbf{X}, \mathbf{Z}, T, S = 0)}{f(Y \mid \mathbf{X}, \mathbf{Z}, T, S = 1)}$$

and

$$\mathbb{E}[\widetilde{\mathrm{OR}}(\mathbf{X}, T, Y) \mid \mathbf{X}, \mathbf{Z}, T, S = 1] \overset{(4)}{=} \frac{f(\mathbf{Z} \mid \mathbf{X}, T, S = 0)}{f(\mathbf{Z} \mid \mathbf{X}, T, S = 1)},$$

where

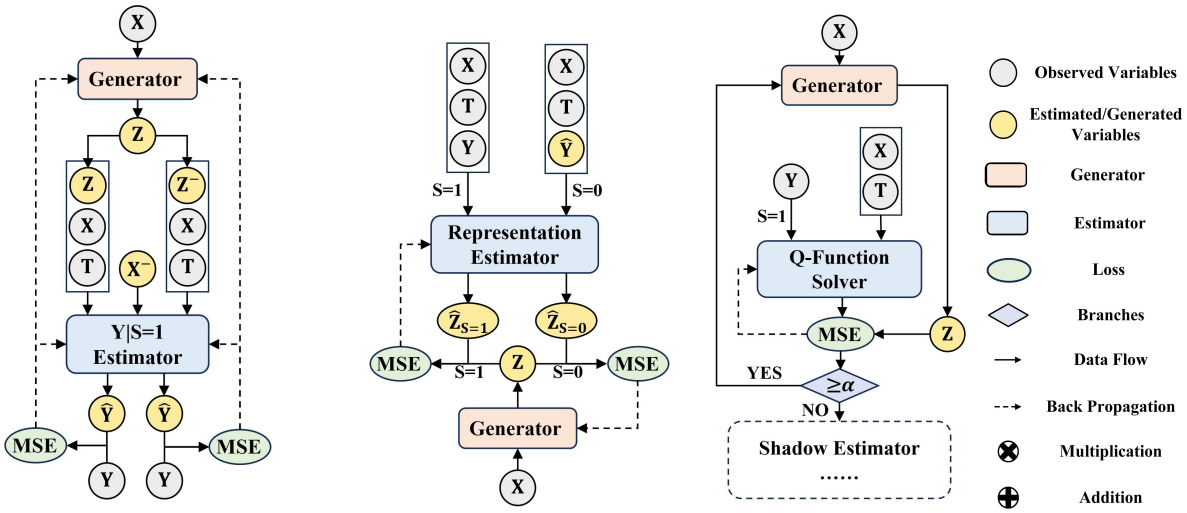
$$\widetilde{\mathrm{OR}}(\mathbf{X}, T, Y) = \frac{\mathrm{OR}(\mathbf{X}, T, Y)}{\mathbb{E}[\mathrm{OR}(\mathbf{X}, T, Y) \mid \mathbf{X}, T, S = 1]}.$$

Eq. (3) shows that the key challenge of collider bias, i.e., $f(Y \mid \mathbf{X}, \mathbf{Z}, T, S = 0)$ is unidentifiable, can be solved under Assumption 2.1 by integrating the odds ratio function with the $S = 1$ data distribution. Since $f(Y \mid \mathbf{X}, \mathbf{Z}, T, S = 1)$ can be obtained from the fully observed $S = 1$ samples, the only problem becomes the identification of the odds ratio function. Fortunately, with $f(\mathbf{Z} \mid \mathbf{X}, S = 0)$ and $f(\mathbf{Z} \mid \mathbf{X}, S = 1)$ obtained from the observed data, Eq. (4) is a Fredholm integral equation of the first kind, with $\widetilde{\mathrm{OR}}(\mathbf{X}, T, Y)$ to be solved for. Because $\mathrm{OR}(\mathbf{X}, T, Y = 0) = 1$, we have the following result (Miao et al., 2024): (the proof is in Appendix D.2)

$$\mathrm{OR}(\mathbf{X}, T, Y) \overset{(5)}{=} \frac{\widetilde{\mathrm{OR}}(\mathbf{X}, T, Y)}{\widetilde{\mathrm{OR}}(\mathbf{X}, T, Y = 0)}.$$

Therefore, identification of $\mathrm{OR}(\mathbf{X}, T, Y)$ is equivalent to finding a unique solution to Eq. (4), which is guaranteed by the following theorem.

Condition 2.3. *(Miao et al., 2024) For all square-integrable functions $h(\mathbf{X}, T, Y)$, $\mathbb{E}[h(\mathbf{X}, T, Y) \mid \mathbf{X}, \mathbf{Z}, T, S = 1] = 0$ almost surely if and only if $h(\mathbf{X}, T, Y) = 0$ almost surely.*



(a) Constraint 1 in the generation phase. (b) Constraint 2 in the generation phase. (c) Hypothesis test phase.

Figure 2. The flowchart of ShadowCatcher, including the generation phase and the test phase.

Theorem 2.4. *(Miao et al., 2024) Under Assumption 2.1 and Condition 2.3., Eq. (4) has a unique solution. Thus* $\mathrm{OR}(\mathbf{X}, T, Y)$ *and* $f(Y \mid \mathbf{X}, \mathbf{Z}, T)$ *can be identified.*

Based on the above theorem, collider bias can be solved with the help of shadow variables by firstly estimating $\mathrm{OR}(\mathbf{X}, T, Y)$ through Eq. (4) and Eq. (5), then recovering $f(Y \mid \mathbf{X}, \mathbf{Z}, T, S = 0)$ through Eq. (3), and finally estimating $f(Y \mid \mathbf{X}, \mathbf{Z}, T)$. However, finding a well-defined shadow variable in real-world scenarios is also challenging because it requires domain-specific knowledge from experts and must be investigated on a case-by-case basis (Li et al., 2023). To relax the assumption that prior knowledge about shadow variables is needed, we propose a novel ShadowCatcher to generate representations serving the role of shadow variables directly from observed covariates without prior knowledge and a novel ShadowEstimator to estimate CATE under collider bias with the help of the generated shadow-variable representations.

2.3. ShadowCatcher

Intuitively, as shown in Figure 1(c), the causal link $\mathbf{X} \to \mathbf{Z}$ indicates that the shadow variable is possible to be learned from the fully observed covariates. Therefore, our proposed ShadowCatcher aims to learn representations $\mathbf{Z}$ that satisfy the shadow variable assumptions from $\mathbf{X}$. To achieve this goal, we must ensure that the generated representations do satisfy Assumption 2.1.

As stated in Assumption 2.1, a valid shadow variable needs to satisfy the conditional dependence assumption $\mathbf{Z} \not\perp\!\!\!\perp Y \mid \mathbf{X}, T, S = 1$ and the conditional independence assumption $\mathbf{Z} \perp\!\!\!\perp S \mid \mathbf{X}, T, Y$. The first assumption can be easily tested with the observed data because only $S = 1$ data is involved. However, the second assumption needs Y to be fully observed, but the fact is that Y values are missing for $S = 0$ data. Fortunately, this assumption is proven refutable with only the observed data.

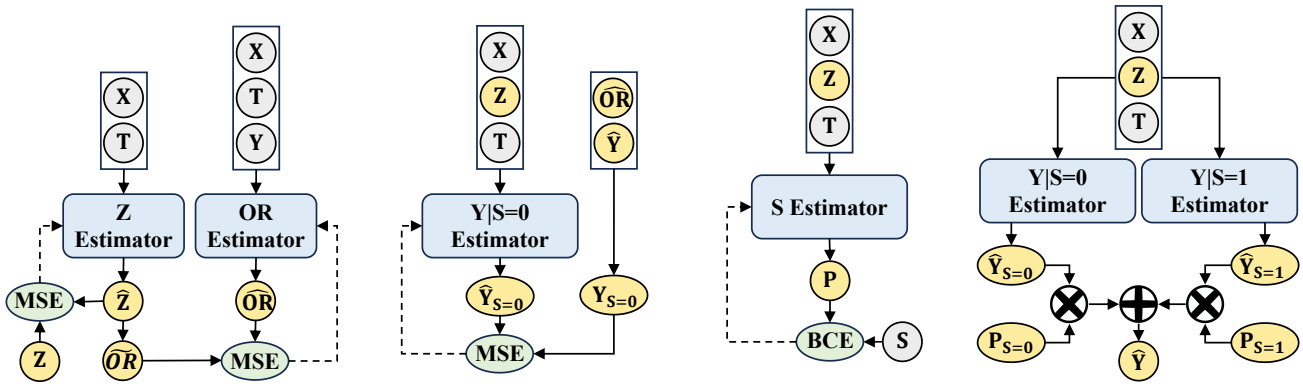
Theorem 2.5. *(d'Haultfoeuille, 2010) Suppose the overlap assumption and* $\mathbf{Z} \not\perp\!\!\!\perp Y \mid \mathbf{X}, T, S = 1$ *hold, then* $\mathbf{Z} \perp\!\!\!\perp S \mid \mathbf{X}, T, Y$ *can be rejected if and only if there does not exist any function* $Q(\cdot)$ *that satisfies the following equation and takes value between* $(0, 1]$*:*

$$\mathbb{E}\left[\frac{S}{Q(\mathbf{X}, T, Y)} - 1 \mid \mathbf{X}, \mathbf{Z}, T\right] \overset{(6)}{=} 0.$$

Note that Eq. (6) only involves the observed data since $\mathbf{X}, \mathbf{Z}, T$ are fully observed and $S/Q(\mathbf{X}, T, Y) = 0$ when $S = 0$. Hence, although we cannot directly test whether the generated $\mathbf{Z}$ satisfies the second assumption, we can test whether the generated $\mathbf{Z}$ can be rejected by Eq. (6). As a result, we can tell ShadowCatcher generates valid shadow-variable representations if and only if the generated $\mathbf{Z}$ is tested to be not refutable.

