A Generative Approach for Treatment Effect Estimation under Collider Bias: From an Out-of-Distribution Perspective

Baohong Li[1] Haoxuan Li[2] Anpeng Wu[1] Minqin Zhu[1] Shiyuan Peng[3] Qingyu Cao[3] Kun Kuang[1]

Abstract

Resulting from non-random sample selection caused by both the treatment and outcome, collider bias poses a unique challenge to treatment effect estimation using observational data whose distribution differs from that of the target population. In this paper, we rethink collider bias from an out-of-distribution (OOD) perspective, considering that the entire data space of the target population consists of two different environments: The observational data selected from the target population belongs to a seen environment labeled with $S = 1$ and the missing unselected data belongs to another unseen environment labeled with $S = 0$. Based on this OOD formulation, we utilize small-scale representative data from the entire data space with no environmental labels and propose a novel method, i.e., Coupled Counterfactual Generative Adversarial Model (C^{2}GAM), to simultaneously generate the missing $S = 0$ samples in observational data and the missing S labels in the small-scale representative data. With the help of C^{2}GAM, collider bias can be addressed by combining the generated $S = 0$ samples and the observational data to estimate treatment effects. Extensive experiments on synthetic and real-world data demonstrate that plugging C^{2}GAM into existing treatment effect estimators achieves significant performance improvements.

1. Introduction

Estimating treatment effects from observational data is crucial for explanatory analysis and decision-making processes (Robins et al., 2000; Angrist & Pischke, 2009; Imbens & Wooldridge, 2009; Emdin et al., 2017). For example, accurately assessing the treatment effect of specific drugs on each patient can help doctors decide how to administer drugs to specific individuals, which is a counterfactual problem since we cannot simultaneously observe the outcomes of an individual taking or not taking the drugs. The critical challenge of estimating treatment effects is eliminating the presence of biases in the observational data (Pearl, 2009).

There are two primary sources for biases: confounding bias and selection bias (Greenland, 2003; Hernán & Robins, 2020). Let T denote the treatment variable, $\mathbf{X}$ denote the pre-treatment variables, Y denote the outcome variable, and S denote the selection indicator. The confounding bias results from common causes of treatments and outcomes ($T \leftarrow \mathbf{X} \rightarrow Y$), and the selection bias results from non-random sample selection caused by some certain variables ($T \dashrightarrow S \dashleftarrow Y$). Most of the previous works focused on addressing confounding bias (Bang & Robins, 2005; Shalit et al., 2017; Louizos et al., 2017; Wager & Athey, 2018) and selection bias caused by only T and $\mathbf{X}$ (Bareinboim & Tian, 2015; Correa et al., 2018), while ignoring collider bias which is a particular form of selection bias ($T \rightarrow S \leftarrow Y$). These methods cannot address collider bias because both T and Y cause S, which introduces spurious correlations between T and Y, resulting in biased estimation of treatment effects not only on $S = 0$ data but also on $S = 1$ data.

Collider bias can be defined as non-random sample selection conditioning on both treatments and outcomes, as shown in Figure 1(b). The observational data is non-randomly sampled from the target data distribution by the sample selection mechanism in Figure 1(b), indicated as $S = 1$, and the unobserved non-selected data is indicated as $S = 0$. In other words, only $S = 1$ samples can be observed, and for $S = 0$ data, the values of $\mathbf{X}$, T and Y are all missing. Due to collider bias, the observed data distribution will differ from the target data distribution. For example, when studying whether vaccination will protect against contracting COVID-19, where T is whether an individual is vaccinated, Y is whether an individual contracts COVID-19 and $\mathbf{X}$ is an individual's covariates like gender, age, etc., we cannot force everyone to test for COVID-19. As a result, we can only observe the data of a specific population who test for

[1] College of Computer Science and Technology, Zhejiang University, Hangzhou, China [2] Center for Data Science, Peking University, Beijing, China [3] Alibaba Group, Hangzhou, China. Correspondence to: Kun Kuang <kunkuang@zju.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

COVID-19. However, whether testing for COVID-19 is not random, people who are vaccinated and who contract COVID-19 are more willing to test, which means the sample selection is conditional on the values of T and Y, leading to collider bias. In fact, without further assumptions about the observational data, treatment effects are unidentifiable with collider bias (Correa & Bareinboim, 2017; Hernán & Robins, 2020), and thus it is necessary to introduce external unbiased data to solve collider bias.