Therefore, ShadowCatcher iteratively generates shadow-variable representations and tests whether the generated representations satisfy Assumption 2.1 until they can pass the hypothesis test, detailed as follows.

Generation Phase. As shown in Figure 2.1, ShadowCatcher uses the representation generator $g : \mathcal{X} \to \mathcal{Z}$ to generate synthetic shadow variables $\mathbf{Z} = g(\mathbf{X})$, maintaining optimal predictive power for T and Y, independent of the selection mechanism, subject to the two constraints ($\mathbf{Z} \not\perp\!\!\!\perp Y \mid \mathbf{X}, T, S = 1$ and $\mathbf{Z} \perp\!\!\!\perp S \mid \mathbf{X}, T, Y$). Let the



(a) Estimation of $\mathrm{OR}(\mathbf{X}, T, Y)$. (b) Estimation of $f(Y|\mathbf{X}, \mathbf{Z}, T, S=0)$. (c) Estimation of $P(S|\mathbf{X}, \mathbf{Z}, T)$. (d) Estimation of $f(Y|\mathbf{X}, \mathbf{Z}, T)$.

Figure 3. The flowchart of ShadowEstimator, including four estimation procedures.

loss function to train g be ℓ_g. This loss function equals to $\ell_g = \ell_{g_y} + \ell_{g_z}$, where ℓ_{g_y} is the loss function from the constraint $\mathbf{Z} \not\perp\!\!\!\perp Y \mid \mathbf{X}, T, S=1$, and ℓ_{g_z} is the loss function from the constraint $\mathbf{Z} \perp\!\!\!\perp S \mid \mathbf{X}, T, Y$. Below we separately provide the expression of ℓ_{g_y} and ℓ_{g_z}.

(1) Constraining the conditional dependence assumption using ℓ_{g_y}. The generated shadow variables should maintain optimal predictive power for Y, and thus we propose an outcome prediction function $h_{y_1} : \mathcal{X} \times \mathcal{Z} \times \mathcal{T} \to \mathcal{Y}$ to estimate $f(Y \mid \mathbf{X}, \mathbf{Z}, T, S=1)$ with *the loss function being* $\ell_{y_1} = \frac{1}{n_1}\sum_{i:s_i=1}[(h_{y_1}(\mathbf{x}_i, \mathbf{z}_i, t_i) - y_i)^2 + (h_{y_1}(\mathbf{x}_i^-, \mathbf{z}_i, t_i) - y_i)^2]$, where n_1 denotes the number of $S=1$ units and $\mathbf{x}_i^-$ denotes random variables that differ from $\mathbf{x}_i$ generated by replacing the original $\mathbf{x}_i$ value with random noise. By minimizing this function, $h_{y_1}(\cdot)$ would maximize the embedding of the predictive information of $\mathbf{X}$ and $\mathbf{Z}$ for Y, separately. Additionally, to constrain the generated $\mathbf{Z}$ satisfying the conditional dependence assumption, i.e., $\mathbf{Z} \not\perp\!\!\!\perp Y \mid \mathbf{X}, T, S=1$, we need to make $f(Y \mid \mathbf{X}, \mathbf{Z}, T, S=1)$ differ from $f(Y \mid \mathbf{X}, \mathbf{Z}^-, T, S=1)$, where $\mathbf{Z}^-$ denotes random variables that differ from $\mathbf{Z}$ generated by replacing the original $\mathbf{z}_i$ value with random noise. Therefore, *one objective of the generator* is to simultaneously *minimize* the Mean-Square Error (MSE) between $h_{y_1}(\mathbf{X}_{S=1}, \mathbf{Z}_{S=1}, T_{S=1})$ and $Y_{S=1}$, and *maximize* the MSE between $h_{y_1}(\mathbf{X}_{S=1}, \mathbf{Z}_{S=1}^-, T_{S=1})$ and $Y_{S=1}$, where $\cdot_{S=1}$ denotes the corresponding variables of the $S=1$ data. The loss function of this constraint on the representation generator is

$$\ell_{g_y} = \frac{1}{n_1}\sum_{i:s_i=1}(h_{y_1}(\mathbf{x}_i, \mathbf{z}_i, t_i) - y_i)^2 - \frac{1}{n_1}\sum_{i:s_i=1}(h_{y_1}(\mathbf{x}_i, \mathbf{z}_i^-, t_i) - y_i)^2,$$

where h_{y_1} is trained using ℓ_{y_1}, and is fixed when training g.

(2) Constraining the conditional independence assumption using ℓ_{g_z}. This estimator aims to estimate $f(\mathbf{Z} \mid \mathbf{X}, T, Y, S=1)$ with $S=1$ samples. That is, we learn a representation prediction function $h_r : \mathcal{X} \times \mathcal{T} \times \mathcal{Y} \to \mathcal{Z}$ with *the loss function of this estimator being* $\ell_r = \frac{1}{n_1}\sum_{i:s_i=1}(h_r(\mathbf{x}_i, t_i, y_i) - \mathbf{z}_i)^2$. To constrain the generated $\mathbf{Z}$ satisfying the conditional independence assumption, i.e., $\mathbf{Z} \perp\!\!\!\perp S \mid \mathbf{X}, T, Y$, we need to make $f(\mathbf{Z} \mid \mathbf{X}, T, Y, S=1)$ the same as $f(\mathbf{Z} \mid \mathbf{X}, T, Y, S=0)$. Therefore, *the other objective of the generator* is to minimize the MSE between $h_r(\mathbf{X}_{S=0}, T_{S=0}, Y_{S=0})$ and $\mathbf{Z}_{S=0}$, where $\cdot_{S=0}$ denotes the corresponding variables of the $S=0$ data. The loss function of this constraint on the representation generator is

$$\ell_{g_z} = \frac{1}{n_0}\sum_{i:s_i=0}(h_r(\mathbf{x}_i, t_i, h_{y_1}(\mathbf{x}_i, \mathbf{z}_i, t_i)) - \mathbf{z}_i)^2,$$

where n_0 denotes the number of $S=0$ units, and h_r is trained using ℓ_r and is fixed when training g. Since the Y values are missing for $S=0$ units, here we use $\hat{Y}_{S=0}$ predicted by h_{y_1} as the substitute. This imputation approach may harm the constraining process, but we can control this impact in the subsequent hypothesis test phase.

Hypothesis Test Phase. In the generation process, the conditional independence assumption is not strictly constrained due to the missing Y values for $S=0$ units. Therefore, ShadowCatcher conducts an additional hypothesis test based on Theorem 2.5 after the generation phase finishes. The tester aims to learn a solution q of $Q(\mathbf{X}, T, Y)$ in Eq. (6) that belongs to $(0, 1]$ which turns into an optimization problem with the loss function being

$$\ell_q = \frac{1}{n}\sum_{i=1}^{n}\left\| \left(\frac{s_i}{q(\mathbf{x}_i, t_i, y_i)} - 1\right) \cdot \begin{pmatrix}\mathbf{x}_i \\ \mathbf{z}_i \\ t_i\end{pmatrix} \right\|_2^2,$$

where $q(\mathbf{x}_i, t_i, y_i)$ is a function from $\mathbb{R}$ to $(0, 1]$ and $\|\cdot\|_2$ denotes the ℓ_2 norm. Note that for $s_i = 0$ units, the value of

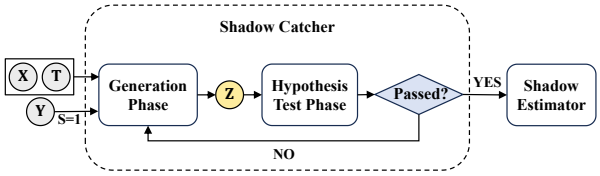


Figure 4. The overall flowchart of ShadowCatcher and ShadowEstimator.

$s_i/q(\mathbf{x}_i, t_i, y_i)$ equals 0, and thus, the entire optimization process does not involve missing y_i values. Therefore, when the loss function converges, if the loss value is greater than a given threshold α, which means it fails to learn a q that satisfies Eq. (6), we can tell that no solution of Eq. (6) belongs to $(0, 1]$ and Assumption 2.1 is rejected. Note that to preempt the possible multiple comparisons issue, we use Bonferroni correction (Dunn, 1961) to dynamically adjust α during training by setting α to α/m in the m-th iteration. As a result, the generated $\mathbf{Z}$ does not satisfy Assumption 2.1, and we need to regenerate it until it can pass the hypothesis test, i.e., the converged loss value is less than α. Finally, the first generated $\mathbf{Z}$ that passes the test can serve the role of shadow variables and be used for treatment effect estimation under collider bias by ShadowEstimator.