Fortunately, we can collect small-scale representative data in real-world applications. The representative data consists of samples randomly selected from the target population. Different from the observational data, these representative samples are collected by interventions such as incentives (Askalidis et al., 2017) and multiple follow-ups (Blank & Schmidt, 2003; Tay et al., 2014) on the randomly selected units to ensure nearly complete responses (Kellerman & Herold, 2001).[1] Through the above way, collider bias is mitigated in the representative data, which means that it can be regarded that it has the same data distribution as that of the target population. However, since collecting representative data requires considerable human and material resources, it is usually carried out on a small scale to ensure quality. As a result, when estimating heterogeneous treatment effects, using only the representative data is insufficient because of the severe overfitting problem. Nevertheless, combining a small-scale representative dataset with a large-scale biased observational dataset to address collider bias is feasible.

In this paper, we present a novel formulation of collider bias as an out-of-distribution (OOD) problem, as illustrated in Figure 1. Specifically, we treat the selection indicator S as the environmental label, such that the observational data and the unselected data, respectively, come from a seen environment labeled with $S = 1$ and an unseen environment labeled with $S = 0$, and the representative data is derived from the entire data space, but with unknown environmental labels. We propose using both datasets to (1) generate the missing $S = 0$ samples in the observational dataset, (2) generate the missing S labels in the representative dataset, and (3) align the distribution of the combined generated $S = 0$ samples with the observational dataset to match that of the entire data space.

To achieve the above objectives, we propose a novel method named **C**oupled **C**ounterfactual **G**enerative **A**dversarial **M**odel, called **C^{2}GAM**, which consists of two generators that respectively generate the missing $S = 0$ samples and the missing S labels, as well as two discriminators that distinguish between the observational data and data with generated $S = 1$ labels, and between the generated unselected samples and data with generated $S = 0$ labels. By optimizing the generators using the discriminators, C^{2}GAM can effectively generate missing data while preserving the original data distribution. Combining the observational data with the unselected samples generated by C^{2}GAM, we can flexibly use any treatment effect estimation methods to achieve an accurate estimate. Extensive experiments on synthetic and real-world datasets have demonstrated the effectiveness of C^{2}GAM. By plugging C^{2}GAM into various treatment effect estimators, we have achieved significant improvement, outperforming existing state-of-the-art methods.

2. Related Works

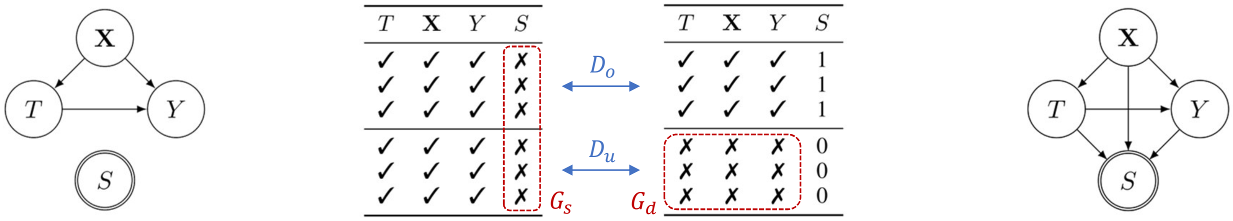
2.1. Methods for Addressing Confounding Bias

Previous works on confounding bias in observational studies include propensity-score-based, confounder balancing, tree-based, representation-learning-based, and generative-model-based methods. The propensity score defined as $\mathbb{P}(T \mid \mathbf{X})$ (Rosenbaum & Rubin, 1983) is widely used for matching (Dehejia & Wahba, 2002), reweighting (Hirano et al., 2003), and doubly robust estimation (Bang & Robins, 2005). Confounder balancing is to learn sample weights that make the confounder distributions of control and treated units similar through sample re-weighting (Hainmueller, 2012; Athey et al., 2018). Tree-based methods like Causal Forest (Wager & Athey, 2018) build a large number of causal trees and estimate heterogeneous treatment effects by taking an average of the outcomes from these causal trees. Methods based on deep representation learning learn a balanced representation of covariates, such as Treatment-Agnostic Representation Network, Balancing Neural Network (BNN) (Johansson et al., 2016), CounterFactual Regression (CFR) (Shalit et al., 2017), Disentangled Representations for CounterFactual Regression (DRCFR) (Hassanpour & Greiner, 2020), and Entire Space CounterFactual Regression (ES-CFR) (Wang et al., 2023). Generative methods include CEVAE (Louizos et al., 2017) that applies variational autoencoders to address hidden confounders, TEDVAE (Zhang et al., 2021) that simultaneously infer latent hidden variables and disentangle them, and GANITE (Yoon et al., 2018) that generates counterfactual outcomes and ITEs. Detailed discussion on the difference between our proposed method and previous generative-model-based methods is in Appendix D.

2.2. Methods for Addressing Selection Bias

The above methods cannot deal with selection bias because the distribution of the observational data differs from that of the target population. Previous works on selection bias mainly focus on sample selection caused by only $\mathbf{X}$ and T. Suppose there are variables in the causal graph that satisfy the selection-backdoor criterion. In that case, selection bias can be addressed by selection-backdoor adjustment

[1] Note that such intervention differs from that in RCTs, as interventions in RCTs means randomly assigning treatments within selective populations.



(a) Representative data without collider bias

(b) Observational data with collider bias

Figure 1. The data form and causal graphs of observation data and representative data where ✓ denotes the data is observable and ✗ denotes the data cannot be observed.

(Bareinboim & Tian, 2015; Correa & Bareinboim, 2017; Correa et al., 2018; Bareinboim et al., 2022). However, these methods cannot solve collider bias because no valid adjustment can block the non-causal path $T \to S \leftarrow Y$. In fact, without further assumptions about the observational data, treatment effects are unidentifiable with collider bias (Correa & Bareinboim, 2017; Hernán & Robins, 2020). To the best of our knowledge, there are currently no methods to solve collider bias without making further assumptions about the observational data.

When a representative dataset is also available, as in the scenario under study in this paper, data fusion methods can be applied to combine the observational and representative datasets to estimate the treatment effects of the target population. These methods aim to make the distribution of the observational dataset match the representative dataset through reweighting (Cole & Stuart, 2010; Lesko et al., 2017; Buchanan et al., 2018; Lee et al., 2023), stratification (Stuart et al., 2011; Tipton, 2013; O'Muircheartaigh & Hedges, 2014), and doubly robust estimation (Dahabreh & Hernán, 2019). Their performance relies on correct model specifications, limiting their applicability in complex real-world scenarios. Our approach, however, addresses this issue by directly learning the distribution from existing data and generating samples without collider bias.

3. Problem and Algorithm

3.1. Problem Formulation

Let $\mathcal{D} = \{\mathbf{x}_i, t_i, y_i, s_i\}_{i=1}^{n}$ be a sample population with n units independently drawn from the true target data distribution $\mathbb{P}$. For a unit i, $t_i \in \{0, 1\}$ is the binary treatment, y_i is the outcome, $\mathbf{x}_i \in \mathbb{R}^d$ is the observed pre-treatment variables with d dimensions, and $s_i \in \{0, 1\}$ is a binary selection indicator. We have a large-scale dataset of observational samples non-randomly drawn from $\mathbb{P}$, denoted as $\mathcal{D}_{\text{obs}}$, and a small-scale representative dataset $\mathcal{D}_{\text{rep}}$ containing units randomly sampled from $\mathbb{P}$. The selection indicator S in $\mathcal{D}$ indicates whether a unit i is selected into $\mathcal{D}_{\text{obs}}$, i.e., $s_i = 1$ if $\{\mathbf{x}_i, t_i, y_i\} \in \mathcal{D}_{\text{obs}}$.

Under the potential outcome framework (Imbens & Rubin, 2015), we define the potential outcomes under treatment as $Y(1)$ and under control as $Y(0)$. Our goal is to estimate the Conditional Average Treatment effect (CATE) on the target population $\mathcal{D}$, which is defined as

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

For a unit i with t_i in $\mathcal{D}$, only the factual outcome $Y(t_i)$ is available. Therefore, to make CATE identifiable, we make the following assumptions (Imbens & Rubin, 2015):

Stable Unit Treatment Value Assumption. The distribution of the potential outcome of one unit is assumed to be 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 pre-treatment variables, i.e., $Y(1), Y(0) \perp\!\!\!\perp T \mid \mathbf{X}$.