2.4. ShadowEstimator

With the help of the shadow-variable representations, we can estimate treatment effects under collider bias through:

- Estimation of the odds ratio function, i.e., estimiting $\widetilde{\mathrm{OR}}(\mathbf{X}, T, Y)$ and $\mathrm{OR}(\mathbf{X}, T, Y)$ by Eq. (4) and Eq. (5);
- Estimation of the conditional distrubution of the unselected outcomes, i.e., using Eq. (3) to recover and estimate $f(Y \mid \mathbf{X}, \mathbf{Z}, T, S = 0)$;
- Estimation of the treatment effects, i.e., estimating $f(Y \mid \mathbf{X}, \mathbf{Z}, T)$ and the CATE using the estimated $f(Y \mid \mathbf{X}, \mathbf{Z}, T, S = 0)$, $f(Y \mid \mathbf{X}, \mathbf{Z}, T, S = 1)$ and $f(S \mid \mathbf{X}, \mathbf{Z}, T)$. Note that the estimated $f(Y \mid \mathbf{X}, \mathbf{Z}, T, S = 1)$ is available from ShadowCatcher.

Estimation of the odds ratio function. With the generated $\mathbf{Z}$ and fully observed $\mathbf{X}$ and T, we first use two shadow-variable estimators $h_{\mathrm{z}_0} : \mathcal{X} \times \mathcal{T} \to \mathcal{Z}$ and $h_{\mathrm{z}_1} : \mathcal{X} \times \mathcal{T} \to \mathcal{Z}$ to estimate $f(\mathbf{Z} \mid \mathbf{X}, T, S = 0)$ and $f(\mathbf{Z} \mid \mathbf{X}, T, S = 1)$ respectively. The loss functions of these estimators are

$$\ell_{\mathrm{z}_0} = \frac{1}{n_0} \sum\nolimits_{i:s_i=0} (h_{\mathrm{z}_0}(\mathbf{x}_i, t_i) - \mathbf{z}_i)^2,$$

and

$$\ell_{\mathrm{z}_1} = \frac{1}{n_1} \sum\nolimits_{i:s_i=1} (h_{\mathrm{z}_1}(\mathbf{x}_i, t_i) - \mathbf{z}_i)^2.$$

Using $\mathbf{X}$, T, and Y of the $S = 1$ units and $h_{\mathrm{z}_0}(\mathbf{X}, T)/h_{\mathrm{z}_1}(\mathbf{X}, T)$ as the ground truths, we estimate $\widetilde{\mathrm{OR}}(\mathbf{X}, T, Y)$ with the loss function being

$$\ell_{\widetilde{\mathrm{or}}} = \frac{1}{n_1} \sum\nolimits_{i:s_i=1} \left(\widetilde{\mathrm{or}}(\mathbf{x}_i, t_i, y_i) - \frac{h_{\mathrm{z}_0}(\mathbf{x}_i, t_i)}{h_{\mathrm{z}_1}(\mathbf{x}_i, t_i)} \right)^2,$$

where $\widetilde{\mathrm{or}}(\cdot)$ is the estimated $\widetilde{\mathrm{OR}}(\cdot)$. Then we can obtain $\mathrm{OR}(\mathbf{X}, T, Y)$ with $\widetilde{\mathrm{or}}(\cdot)$ by Eq. (5).

Estimation of the conditional distribution of the unselected outcomes. With the estimated $\mathrm{OR}(\mathbf{X}, T, Y)$ and $f(Y \mid \mathbf{X}, \mathbf{Z}, T, S = 1)$, the ground truth counterfactual outcomes of the $S = 1$ samples, i.e., the outcome values of the $S = 1$ samples result from $f(Y \mid \mathbf{X}, \mathbf{Z}, T, S = 0)$, can be obtained by Eq. (3). Therefore, we can learn another outcome prediction function $h_{\mathrm{y}_0} : \mathcal{X} \times \mathcal{Z} \times \mathcal{T} \to \mathcal{Y}$ to estimate $f(Y \mid \mathbf{X}, \mathbf{Z}, T, S = 0)$ using $S = 1$ samples. The loss function of this estimator is

$$\ell_{\mathrm{y}_0} = \frac{1}{n_1} \sum\nolimits_{i:s_i=1} (h_{\mathrm{y}_0}(\mathbf{x}_i, \mathbf{z}_i, t_i) - \frac{\widetilde{\mathrm{or}}(\mathbf{x}_i, t_i, y_i)}{\widetilde{\mathrm{or}}(\mathbf{x}_i, t_i, h_{\mathrm{y}_1}(\mathbf{x}_i, \mathbf{z}_i, t_i))} \cdot h_{\mathrm{y}_1}(\mathbf{x}_i, \mathbf{z}_i, t_i))^2.$$

Estimation of the treatment effects. Now that $f(Y \mid \mathbf{X}, \mathbf{Z}, T, S = 0)$ and $f(Y \mid \mathbf{X}, \mathbf{Z}, T, S = 1)$ are both estimated, estimation of $f(Y \mid \mathbf{X}, \mathbf{Z}, T)$ becomes estimation of $f(S \mid \mathbf{X}, \mathbf{Z}, T)$, which can be achieved by learning an S prediction function $h_{\mathrm{s}} : \mathcal{X} \times \mathcal{Z} \times \mathcal{T} \to \mathcal{S}$, where $\mathcal{S}$ is the space of S. The loss function of this estimator is

$$\ell_{\mathrm{s}} = -\frac{1}{n} \sum_{i=1}^{n} (s_i \cdot \log(h_{\mathrm{s}}(\mathbf{x}_i, \mathbf{z}_i, t_i)) + (1 - s_i) \cdot \log(1 - h_{\mathrm{s}}(\mathbf{x}_i, \mathbf{z}_i, t_i))),$$

and then we can obtain $f(Y \mid \mathbf{X}, \mathbf{Z}, T)$ by

$$f(Y \mid \mathbf{X}, \mathbf{Z}, T) = \sum\nolimits_{s \in \{0,1\}} f(Y \mid \mathbf{X}, \mathbf{Z}, T, S = s) \cdot f(S = s \mid \mathbf{X}, \mathbf{Z}, T).$$

Consequently, we can use Eq. (1) to achieve CATE estimation. Additionally, we add an Integral Probability Metric (IPM) term to the outcome estimators following Shalit et al. (2017) to address possible confounding bias.

In summary, the overall framework of ShadowCatcher and ShadowEstimator is shown in Figure 4: ShadowCatcher first takes the fully observed $\mathbf{X}$ and T, and the observed Y of $S = 1$ units as inputs to generate shadow-variable representations $\mathbf{Z}$. Subsequently, it tests whether the generated $\mathbf{Z}$ satisfies Assumption 2.1. If the generated $\mathbf{Z}$ does not pass the hypothesis test, ShadowCatcher should re-generate new shadow-variable representations until the generated $\mathbf{Z}$ finally passes the test. After that, ShadowEstimator uses the generated $\mathbf{Z}$ to estimate treatment effects with observational samples. The pseudo-codes are in Appendix A, and the source code is available at `https://github.com/ZJUBaohongLi/ShadowCatcher-ShadowEstimator`.

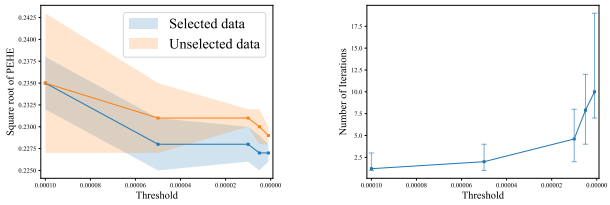
Table 1. The results of CATE estimation ($\sqrt{\text{PEHE}}$) on synthetic datasets under different β.

	$\beta = 1$		$\beta = 3$		$\beta = 5$	
ESTIMATOR	SELECTED DATA	UNSELECTED DATA	SELECTED DATA	UNSELECTED DATA	SELECTED DATA	UNSELECTED DATA
HECKIT	0.323±0.065	0.330±0.046	0.340±0.055	0.352±0.042	0.349±0.069	0.413±0.048
DR	0.298±0.032	0.316±0.042	0.331±0.048	0.357±0.053	0.367±0.033	0.448±0.017
IPSW	0.328±0.048	0.348±0.049	0.328±0.031	0.353±0.034	0.465±0.011	0.545±0.014
BNN	0.290±0.011	0.306±0.012	0.329±0.048	0.354±0.033	0.359±0.011	0.439±0.015
TARNET	0.295±0.012	0.312±0.011	0.329±0.030	0.357±0.053	0.366±0.071	0.436±0.087
CFR	0.290±0.009	0.307±0.008	0.324±0.009	0.350±0.013	0.359±0.008	0.436±0.030
CFOREST	0.310±0.030	0.331±0.038	0.338±0.019	0.368±0.022	0.373±0.026	0.453±0.043
DR-CFR	0.284±0.038	0.307±0.040	0.340±0.055	0.355±0.064	0.366±0.051	0.435±0.060
TEDVAE	0.281±0.056	0.419±0.070	0.378±0.063	0.420±0.059	0.394±0.054	0.431±0.067
DER-CFR	0.291±0.010	0.309±0.014	0.323±0.015	0.348±0.017	0.358±0.011	0.439±0.013
DESCN	0.295±0.002	0.312±0.002	0.326±0.003	0.357±0.004	0.365±0.003	0.449±0.011
ES-CFR	0.289±0.003	0.305±0.004	0.331±0.003	0.359±0.003	0.369±0.003	0.448±0.005
OURS	**0.227±0.001**	**0.229±0.001**	**0.249±0.013**	**0.255±0.021**	**0.299±0.008**	**0.300±0.008**

3. Experiments

3.1. Baselines

Currently, no causal inference method can solve collider bias without introducing additional assumptions and prior knowledge. Therefore, we implemented the following treatment effect estimators that focus on confounding bias and sample selection bias caused by $\mathbf{X}$ and T as our baselines, including three groups. (1) Statistical estimators: Heckman's Correction (Heckit) (Heckman, 1979), Doubly Robust (Bang & Robins, 2005), Inverse Probability of Sampling Weights (IPSW) (Cole & Stuart, 2010), and Causal Forest (CForest) (Wager & Athey, 2018). (2) Balanced representation learning estimators: Balancing Neural Network (BNN), Treatment-Agnostic Representation Network (TARNet) (Johansson et al., 2016), CounterFactual Regression (CFR) (Shalit et al., 2017), Deep Entire Space Cross Networks (DESCN) (Zhong et al., 2022), and Entire Space CounterFactual Regression (ES-CFR) (Wang et al., 2023). (3) Disentangled representation learning estimators: Disentangled Representations for CounterFactual Regression (DR-CFR) (Greiner, 2020), TEDVAE (Zhang et al., 2021), and Decomposed Representations for CounterFactual Regression (DeR-CFR) (Wu et al., 2022). We used the above baselines to estimate and compare the CATE with our proposed methods. Based on the estimated CATE, we use the Precision in Estimation of Heterogeneous Effect (PEHE) (Shalit et al., 2017; Louizos et al., 2017) to evaluate the performance of the estimators, where $\text{PEHE} = \frac{1}{N} \cdot \sum_{i=1}^{N} ((\hat{y}_i(1) - \hat{y}_i(0)) - (y_i(1) - y_i(0))^2$. We split each dataset into 60/20/20 train/validation/test datasets, independently repeated 20 times, and report the mean and standard deviation (std) of $\sqrt{\text{PEHE}}$ for all experiments, formed as mean ± std in the tables.



(a) CATE estimates ($\sqrt{\text{PEHE}}$). (b) Number of iterations.

Figure 5. Ablation studies of the reject threshold α.

3.2. Experiments on Synthetic Data

3.2.1. DATASETS

In order to better evaluate the performance of each estimator under collider bias, we generated synthetic datasets with different collider bias strengths, denoted by β, which affects the impact of Y on S. The size n of all datasets was 10,000, and the dimension d of the covariates was 10. To compare our methods with the baselines under different strengths of collider bias, we evaluated the performance of each estimator under $\beta = \{1, 3, 5\}$. Moreover, we aim to prove the effectiveness of the constraints on the shadow-variable assumptions in ShadowCatcher. Therefore, we conducted ablation studies, including a comparison between ShadowCatcher and an ablation version without the conditional dependence constraint and another version without the hypothesis tester that guarantees the conditional independence constraint. Moreover, we performed additional experiments on synthetic data to evaluate the impact of different proportions of non-shadow variables in the covariates. We also compared the proposed method with shadow-variable regression using correctly specified shadow variables (Miao & Tchetgen Tchetgen, 2016). The data generation process

Table 2. The results of CATE estimation on three real-world datasets.

	IHDP ($\sqrt{\text{PEHE}}$)		ACIC 2016 ($\sqrt{\text{PEHE}}$)		JOBS ($\hat{R}_{\text{Pol}}$)	
ESTIMATOR	WITHIN-SAMPLE	OUT-OF-SAMPLE	WITHIN-SAMPLE	OUT-OF-SAMPLE	WITHIN-SAMPLE	OUT-OF-SAMPLE
HECKIT	1.587±0.065	1.621±0.041	3.106±0.444	3.340±0.111	0.328±0.050	0.331±0.052
DR	1.355±0.123	1.572±0.205	2.346±0.129	2.653±0.222	0.316±0.007	0.317±0.036
IPSW	2.118±0.344	2.129±0.295	4.244±0.145	5.411±0.073	0.284±0.051	0.289±0.063
BNN	1.308±0.298	1.457±0.339	2.173±0.150	2.586±0.486	0.303±0.025	0.304±0.041
TARNET	1.240±0.158	1.416±0.154	2.275±0.756	2.805±0.766	0.315±0.012	0.316±0.050
CFR	1.283±0.186	1.401±0.238	2.107±0.297	2.361±0.587	0.313±0.018	0.314±0.072
CFOREST	1.702±0.292	1.948±0.429	4.137±0.295	4.605±0.137	0.326±0.012	0.326±0.059
DR-CFR	1.299±0.087	1.399±0.171	2.240±0.691	2.340±0.663	0.322±0.022	0.323±0.099
TEDVAE	4.246±0.394	4.347±0.563	3.501±0.708	4.468±0.813	0.296±0.046	0.300±0.031
DER-CFR	1.446±0.345	1.571±0.371	2.214±0.204	2.246±0.598	0.309±0.023	0.311±0.029
DESCN	1.193±0.057	1.665±0.246	2.185±0.150	2.306±0.236	0.331±0.010	0.331±0.051
ES-CFR	1.499±0.096	1.436±0.095	3.875±0.224	4.494±0.214	0.290±0.045	0.293±0.046
OURS	**0.703±0.106**	**0.723±0.102**	**1.911±0.126**	**2.047±0.351**	**0.279±0.017**	**0.280±0.018**

and the ablations are detailed in Appendix C.

3.2.2. RESULTS

We separately report the results of the selected data ($S = 1$) and unselected data ($S = 0$) in Table 1 under different collider bias strengths with $\beta = \{1, 3, 5\}$. We observe that: (1) The overall performance of DR, BNN, CFR, CForest, TEDVAE, DR-CFR, DESCN, DeR-CFR and ES-CFR is not good because they all focus on confounding bias and thus cannot deal with sample selection bias. (2) The performance of Heckit and IPSW is also poor because they can only address sample selection bias caused by T and $\mathbf{X}$ and cannot address collider bias because of the spurious association $T \rightarrow S \leftarrow Y$. (3) Our method outperforms all baselines under all β settings because the generated representations by ShadowCatcher make identification under collider bias possible, and ShadowEstimator provides a practical solution. (4) As collider bias strengthens, the performance gap between selected and unselected data increases. However, this gap for our method is much smaller than that of other baselines, which demonstrates that our proposed approaches can practically address collider bias.

In ShadowCatcher, we introduce a hyperparameter α that is the rejection threshold of the test phase. The choice of the reject threshold α is a tradeoff between efficiency and performance during the generation process of ShadowCatcher: if the reject threshold is too small, the generated representations may be too weak to be a valid shadow variable; if the threshold is too large, it may needs more iterations for the generated representations to pass the test. To further study the impact of different options of α on the efficiency and performance of ShadowCatcher, we conducted experiments with $\alpha = \{10^{-4}, 5 \times 10^{-5}, 10^{-5}, 5 \times 10^{-6}, 10^{-6}\}$ on the synthetic dataset in Section 3.2.1 with $d_s = 0.9 \cdot d$ and $\beta = 1$. The results are in Figure 5. It shows that the performance of ShadowCatcher improves as the reject threshold decreases because the hypothesis test gets more strict, which means the constraint gets more reliable. However, the number of iterations ShadowCatcher requires to pass the hypothesis test also increases quickly, reducing its efficiency. Therefore, choosing an appropriate α is a tradeoff between efficiency and performance and depends on the application scenarios.

3.3. Experiments on Real-World Data

3.3.1. DATASETS

In order to evaluate the proposed method in real-world scenarios, we conducted experiments on three well-known datasets: **the IHDP dataset** (Hill, 2011)[3], **the ACIC 2016 dataset** (Dorie et al., 2019)[4], and **the Jobs dataset** (Shalit et al., 2017)[5]. For the IHDP and ACIC 2016 datasets, the treatment assignment and outcome generation process are simulated based on covariates collected from real-world applications. Therefore, the ground truth CATE is known, and we use the same metric as those in the experiments on the synthetic data. For the Jobs dataset, because the ground truth CATE is unknown, we follow Shalit et al. (2017) to use the policy risk to evaluate the quality of CATE estimation. The policy risk is defined as the average loss in value when treating according to the policy implied by a CATE estimator: $\hat{R}_{\text{Pol}} = 1 - (\mathbb{E}[Y(1) \mid \tau(\mathbf{x}) > 0, T = 1] \cdot \mathbb{P}(\tau(\mathbf{x}) > 0) + \mathbb{E}[Y(0) \mid \tau(\mathbf{x}) \leq 0, T = 0] \cdot \mathbb{P}(\tau(\mathbf{x}) \leq 0)$. We report the mean and std of the policy risk formed as mean ± std in the table. The original three datasets exhibit confounding bias but lack collider bias, so we introduced collider bias into them. For the IHDP and Jobs datasets, we selectively omitted the outcome values of certain sub-samples that met specific criteria. Regarding the ACIC 2016 dataset, we employed the same simulation method used for generating

[3] http://www.fredjo.com/

[4] https://github.com/vdorie/aciccomp/tree/master/2016

[5] https://users.nber.org/~rdehejia/nswdata2.html

synthetic data to obtain S. More details about these datasets and the simulation process are provided in Appendix C.3.