Based on the above assumptions, CATE is estimated by

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

However, because the sample selection mechanism of $\mathcal{D}_{\text{obs}}$ is not random but is jointly determined by T, $\mathbf{X}$ and Y, $\mathbb{P}_{\{\mathbf{x},t,y\}\sim\mathcal{D}_{\text{obs}}}(\mathbf{x}, t, y) \neq \mathbb{P}_{\{\mathbf{x},t,y\}\sim\mathcal{D}}(\mathbf{x}, t, y)$, i.e. $\mathbb{P}(\mathbf{X}, T, Y \mid S = 1) \neq \mathbb{P}(\mathbf{X}, T, Y)$, resulting in collider bias. Therefore, estimating the CATE on $\mathcal{D}$ using only $\mathcal{D}_{\text{obs}}$ brings the following problems:

- **Distribution shift.** Because $\mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}, T = t, S = 1] \neq \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}, T = t]$, the performance of CATE estimation models trained on $\mathcal{D}_{\text{obs}}$ degrades on the target data $\mathcal{D}$.
- **Biased estimation.** $T \to S \leftarrow Y$ causes a spurious correlation between T and Y, leading to a biased estimate of CATE using only $\mathcal{D}_{\text{obs}}$.

Using only $\mathcal{D}_{\text{rep}}$ to estimate the CATE on $\mathcal{D}$ is not applicable either because the sample size of $\mathcal{D}_{\text{rep}}$ is too small, which makes estimators suffer from the severe overfitting problem. Therefore, we need to leverage both $\mathcal{D}_{\text{obs}}$ and $\mathcal{D}_{\text{rep}}$ to help solve collider bias.

3.2. Motivation

To address collider bias, we formulate it as an OOD problem (Zhang et al., 2023; 2024). As shown in Figure 1, in $\mathcal{D}_{\text{obs}}$, the non-random sample selection caused by collider bias mainly results in missing $S = 0$ data, i.e., the unselected sub-population of the target population is not available; In $\mathcal{D}_{\text{rep}}$, the selection indicators S are unknown, i.e., we cannot tell the probability that a unit in $\mathcal{D}_{\text{rep}}$ would be selected into $\mathcal{D}_{\text{obs}}$ in observational studies. Therefore, we consider S as the environmental labels. In this way, the observational data can be regarded as samples from a seen environment labeled with $S = 1$, the missing unselected data can be regarded as samples from an unseen environment labeled with $S = 0$, and the representative data can be regarded as samples from the entire data space but the environmental labels are unknown. From an OOD perspective, we wish to recover the distribution of $\mathcal{D}$ from $\mathcal{D}_{\text{obs}}$ and $\mathcal{D}_{\text{rep}}$ as much as possible, which means we need to recover the missing parts of $\mathcal{D}_{\text{obs}}$ and $\mathcal{D}_{\text{rep}}$ by two generators respectively:

- **Unselected samples generator** $\mathrm{G_d}$**.** It generates the missing $S = 0$ data in $\mathcal{D}_{\text{obs}}$ from random noises Z. The generated samples are denoted as $\mathcal{D}_{\text{gen}}$
- **Selection indicator generator** $\mathrm{G_s}$**.** It generates the missing S labels of data in $\mathcal{D}_{\text{rep}}$ from the corresponding $\mathbf{X}, T$ and Y.

To optimize the above generators, we need (1) a discriminator to align the distribution of the data with generated $S = 1$ labels and that of the ground truth $S = 1$ data and (2) a discriminator to align the distribution of the generated $S = 0$ data and that of the ground truth $S = 0$ data. Since the latter involves ground truth $S = 0$ data, which is not available in $\mathcal{D}_{\text{obs}}$, we use data in $\mathcal{D}_{\text{rep}}$ with $S = 0$ labels generated by $\mathrm{G_s}$ as an approximation. As a result, the two discriminators perform the following tasks respectively:

- **Selected data discriminator** $\mathrm{D_o}$**.** It makes the distribution of $\mathcal{D}_{\text{obs}}$ the same as that of data in $\mathcal{D}_{\text{rep}}$ with $S = 1$ labels generated by $\mathrm{G_s}$.
- **Unselected data discriminator** $\mathrm{D_u}$**.** It makes the distribution of the $S = 0$ samples generated by $\mathrm{G_d}$ the same as that of data in $\mathcal{D}_{\text{rep}}$ with $S = 0$ labels generated by $\mathrm{G_s}$.

To further ensure that the distribution of the combination of the generated $S = 0$ samples and $\mathcal{D}_{\text{obs}}$ is the same as that of $\mathcal{D}$, we need an additional constraint to make $\mathbb{P}_{\{s\}\sim\mathcal{D}}(s) = \mathbb{P}_{\{\mathbf{x},t,y\}\sim\mathcal{D}_{\text{rep}}}(\mathrm{G_s}(\mathbf{x}, t, y))$. To accomplish this, we make the ratio of the generated samples to the observational data the same as that of the representative data with generated $S = 0$ labels to the representative data with generated $S = 1$ labels. We also constrain $\mathrm{G_d}$ to minimize the distance between $\mathbb{P}_{\{\mathbf{x},t,y\}\sim\mathcal{D}_{\text{rep}}}(\mathbf{x}, t, y)$ and $\mathbb{P}_{\{\mathbf{x},t,y\}\sim\mathcal{D}_{\text{obs}}\cup\mathcal{D}_{\text{gen}}}(\mathbf{x}, t, y)$ during optimization.