3.3.2. RESULTS

We separately report the results of within-sample data and out-of-sample data in Table 2, where within-sample means that the (factual) outcome of one treatment is observed, i.e., the $S = 1$ samples for training, and out-of-sample means no observed outcomes, i.e., the $S = 1$ samples for testing and all $S = 0$ samples (Shalit et al., 2017). From the results, we observe that: (1) The performance of the methods on confounding bias is not good because they cannot address sample selection bias. (2) The performance of the methods on sample selection bias is also poor because they can only address the cases that $\mathbf{X}$ and T cause S and thus cannot achieve a better estimate under collider bias. (3) Our method outperforms all baselines on both datasets because ShadowCatcher and ShadowEstimator effectively address collider bias in data. (4) The performance gap between our proposed method's within-sample and out-of-sample data is also overall the lowest, proving the ability to counterfactual prediction of our method. (5) Our method shows the lowest policy risk on the Jobs dataset, which demonstrates the effectiveness of our methods in real-world applications.

4. Conclusion

In this paper, we overcome the challenge of finding valid shadow variables to estimate treatment effects under collider bias in observational studies. We propose a novel ShadowCatcher that can generate representations serving the role of shadow variables and a novel ShadowEstimator that uses the generated representations to estimate CATE under collider bias. Experimental results demonstrate the effectiveness and application value of ShadowCatcher and ShadowEstimator.

One main limitation of our work is that the choice of the reject threshold α is a tradeoff between efficiency and performance during the generation process of ShadowCatcher as analyzed in Section 3.2.2. Another limitation is that the performance of ShadowCatcher depends on the extent to which covariates are involved in the sample selection. If no latent information satisfying the shadow variable assumption exists in the raw data covariates, extracting representations that satisfy the assumption of shadow variables (conditional independent of S) can be challenging. However, in most real-world scenarios, it is common that some covariates do not cause the sample selection directly. Therefore, the proposed method is not typically susceptible to this issue in most practical contexts.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62376243, 62441605, U20A20387, 623B2002), and the Starry Night Science Fund of Zhejiang University Shanghai Institute for Advanced Study (SN-ZJU-SIAS-0010).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Almond, D., Chay, K. Y., and Lee, D. S. The Costs of Low Birth Weight. *The Quarterly Journal of Economics*, 120 (3):1031–1083, 2005.

Athey, S., Imbens, G. W., and Wager, S. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(4): 597–623, 2018.

Bang, H. and Robins, J. M. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61 (4):962–73, 2005.

Bareinboim, E. and Tian, J. Recovering causal effects from selection bias. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Bareinboim, E., Tian, J., and Pearl, J. Recovering from selection bias in causal and statistical inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014.

Bottou, L., Peters, J., Quiñonero-Candela, J., Charles, D. X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P., and Snelson, E. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(11), 2013.

Brooksgunn, J., Liaw, F. R., and Klebanov, P. K. Effects of early intervention on cognitive function of low-birthweight preterm infants. *Journal of Pediatrics*, 120(3): 350–359, 1992.

Cole, S. R. and Stuart, E. A. Generalizing evidence from randomized clinical trials to target populations. *American Journal of Epidemiology*, 172(1):107–115, 2010.

Cuturi, M. and Doucet, A. Fast computation of wasserstein barycenters. In *International Conference on Machine Learning*, volume 32, pp. 685–693. JMLR, 2014.

Dehejia, R. H. and Wahba, S. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, 84(1):151–161, 2002.

Ding, P. Bayesian robust inference of sample selection using selection-tmodels. *Journal of Multivariate Analysis*, 124: 451–464, 2014.

Dorie, V., Hill, J., Shalit, U., Scott, M., and Cervone, D. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68, 2019.

Dunn, O. J. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961.

d'Haultfoeuille, X. A new instrumental method for dealing with endogenous selection. *Journal of Econometrics*, 154 (1):1–15, 2010.

Greenland, S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*, 14(3):300–6, 2003.

Greiner, N. H. R. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2020.

Guo, R. C., Cheng, L., Li, J. D., Hahn, P. R., and Liu, H. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys*, 53(4):75:1–75:37, 2020.

Hainmueller, J. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20 (1):25–46, 2012.

Heckman, J. J. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, 47(1): 153–161, 1979.

Hernán, M. A. and Robins, J. M. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.

Hill, J. L. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Ibrahim, J. G., Lipsitz, S. R., and Horton, N. Using auxiliary data for parameter estimation with non-ignorably missing outcomes. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 50(3):361–373, 2001.

Imbens, G. W. and Rubin, D. B. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, volume 37, pp. 448–456, 2015.

Johansson, F., Shalit, U., and Sontag, D. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, volume 48, pp. 3020–3029. PMLR, 2016.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Kohavi, R. and Longbotham, R. Unexpected results in online controlled experiments. *ACM SIGKDD Explorations Newsletter*, 12(2):31–35, 2011.

LaLonde, R. Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76(4):604–20, 1986.

Li, W., Miao, W., and Tchetgen Tchetgen, E. Nonparametric inference about mean functionals of nonignorable non-response data without identifying the joint distribution. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):913–935, 2023.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. Causal effect inference with deep latent-variable models. *Advances in Neural Information Processing Systems*, 30, 2017.

Marchenko, Y. V. and Genton, M. G. A heckman selection-tmodel. *Journal of the American Statistical Association*, 107(497):304–317, 2012.

Miao, W. and Tchetgen Tchetgen, E. J. On varieties of doubly robust estimators under missingness not at random with a shadow variable. *Biometrika*, 103(2):475–482, 2016.

Miao, W., Liu, L., Li, Y., Tchetgen Tchetgen, E. J., and Geng, Z. Identification and semiparametric efficiency theory of nonignorable missing data with a shadow variable. *ACM/JMS Journal of Data Science*, 1(2):1–23, 2024.

Ogundimu, E. O. and Hutton, J. L. A sample selection model with skew-normal distribution. *Scandinavian Journal of Statistics*, 43(1):172–190, 2016.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, volume 70, pp. 3076–3085. PMLR, 2017.

Wager, S. and Athey, S. Estimation and inference of heterogeneous treatment effects using random forests. *Journal*

of the American Statistical Association, 113(523):1228–1242, 2018.

Wang, H., Fan, J., Chen, Z., Li, H., Liu, W., Liu, T., Dai, Q., Wang, Y., Dong, Z., and Tang, R. Optimal transport for treatment effect estimation. *Advances in Neural Information Processing Systems*, 2023.

Wang, S., Shao, J., and Kim, J. k. An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica*, 2014.

Wang, X., Wu, Y., Zhang, A., Feng, F., He, X., and Chua, T.-S. Reinforced causal explainer for graph neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2297–2309, 2022.

Wiemann, P. F. V., Klein, N., and Kneib, T. Correcting for sample selection bias in bayesian distributional regression models. *Computational Statistics & Data Analysis*, 168: 107382, 2022.

Wu, A., Yuan, J., Kuang, K., Li, B., Wu, R., Zhu, Q., Zhuang, Y., and Wu, F. Learning decomposed representations for treatment effect estimation. *IEEE Transactions on Knowledge and Data Engineering*, 35(5):4989–5001, 2022.

Yoon, J., Jordon, J., and van der Schaar, M. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018.

Zhang, M., Yuan, J., He, Y., Li, W., Chen, Z., and Kuang, K. MAP: Towards balanced generalization of iid and ood through model-agnostic adapters. In *Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV*, pp. 11921–11931, 2023.

Zhang, M., Li, H., Wu, F., and Kuang, K. Metacoco: A new few-shot classification benchmark with spurious correlation. In *International Conference on Learning Representations, ICLR*, 2024.

Zhang, W., Liu, L., and Li, J. Treatment effect estimation with disentangled latent factors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2021.

Zhao, J. and Shao, J. Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. *Journal of the American Statistical Association*, 110 (512):1577–1590, 2016.

Zhong, K., Xiao, F., Ren, Y., Liang, Y., Yao, W., Yang, X., and Cen, L. DESCN: Deep entire space cross networks for individual treatment effect estimation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4612–4620, 2022.

A. Pseudo-Codes of ShadowCatcher and ShadowEstimator

As stated in Section 2, we propose a novel ShadowCatcher that generates representations serving the role of shadow variables and a novel ShadowEstimator that estimates treatment effects under collider bias with the help of the generated representations. The pseudo-codes of ShadowCatcher and ShadowEstimator are detailed in Algorithm 1 and 2, where g denotes the representations generator, h_{y_1} denotes the selected outcome estimator, h_{y_0} denotes the unselected outcome estimator, h_{r} denotes the representations estimator, h_{z_1} and h_{z_0} denote the shadow-variable estimators, $\widetilde{\mathrm{or}}$ denotes the odds ratio estimator, h_{s} denotes the sample selection estimator, and q denotes the Q function solver.