Such coupled design is reasonable for the following reasons: the distribution of the ground truth $S = 1$ data is known, i.e., the observational data is available, hence $\mathrm{D_o}$ can assist $\mathrm{G_s}$ in generating correct $S = 1$ labels for data in $\mathcal{D}_{\text{rep}}$. Additionally, with the constraint on the ratio of $S = 0$ data to $S = 1$ data, $\mathrm{D_u}$ can make $\mathrm{G_d}$ generate $S = 0$ samples with the same distribution as the representative dataset labeled with $S = 0$ by $\mathrm{G_s}$. Consequently, the combination of $\mathrm{G_d}$ generated $S = 0$ samples with the original $S = 1$ observational data results in a distribution consistent with that of the representative dataset.

By jointly optimizing the two generators and the two discriminators with the above constraints, we can achieve the objective of recovering the distribution of $\mathcal{D}$ by combining the generated $S = 0$ samples with the original observational data, which can be used as the training data for any treatment effect estimation methods to achieve a better CATE estimate. Naturally, a Generative Adversarial Nets (GAN) (Goodfellow et al., 2014) framework is suitable for this task.

3.3. C^{2}GAM: Coupled Counterfactual Generative Adversarial Model

Based on the above motivation, we propose a novel method named Coupled Counterfactual Generative Adversarial Model (C^{2}GAM), as shown in Figure 2. C^{2}GAM consists of two generators $\mathrm{G_d}$ and $\mathrm{G_s}$ and two discriminators $\mathrm{D_o}$ and $\mathrm{D_u}$, as mentioned earlier. The details are as follows:

Selection indicator generator $\mathrm{G_s}$**.** This generator aims to generate selection indicators S for data in $\mathcal{D}_{\text{rep}}$. It takes $(\mathbf{x}, t, y) \sim \mathcal{D}_{\text{rep}} \cup \mathcal{D}_{\text{gen}}$ as inputs to generate the corresponding S labels, denoted as $\mathrm{G_s}(\mathbf{x}, t, y)$. The objective is to optimize $\mathrm{G_s}$ to maximize the probability of correctly labeling the data with S ,i.e., to make $\mathbb{P}_{\{\mathbf{x},t,y\}\sim\mathcal{D}_{\text{rep}}}(\mathbf{x}, t, y, \mathrm{G_s}(\mathbf{x}, t, y)) = \mathbb{P}_{\{\mathbf{x},t,y,s\}\sim\mathcal{D}}(\mathbf{x}, t, y, s)$.

Unselected samples generator $\mathrm{G_d}$**.** This generator aims to generate samples whose distribution is the same as that of $S = 0$ data in $\mathcal{D}$. It takes random noises $Z = \{z_i \sim \mathcal{N}(0, 1)\}_{i=1}^{n_{\text{gen}}}$ as inputs to generate $\mathcal{D}_{\text{gen}} = \{\mathbf{x}_i, t_i, y_i\}_{i=1}^{n_{\text{gen}}}$, denoted as $\{\mathrm{G_d}(z_i)\}_{i=1}^{n_{\text{gen}}}$, where n_{gen} is the size of the generated samples. The objective is to optimize $\mathrm{G_d}$ to make $\mathbb{P}_{\{\mathbf{x},t,y\}\sim\mathcal{D}_{\text{gen}}\cup\mathcal{D}_{\text{obs}}}(\mathbf{x}, t, y) = \mathbb{P}_{\{\mathbf{x},t,y\}\sim\mathcal{D}}(\mathbf{x}, t, y)$, i.e., to minimize the distance between $\mathbb{P}_{z\sim\mathcal{N}(0,1)^{n_{\text{gen}}}}(\mathrm{G_d}(z))$ and $\mathbb{P}_{\{\mathbf{x},t,y,s\}\sim\mathcal{D}}(\mathbf{x}, t, y \mid s = 0)$.