Algorithm 1 ShadowCatcher

Input: the observational samples with $\mathbf{X}, T, Y$, and S, reject threshold α.
Output: generated shadow-variable representations $\mathbf{Z}$.

```
m ← 1.
initialization of parameters in h_{y_1}, h_r, q and g.
repeat
  α ← α/m.
  m ← m + 1.
  repeat
    Z ← g(X).
    optimize h_{y_1} by ℓ_{y_1} and h_r by ℓ_r with S = 1 units.
    use h_{y_1} to predict the missing Y values for S = 0 units as replacements of the true values.
    optimize g by ℓ_g = ℓ_{g_y} + ℓ_{g_z} with all units.
  until convergence
  repeat
    optimize q by ℓ_q with all units.
  until convergence
  update l_q with the final output of ℓ_q.
until l_q ≥ α
Z ← g(X).
return Z
```

Algorithm 2 ShadowEstimator

Input: the observational samples with $\mathbf{X}, T, Y, S$, and $\mathbf{Z}$.
Output: the CATE of all units in the observational samples.

```
initialization of parameters in h_{y_0}, h_{z_0}, h_{z_1}, or~ and h_s.
repeat
  optimize h_{z_0} and h_{z_1} by minimizing ℓ_{z_0} and ℓ_{z_1}.
until convergence
calculate the "ground truth" values of ~OR(X, T, Y) by Eq. (4).
repeat
  optimize or~ by minimizing ℓ_or with S = 1 units and the calculated "ground truth" values.
until convergence
calculate the "ground truth" values of OR(X, T, Y) by Eq. (5) and Y values of S = 0 units by Eq. (3).
repeat
  optimize h_{y_0} by minimizing ℓ_{y_0} with S = 0 units and the calculated "ground truth" values.
  optimize h_s by minimizing ℓ_s with all units.
until convergence
calculate the CATE of all units with the optimized h_{y_0}, h_{y_1} and h_s.
return CATE
```

Table 3. The hyperparameters of ShadowCatcher and ShadowEstimator on different datasets.

DATASET	EPOCHS	BATCH SIZE	LEARNING RATE	WEIGHT DECAY	IPM WEIGHT	α
SYNTHETIC DATASETS	100	1024	0.03	0.01	0.001	1E-6
THE IHDP DATASET	100	128	0.03	0.01	0.001	0.01
THE TWINS DATASET	100	1024	0.03	0.01	0.1	0.1
THE JOBS DATASET	100	256	0.003	0.001	0.1	0.1
THE ACIC 2016 DATASETS	100	256	0.01	0.001	0.001	100

B. Related Work

Previous works on treatment effect estimation mainly focus on confounding bias in observational studies. Reweighting methods either use the inverse propensity score (Dehejia & Wahba, 2002) or learn a balancing weight from data (Hainmueller, 2012; Athey et al., 2018) to make T and $\mathbf{X}$ of the reweighted samples independent. Balanced representation learning methods (Johansson et al., 2016; Shalit et al., 2017; Greiner, 2020; Wang et al., 2023) learn representations of covariates so that the learned representations are independent of the treatment variable. Causal Forest (Wager & Athey, 2018) builds a large number of causal trees and then estimates heterogeneous treatment effects by taking an average of the outcomes from these causal trees. Generative methods (Louizos et al., 2017; Yoon et al., 2018; Zhang et al., 2021) utilize generative models to generate counterfactual data. However, all the above methods suffer from sample selection bias because of the distribution shift problem.

To address sample selection bias, Heckman (1979) proposed a two-stage regression method with many extensions (Marchenko & Genton, 2012; Ding, 2014; Ogundimu & Hutton, 2016; Wiemann et al., 2022). Cole & Stuart (2010) proposed a sample reweighting method, which reweights the selected samples by estimating the inverse conditional probability of the sample selection as weights. Bareinboim et al. (2014); Bareinboim & Tian (2015) proposed the selection-backdoor adjustment approach by blocking the selection-backdoor paths. These methods can only solve selection bias caused by covariates and the treatment. However, these methods cannot solve collider bias, which is more likely to appear in real-world scenarios because Y also causes S.

Fortunately, treatment effects are identifiable under collider bias if some shadow variables are available in the observational data (d'Haultfoeuille, 2010; Miao & Tchetgen Tchetgen, 2016; Miao et al., 2024). Shadow variables are assumed to satisfy that $\mathbf{Z} \not\perp\!\!\!\perp Y \mid \mathbf{X}, T, S=1$ and $\mathbf{Z} \perp\!\!\!\perp S \mid \mathbf{X}, T, Y$. With the help of shadow variables, various estimators, including regression-based (d'Haultfoeuille, 2010; Zhao & Shao, 2016), IPSW-based (Wang et al., 2014), and doubly robust-based (Miao & Tchetgen Tchetgen, 2016) were proposed to solve collider bias. However, the accessibility of valid shadow variables itself is a strong assumption because finding a well-defined shadow variable requires domain-specific knowledge from experts and needs to be investigated on a case-by-case basis (Li et al., 2023). Therefore, our proposed method that automatically generates representations serving the role of shadow variables can effectively relax the assumptions of solving collider bias and has excellent application values.

C. Supplement to the Experiments Section

C.1. Implementation Details

We utilized 3-layer Neural Networks to implement each module in ShadowEstimator and ShadowCatcher. We used the Adam optimizer (Kingma & Ba, 2015) with batch normalization (Ioffe & Szegedy, 2015) in the training process, and we used the Wasserstein distance (Cuturi & Doucet, 2014) as the Integral Probability Metric (IPM) to implement all the methods that need IPM to balance representations. The hyperparameters of our methods on different datasets are detailed in Table 3. We implemented all the methods in the PyTorch environment with Python 3.9. The CPU was 13th Gen Intel(R) Core(TM) i7-13700K, and the GPU was NVIDIA GeForce RTX 3080 with CUDA 12.1.

C.2. Data Generation Process of the Synthetic Datasets

We first generated the continuous covariates $\mathbf{X} \in \mathbb{R}^{n \times d}$ with independent Gaussian distributions as $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$, and then generated the binary treatment variable $T \in \mathbb{R}^n$ from a logistic function as $T \sim \text{Bernoulli}(1/(1+e^{-t(\mathbf{X})}))$, where $\text{Bernoulli}(\cdot)$ denotes the Bernoulli distribution, $t(\mathbf{X}) = \sum_{i=1}^{d}(\mathbf{1}(\text{mod}(i,2) \neq 1) - \mathbf{1}(\text{mod}(i,2) \equiv 1)) \cdot X_i/d) + \epsilon_{\text{t}}$, $\mathbf{1}(\cdot)$ is

Table 4. The results of CATE estimation ($\sqrt{\text{PEHE}}$) on the Twins datasets.

ESTIMATORS	WITHIN-SAMPLE	OUT-OF-SAMPLE
HECKIT	0.345±0.023	0.357±0.023
DR	0.476±0.010	0.487±0.007
IPSW	0.339±0.009	0.344±0.021
BNN	0.358±0.021	0.373±0.021
TARNET	0.401±0.049	0.407±0.058
CFR	0.361±0.040	0.369±0.040
CFOREST	0.356±0.034	0.421±0.035
DR-CFR	0.340±0.028	0.350±0.028
TEDVAE	0.319±0.003	0.337±0.008
DER-CFR	0.316±0.009	0.321±0.013
DESCN	0.401±0.021	0.432±0.029
ES-CFR	0.312±0.010	0.320±0.023
OURS (NEW)	**0.294±0.008**	**0.304±0.015**

the indicator function, function $\mathrm{mod}(a, b)$ returns the modulus after division of a by b and $\epsilon_{\mathrm{t}} \sim \mathcal{N}(0, 1)$. Next, we generated the continuous outcome variable $Y \in \mathbb{R}^n$ from a non-linear function as $Y = \mathrm{Sigmoid}(T + \sum_{i=1}^{d}(T \cdot X_i + (\mathbf{1}(\mathrm{mod}(i, 2) \neq 1) - \mathbf{1}(\mathrm{mod}(i, 2) \equiv 1)) \cdot (X_i + X_i^2)/d) + \epsilon_{\mathrm{y}})$, where Sigmoid denotes the sigmoid function and $\epsilon_{\mathrm{y}} \sim \mathcal{N}(0, 1)$. To introduce collider bias with strength β and implicit shadow variables into datasets, we generated the binary selection variable $S \in \mathbb{R}^n$ from a logistic function $S \sim \mathrm{Bernoulli}(1/(1 + e^{-s(\mathbf{X},T)}))$, where $s(\mathbf{X}, T) = T - \beta \cdot Y + \sum_{i=1}^{d_{\mathrm{s}}}(\mathbf{1}(\mathrm{mod}(i, 2) \equiv 1) - \mathbf{1}(\mathrm{mod}(i, 2) \neq 1)) \cdot X_i/d) + \epsilon_{\mathrm{s}}$ with $\epsilon_{\mathrm{s}} \sim \mathcal{N}(0, 1)$. Note that $d_{\mathrm{s}} \leq d$ denotes the dimension of $\mathbf{X}$ that contributes to S, and the remaining covariates not involved in the sample selection are implicit shadow variables. A unit is selected into the sample when $S = 1$, i.e., the outcome values can be observed only when $S = 1$. The ground truth CATE can be calculated easily by the above functions. Note that the results in Table 1 are under the setting of $d_{\mathrm{s}} = 0.9 \cdot d$.

C.3. Real-World Datasets Details

The IHDP dataset is from a study evaluating the effect of specialist home visits on the future cognitive test scores of premature infants (Brooksgunn et al., 1992), where confounding bias is introduced by removing a non-random subset of the treated group and using simulated outcomes to replace the original ones. To further introduce collider bias into the IHDP dataset, we set $S = 0$ for $T = 0$ units that the mother boozes and the infant's score is lower than the mean value. Intuitively, unlike the treated group, which can carefully design and regularly follow up to ensure the collection of effective test results, the control group is more likely to have sample selection bias. For those mothers with boozing problems and mothers whose children have weaker cognitive abilities, it is more likely that they will not take their children to participate in the cognitive test, resulting in collider bias. The final IHDP dataset comprises 747 units (557 selected, 190 unselected) with 25 covariates. The ground truth CATE is known because the outcomes are simulated, and both the factual and counterfactual outcomes are available.

The 2016 Atlantic Causal Inference Challenge (ACIC 2016) (Dorie et al., 2019) contains various settings of benchmark datasets with confounding bias simulated by comprehensive data generation processes. To introduce collider bias into the ACIC 2016 datasets, we used the same simulation of S as stated in Appendix C.2: $S \sim \mathrm{Bernoulli}(1/(1 + e^{-s(\mathbf{X},T)}))$, where $s(\mathbf{X}, T) = T - Y + \sum_{i=1}^{d_{\mathrm{s}}}(\mathbf{1}(\mathrm{mod}(i, 2) \equiv 1) - \mathbf{1}(\mathrm{mod}(i, 2) \neq 1)) \cdot X_i/d) + \epsilon_{\mathrm{s}}$ with $\epsilon_{\mathrm{s}} \sim \mathcal{N}(0, 1)$ and $d = 58$.

The Jobs dataset combines a randomized study based on the National Supported Work (NSW) program with observational data to form a larger confounding biased dataset that focuses on estimating the effects of a job training program on future employment situation (LaLonde, 1986; Shalit et al., 2017). To introduce collider bias into the Jobs dataset, we set $S = 0$ for $T = 0$ units that used to have a job but became unemployed. Intuitively, for those who used to have a job and have not participated in job training programs, it is more likely that they are unwilling to report their current employment situation if they lose their job, leading to collider bias. The final Jobs dataset comprises 2675 units (2494 selected, 181 unselected) with 10 covariates.

The Twins data is from a study evaluating the effect of low birth weight on the mortality of infants in their first year of life (Almond et al., 2005), where confounding bias is introduced by using simulated treatments to replace the original ones (Louizos et al., 2017; Yoon et al., 2018). To introduce collider bias into the Twins dataset, we set $S = 0$ for $T = 1$ units that

Table 5. The results of CATE estimation ($\sqrt{\text{PEHE}}$) on synthetic datasets under different d_{s}.

	$d_{\text{s}} = 0.1 \cdot d$		$d_{\text{s}} = 0.5 \cdot d$		$d_{\text{s}} = 0.9 \cdot d$	
ESTIMATOR	SELECTED DATA	UNSELECTED DATA	SELECTED DATA	UNSELECTED DATA	SELECTED DATA	UNSELECTED DATA
HECKIT	0.100±0.013	0.120±0.016	0.359±0.044	0.367±0.092	0.349±0.069	0.413±0.048
DR	0.129±0.022	0.130±0.030	0.315±0.038	0.368±0.058	0.367±0.033	0.448±0.017
IPSW	0.604±0.284	0.627±0.287	0.331±0.060	0.353±0.064	0.465±0.011	0.545±0.014
BNN	0.103±0.014	0.110±0.016	0.305±0.007	0.358±0.009	0.359±0.011	0.439±0.015
TARNET	0.105±0.015	0.106±0.021	0.307±0.056	0.360±0.056	0.366±0.071	0.436±0.087
CFR	0.104±0.005	0.105±0.017	0.307±0.041	0.358±0.055	0.359±0.008	0.436±0.030
CFOREST	0.105±0.011	0.109±0.012	0.312±0.022	0.363±0.026	0.373±0.026	0.453±0.043
DR-CFR	0.106±0.005	0.113±0.011	0.287±0.045	0.361±0.057	0.366±0.051	0.435±0.060
TEDVAE	0.227±0.018	0.257±0.021	0.283±0.052	0.378±0.059	0.394±0.054	0.431±0.067
DER-CFR	0.095±0.011	0.097±0.011	0.319±0.050	0.348±0.017	0.358±0.011	0.439±0.013
DESCN	0.107±0.002	0.109±0.002	0.311±0.004	0.367±0.004	0.365±0.003	0.449±0.011
ES-CFR	0.094±0.004	0.098±0.005	0.308±0.002	0.360±0.004	0.369±0.003	0.448±0.005
OURS	**0.085±0.001**	**0.086±0.002**	**0.228±0.006**	**0.256±0.009**	**0.299±0.008**	**0.300±0.008**

Table 6. The comparison of CATE estimation ($\sqrt{\text{PEHE}}$) between the proposed method and shadow-variable regression using correctly specified shadow variables.

VERSION OF SHADOWCATCHER	SELECTED DATA	UNSELECTED DATA
SHADOW-VARIABLE REGRESSION	0.285 ± 0.012	0.290 ± 0.015
THE PROPOSED METHOD	0.299 ± 0.008	0.300 ± 0.008

both the mother uses tobacco and the twin is alive. Intuitively, parents seldom take relatively healthy infants to the hospital, so it is more difficult to record the data of these infants, resulting in collider bias. The final Twins dataset comprises 9642 units (8804 selected, 838 unselected) with 48 covariates. The ground truth CATE is known because, for each twin pair, we observed both the case $T = 0$ (lighter twin) and $T = 1$ (heavier twin) (Yoon et al., 2018). The results are reported in Table 4.

C.4. Ablation Studies

In addition to the results stated in Section 3.2, we also conducted more experiments detailed as follows:

C.4.1. STUDIES OF THE IMPACT OF DIFFERENT NON-SHADOW-VARIABLE PROPORTIONS IN THE COVARIATES

In Section 3.2, we generated synthetic datasets to evaluate the performance of our proposed ShadowCatcher and ShadowEstimator under different strengths of collider bias, i.e., β that affects the impact of Y on S. To ensure that the strength of collider bias was only determined by β, we fixed the proportion of non-shadow variables in covariates by setting $d_{\text{s}} = 0.9 \cdot d$. Intuitively, this proportion can also determine the strength of collider bias because it affects how many covariates are involved in the sample selection. The smaller d_{s} is, the weaker the collider bias is. Therefore, we also conducted experiments under different d_{s} settings with a fixed $\beta = 5$. The results are in Table 5.

Our observations and analyses are as follows: (1) In general, the performance of all estimators gradually decreases as the proportion of non-shadow variables in covariates increases because the impact of $\mathbf{X}$ on S increases. (2) The performance of IPSW under $d_{\text{s}} = 0.1 \times d$ is abnormally poor because IPSW estimates $\mathbb{P}(S \mid \mathbf{X}, T)$ instead of the ideal $\mathbb{P}(S \mid \mathbf{X}, T, Y)$ for reweighting, the difference of which is significant when the impact of Y on S far exceeds that of $\mathbf{X}$ and T on S, leading to an inaccurate estimate. (3) The overall performance of all estimators on selected data is better than unselected data because collider bias results in $\mathbb{E}[Y \mid \mathbf{X}, T, S = 1] \neq \mathbb{E}[Y \mid \mathbf{X}, T, S = 0]$. (4) As the proportion of non-shadow variables in covariates increases, the performance gap between selected and unselected data increases because the more substantial the collider bias is, the more significant the distribution shift problem is. Especially when only one covariate is involved in the sample selection, the gap nearly disappears for most estimators. (5) Our method outperforms all baselines under all d_{s} settings, and the performance gap between selected data and unselected data, though it still exists, is much smaller than that of other baselines, which demonstrates that our proposed approaches can practically address collider bias in CATE estimation.

C.4.2. Comparison between the Proposed Method and Shadow Variable Regression with Correctly Specified Shadow Variables

To verify the validity of the generated shadow-variable representations by ShadowCatcher, we compared the proposed method with shadow-variable regression using correctly specified shadow variables (Miao & Tchetgen Tchetgen, 2016). Specifically, we conducted this ablation on the synthetic dataset in Section 3.2.1 with $d_{\mathrm{s}} = 0.9 \cdot d$, $\alpha = 1e-6$, and $\beta = 5$, where a valid shadow variable is available.

From the experimental results shown in Table 6, we can observe that while our method performs less favorably compared to shadow-variable regression with correctly specified shadow variables, which can be considered as an optimal scenario for our method, the performance of the two methods is very close. It demonstrates that the proposed method can learn valid shadow-variable representations.

C.4.3. Ablation Studies of ShadowCatcher

Ablation study of the generation phase. During the generation phase of ShadowCatcher, we make two constraints on the representations generator to ensure that the learned representations satisfy the assumptions of shadow variables. The constraint on $\mathbf{Z} \perp\!\!\!\perp S \mid \mathbf{X}, T, Y$ assumption is already guaranteed effective by the hypothesis test phase. However, the effectiveness of the constraint on $\mathbf{Z} \not\perp\!\!\!\perp Y \mid \mathbf{X}, T, S = 1$ assumption still needs to be proved. Therefore, we conducted ablation studies by comparing the performance of ShadowCatcher with and without the constraint on $\mathbf{Z} \not\perp\!\!\!\perp Y \mid \mathbf{X}, T, S = 1$. Specifically, the ablation version of ShadowCatcher optimizes the generator and the selected outcome estimator only by minimizing $\ell_{\mathrm{g_z}}$ and $\ell_{\mathrm{y_1}}$. We conducted the experiments on the synthetic dataset in Section 3.2.1 with $d_{\mathrm{s}} = 0.9 \cdot d$, $\alpha = 1e-6$, and $\beta = 1$. The results are in Table 7. The results show that the performance of this ablation version of ShadowCatcher gets worse, though still better than other baselines reported in Table 1, proving the effectiveness and necessity of the constraints in the generation phase of ShadowCatcher.