Selected data discriminator $\mathrm{D_o}$**.** This discriminator aims to discriminate between data with generated $S = 1$ labels in $\mathcal{D}_{\text{rep}}$ and the observed $S = 1$ data in $\mathcal{D}_{\text{obs}}$. We regard $\mathcal{D}_{\text{obs}}$

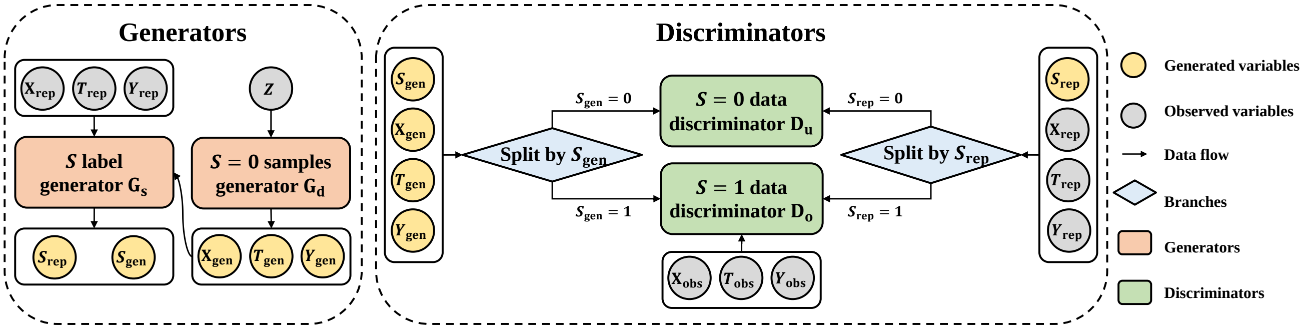


Figure 2. Overview of the architecture of C^{2}GAM.

as the original dataset, and data in $\mathcal{D}_{\text{rep}}$ and $\mathcal{D}_{\text{gen}}$ labeled with $S = 1$ by G_s as the generated dataset. Therefore, D_o takes $(\mathbf{x}, t, y) \sim \mathcal{D}_{\text{obs}} \cup ((\mathcal{D}_{\text{rep}} \cup \mathcal{D}_{\text{gen}}) \mid \text{G}_\text{s}(\mathbf{x}, t, y) = 1)$ as inputs and returns the probability that $(\mathbf{x}, t, y)$ is from $\mathcal{D}_{\text{obs}}$, denoted as $\text{D}_\text{o}(\mathbf{x}, t, y)$. The objective is to optimize D_o to maximize the probability of correctly determining whether a sample comes from $\mathcal{D}_{\text{obs}}$ or $(\mathcal{D}_{\text{rep}} \cup \mathcal{D}_{\text{gen}}) \mid \text{G}_\text{s}(\mathbf{x}, t, y) = 1$. The objective function is

$$\begin{aligned}
&\min_{\text{G}_\text{s}, \text{G}_\text{d}} \max_{\text{D}_\text{o}} \mathbb{E}_{\{\mathbf{x},t,y\}\sim\mathcal{D}_{\text{obs}}}[\log(\text{D}_\text{o}(\mathbf{x}, t, y))] \\
&+ \mathbb{E}_{\{\mathbf{x},t,y\}\sim\mathcal{D}_{\text{rep}}}[\text{G}_\text{s}(\mathbf{x}, t, y) \cdot \log(1 - \text{D}_\text{o}(\mathbf{x}, t, y))] \\
&+ \mathbb{E}_{z\sim\mathcal{N}(0,1)^{n_{\text{gen}}}}[\text{G}_\text{s}(\text{G}_\text{d}(z)) \cdot \log(1 - \text{D}_\text{o}(\text{G}_\text{d}(z)))].
\end{aligned}$$

Unselected data discriminator D_u. This discriminator aims to discriminate between the generated $S = 0$ samples and the $S = 0$ data in $\mathcal{D}$. However, since $S = 0$ data in $\mathcal{D}$ is not available, we can only use data with generated $S = 0$ labels as approximations. We regard data from $\mathcal{D}_{\text{gen}}$ and $\mathcal{D}_{\text{rep}}$ labeled with $S = 0$ by G_s as the original dataset and all data from $\mathcal{D}_{\text{gen}}$ as the generated dataset. Therefore, D_u takes $(\mathbf{x}, t, y) \sim \mathcal{D}_{\text{gen}} \cup (\mathcal{D}_{\text{rep}} \mid \text{G}_\text{s}(\mathbf{x}, t, y) = 0)$ as inputs and returns the probability that $(\mathbf{x}, t, y)$ is from $(\mathcal{D}_{\text{gen}} \cup \mathcal{D}_{\text{rep}}) \mid \text{G}_\text{s}(\mathbf{x}, t, y) = 0$, denoted as $\text{D}_\text{u}(\mathbf{x}, t, y)$. The objective is to optimize D_u to maximize the probability of correctly determining whether a sample comes from $(\mathcal{D}_{\text{gen}} \cup \mathcal{D}_{\text{rep}}) \mid \text{G}_\text{s}(\mathbf{x}, t, y) = 0$ or $\mathcal{D}_{\text{gen}}$. The objective function is