Ablation study of the hypothesis test phase. We also want to prove the necessity of the hypothesis test phase in ShadowCatcher. Therefore, we conducted ablation studies by comparing the performance of ShadowCatcher with and without the hypothesis tester. The results are in Table 7. The results show that the performance of this ablation version of ShadowCatcher is poor, and the std is very large. It proves that without the hypothesis tester that guarantees the conditional independent assumption satisfied, i.e., $\mathbf{Z} \perp\!\!\!\perp S \mid \mathbf{X}, T, Y$, the generator cannot constrain this assumption well because it uses biased predicted missing outcome values, resulting in an unstable generation.

D. Further Explanations of Some Formulas

D.1. An Explanation of Eq. (2)

In Eq. (2), the original odds ratio function is

$$
\begin{aligned}
\mathrm{OR}(\mathbf{X}, \mathbf{Z}, T, Y) =& \frac{f(Y \mid \mathbf{X}, \mathbf{Z}, T, S=0) \cdot f(Y=0 \mid \mathbf{X}, \mathbf{Z}, T, S=1)}{f(Y \mid \mathbf{X}, \mathbf{Z}, T, S=1) \cdot f(Y=0 \mid \mathbf{X}, \mathbf{Z}, T, S=0)} \\
=& \frac{f(Y \mid \mathbf{X}, \mathbf{Z}, T, S=0) \cdot f(\mathbf{X}, \mathbf{Z}, T, S=0) \cdot f(Y=0 \mid \mathbf{X}, \mathbf{Z}, T, S=1) \cdot f(\mathbf{X}, \mathbf{Z}, T, S=1)}{f(Y \mid \mathbf{X}, \mathbf{Z}, T, S=1) \cdot f(\mathbf{X}, \mathbf{Z}, T, S=1) \cdot f(Y=0 \mid \mathbf{X}, \mathbf{Z}, T, S=0) \cdot f(\mathbf{X}, \mathbf{Z}, T, S=0)} \\
=& \frac{f(Y, \mathbf{X}, \mathbf{Z}, T, S=0) \cdot f(Y=0, \mathbf{X}, \mathbf{Z}, T, S=1)}{f(Y, \mathbf{X}, \mathbf{Z}, T, S=1) \cdot f(Y=0, \mathbf{X}, \mathbf{Z}, T, S=0)} \\
=& \frac{f(Y, \mathbf{X}, \mathbf{Z}, T, S=0)/f(Y, \mathbf{X}, \mathbf{Z}, T) \cdot f(Y=0, \mathbf{X}, \mathbf{Z}, T, S=1)/f(Y=0, \mathbf{X}, \mathbf{Z}, T)}{f(Y, \mathbf{X}, \mathbf{Z}, T, S=1)/f(Y, \mathbf{X}, \mathbf{Z}, T) \cdot f(Y=0, \mathbf{X}, \mathbf{Z}, T, S=0)/f(Y=0, \mathbf{X}, \mathbf{Z}, T)} \\
=& \frac{f(S=0 \mid \mathbf{X}, \mathbf{Z}, T, Y) \cdot f(S=1 \mid \mathbf{X}, \mathbf{Z}, T, Y=0)}{f(S=0 \mid \mathbf{X}, \mathbf{Z}, T, Y=0) \cdot f(S=1 \mid \mathbf{X}, \mathbf{Z}, T, Y)}
\end{aligned}
$$

Under Assumption 2.1, because $\mathbf{Z} \perp\!\!\!\perp S \mid \mathbf{X}, T, Y$, the above equation equals $\mathrm{OR}(\mathbf{X}, T, Y)$ in Eq. (2). It indicates that the odds ratio function captures the impact of the outcome itself on the sample selection mechanism and is thus a measure of collider bias (Miao & Tchetgen Tchetgen, 2016).

Table 7. The results of CATE estimation ($\sqrt{\text{PEHE}}$) by different versions of ShadowCatcher.

VERSION OF SHADOWCATCHER	SELECTED DATA	UNSELECTED DATA
SHADOWCATCHER WITHOUT THE HYPOTHESIS TEST PHASE	0.486±0.416	0.489±0.434
SHADOWCATCHER WITHOUT THE CONSTRAINT ON $\mathbf{Z} \not\perp\!\!\!\perp Y \mid \mathbf{X}, T, S=1$	0.288±0.056	0.306±0.076
SHADOWCATCHER WITH THE CONSTRAINT ON $\mathbf{Z} \not\perp\!\!\!\perp Y \mid \mathbf{X}, T, S=1$	**0.227±0.001**	**0.229±0.001**

D.2. An Explanation of Eq. (5)

By Eq. (2), $\text{OR}(\mathbf{X}, T, Y=0) = 1$ because

$$\begin{aligned}\text{OR}(\mathbf{X}, T, Y=0) &= \frac{f(S=0 \mid \mathbf{X}, T, Y=0) \cdot f(S=1 \mid \mathbf{X}, T, Y=0)}{f(S=0 \mid \mathbf{X}, T, Y=0) \cdot f(S=1 \mid \mathbf{X}, T, Y=0)} \\ &= 1.\end{aligned}$$

Therefore, by the definition of $\widetilde{\text{OR}}(\mathbf{X}, T, Y)$ that

$$\widetilde{\text{OR}}(\mathbf{X}, T, Y) = \text{OR}(\mathbf{X}, T, Y) / \mathbb{E}[\text{OR}(\mathbf{X}, T, Y) \mid \mathbf{X}, T, S=1],$$

the right hand side of Eq. (5) equals to

$$\begin{aligned}\frac{\widetilde{\text{OR}}(\mathbf{X}, T, Y)}{\widetilde{\text{OR}}(\mathbf{X}, T, Y=0)} =& \frac{\text{OR}(\mathbf{X}, T, Y) \cdot \mathbb{E}[\text{OR}(\mathbf{X}, T, Y) \mid \mathbf{X}, T, S=1]}{\text{OR}(\mathbf{X}, T, Y=0) \cdot \mathbb{E}[\text{OR}(\mathbf{X}, T, Y) \mid \mathbf{X}, T, S=1]} \\ =& \frac{\text{OR}(\mathbf{X}, T, Y)}{\text{OR}(\mathbf{X}, T, Y=0)} \\ =& \text{OR}(\mathbf{X}, T, Y),\end{aligned}$$

which is exactly the left hand side of Eq. (5) (Miao et al., 2024).

D.3. An Explanation of ℓ_{q}

As stated in Section 1, ShadowCatcher conducts an additional hypothesis test based on Theorem 2.5 after the generation phase finishes.

Theorem 2.5 (d'Haultfoeuille, 2010). Suppose the overlap assumption and $\mathbf{Z} \not\perp\!\!\!\perp Y \mid \mathbf{X}, T, S=1$ hold, then $\mathbf{Z} \perp\!\!\!\perp S \mid \mathbf{X}, T, Y$ can be rejected if and only if there does not exist any function $Q(\cdot)$ that satisfies the following equation and takes value between $(0, 1]$:

$$\mathbb{E}\left[\frac{S}{Q(\mathbf{X}, T, Y)} - 1 \mid \mathbf{X}, \mathbf{Z}, T\right] = 0\,.$$

The tester aims to learn a solution q of $Q(\mathbf{X}, T, Y)$ in Eq. (6) that belongs to $(0, 1]$ which turns into an optimization problem by minimizing

$$\ell_{\text{q}} = \frac{1}{n}\sum_{i=1}^{n}\left\|\left(\frac{s_i}{q(\mathbf{x}_i, t_i, y_i)} - 1\right) \cdot \begin{pmatrix}\mathbf{x}_i \\ \mathbf{z}_i \\ t_i\end{pmatrix}\right\|_2^2,$$

where $q(\mathbf{x}_i, t_i, y_i)$ is a function from $\mathbb{R}$ to $(0, 1]$ and $||\cdot||_2$ denotes the ℓ_2 norm.

Specifically, if $\mathbb{E}[S/Q(\mathbf{X}, T, Y) - 1 \mid \mathbf{X}, \mathbf{Z}, T] = 0$ (by Theorem 2.5 in Eq. (6)), then

$$\mathbb{E}\left[\mathbb{E}\left[\frac{S}{Q(\mathbf{X}, T, Y)} - 1 \mid \mathbf{X}, \mathbf{Z}, T\right] \cdot \begin{pmatrix}\mathbf{X} \\ \mathbf{Z} \\ T\end{pmatrix}\right] = 0.$$

The left hand side equals to

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}\left[\frac{S}{Q(\mathbf{X},T,Y)}-1 \mid \mathbf{X},\mathbf{Z},T\right]\cdot\begin{pmatrix}\mathbf{X}\\ \mathbf{Z}\\ T\end{pmatrix}\right] =&\mathbb{E}\left[\mathbb{E}\left[\left(\frac{S}{Q(\mathbf{X},T,Y)}-1\right)\cdot\begin{pmatrix}\mathbf{X}\\ \mathbf{Z}\\ T\end{pmatrix} \mid \mathbf{X},\mathbf{Z},T\right]\right]\\ =&\mathbb{E}\left[\left(\frac{S}{Q(\mathbf{X},T,Y)}-1\right)\cdot\begin{pmatrix}\mathbf{X}\\ \mathbf{Z}\\ T\end{pmatrix}\right]=0.\end{aligned}$$

Then ℓ_{q} is just to minimize the square of the ℓ_2 norm of the last equation.