$$\begin{aligned}
&\min_{\text{G}_\text{s}, \text{G}_\text{d}} \max_{\text{D}_\text{u}} \mathbb{E}_{z\sim\mathcal{N}(0,1)^{n_{\text{gen}}}}[\log(1 - \text{D}_\text{u}(\text{G}_\text{d}(z))] \\
&+ \mathbb{E}_{z\sim\mathcal{N}(0,1)^{n_{\text{gen}}}}[(1 - \text{G}_\text{s}(\text{G}_\text{d}(z))) \cdot \log(\text{D}_\text{u}(\text{G}_\text{d}(z)))] \\
&+ \mathbb{E}_{\{\mathbf{x},t,y\}\sim\mathcal{D}_{\text{rep}}}[(1 - \text{G}_\text{s}(\mathbf{x}, t, y)) \cdot \log(\text{D}_\text{u}(\mathbf{x}, t, y))].
\end{aligned}$$

Following (Goodfellow et al., 2014), with the above objective functions, the discriminators D_o and D_u and the generators G_s and G_d can be iteratively optimized using mini-batch gradient descent. In each batch, we first fix the parameters of both generators to optimize both discriminators simultaneously, then fix the parameters of both discriminators to optimize both generators simultaneously.

Specifically, when fixing the parameters of the generators, the discriminators are optimized using the loss function being $\mathcal{L}_{\text{D}_\text{o}} + \mathcal{L}_{\text{D}_\text{u}}$, where

$$\begin{aligned}
\mathcal{L}_{\text{D}_\text{o}} &= -(\mathbb{E}_{\{\mathbf{x},t,y\}\sim\mathcal{D}_{\text{obs}}}[\log(\text{D}_\text{o}(\mathbf{x}, t, y))] \\
&\quad + \mathbb{E}_{\{\mathbf{x},t,y\}\sim\mathcal{D}_{\text{rep}}}[\text{G}_\text{s}(\mathbf{x}, t, y) \cdot \log(1 - \text{D}_\text{o}(\mathbf{x}, t, y))] \\
&\quad + \mathbb{E}_{z\sim\mathcal{N}(0,1)^{n_{\text{gen}}}}[\text{G}_\text{s}(\text{G}_\text{d}(z)) \cdot \log(1 - \text{D}_\text{o}(\text{G}_\text{d}(z)))]), \\
\mathcal{L}_{\text{D}_\text{u}} &= -(\mathbb{E}_{z\sim\mathcal{N}(0,1)^{n_{\text{gen}}}}[\log(1 - \text{D}_\text{u}(\text{G}_\text{d}(z))] \\
&\quad + \mathbb{E}_{z\sim\mathcal{N}(0,1)^{n_{\text{gen}}}}[(1 - \text{G}_\text{s}(\text{G}_\text{d}(z))) \cdot \log(\text{D}_\text{u}(\text{G}_\text{d}(z)))] \\
&\quad + \mathbb{E}_{\{\mathbf{x},t,y\}\sim\mathcal{D}_{\text{rep}}}[(1 - \text{G}_\text{s}(\mathbf{x}, t, y)) \cdot \log(\text{D}_\text{u}(\mathbf{x}, t, y))]).
\end{aligned}$$

Given the parameters of the discriminators, the two generators are optimized using the loss function being $\mathcal{L}_{\text{G}_\text{s}} + \mathcal{L}_{\text{G}_\text{d}}$, where

$$\begin{aligned}
\mathcal{L}_{\text{G}_\text{s}} &= \mathbb{E}_{z\sim\mathcal{N}(0,1)^{n_{\text{gen}}}}[(1 - \text{G}_\text{s}(\text{G}_\text{d}(z))) \cdot \log(\text{D}_\text{u}(\text{G}_\text{d}(z)))] \\
&\quad + \mathbb{E}_{\{\mathbf{x},t,y\}\sim\mathcal{D}_{\text{rep}}}[\text{G}_\text{s}(\mathbf{x}, t, y) \cdot \log(1 - \text{D}_\text{o}(\mathbf{x}, t, y))] \\
&\quad + \mathbb{E}_{\{\mathbf{x},t,y\}\sim\mathcal{D}_{\text{rep}}}[(1 - \text{G}_\text{s}(\mathbf{x}, t, y)) \cdot \log(\text{D}_\text{u}(\mathbf{x}, t, y))] \\
&\quad + \mathbb{E}_{z\sim\mathcal{N}(0,1)^{n_{\text{gen}}}}[\text{G}_\text{s}(\text{G}_\text{d}(z)) \cdot \log(1 - \text{D}_\text{o}(\text{G}_\text{d}(z)))], \\
\mathcal{L}_{\text{G}_\text{d}} &= \mathbb{E}_{z\sim\mathcal{N}(0,1)^{n_{\text{gen}}}}[\text{G}_\text{s}(\text{G}_\text{d}(z)) \cdot \log(1 - \text{D}_\text{o}(\text{G}_\text{d}(z)))] \\
&\quad + \mathbb{E}_{z\sim\mathcal{N}(0,1)^{n_{\text{gen}}}}[\log(1 - \text{D}_\text{u}(\text{G}_\text{d}(z))] \\
&\quad + \mathbb{E}_{z\sim\mathcal{N}(0,1)^{n_{\text{gen}}}}[(1 - \text{G}_\text{s}(\text{G}_\text{d}(z))) \cdot \log(\text{D}_\text{u}(\text{G}_\text{d}(z)))].
\end{aligned}$$

We iteratively optimize the discriminators and the generators and update n_{gen} with $n_{\text{obs}} \cdot n_0/n_1$, where n_{obs} is the sample size of $\mathcal{D}_{\text{obs}}$, n_0 and n_1 is the count of units in $\mathcal{D}_{\text{rep}}$ with $\text{G}_\text{s}(\mathbf{x}, t, y) = 0$ and $\text{G}_\text{s}(\mathbf{x}, t, y) = 1$, respectively. Note that n_{gen} is initialized as $\mathcal{D}_{\text{obs}}$. At the end of each iteration, we compute the Wasserstein distance (Cuturi & Doucet, 2014) between $\mathbb{P}_{\{\mathbf{x},t,y\}\sim\mathcal{D}_{\text{rep}}}(\mathbf{x}, t, y)$ and $\mathbb{P}_{\{\mathbf{x},t,y\}\sim\mathcal{D}_{\text{obs}}\cup\mathcal{D}_{\text{gen}}}(\mathbf{x}, t, y)$ and optimize G_d to minimize this distance. Iterations terminate when this distance falls below a predefined threshold, or the maximum number of iterations reaches. Combining the generated samples $\mathcal{D}_{\text{gen}}$ and the observational data $\mathcal{D}_{\text{obs}}$, we then fit any CATE estimator to achieve CATE estimation. The pseudo-code of C^{2}GAM is in Appendix A, and the source code is available at `https://github.com/ZJUBaohongLi/C2GAM`.

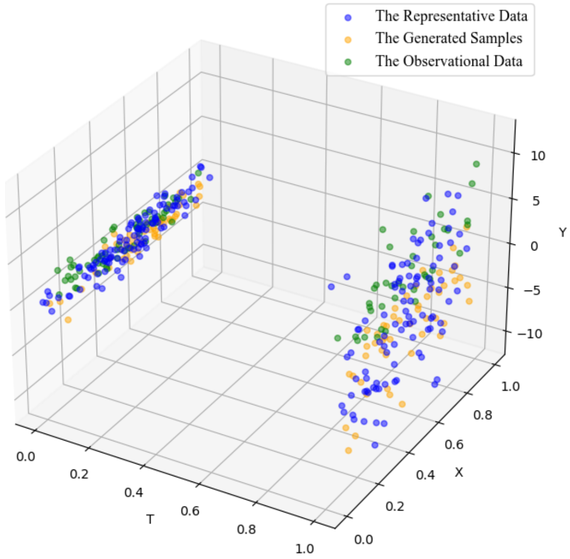


Figure 3. The scatter plot showing the joint distribution of T, $\mathbf{X}$, and Y across $\mathcal{D}_{\text{rep}}$, $\mathcal{D}_{\text{obs}}$, and $\mathcal{D}_{\text{gen}}$.

Table 1. The results (mean ± std of $\sqrt{\text{PEHE}}$) of CATE estimation on synthetic data. The best results are in bold, and the second-best ones are underlined.

METHOD	$S=1$ SAMPLES	$S=0$ SAMPLES
DOUBLY ROBUST	7.410±4.602	8.496±2.995
CAUSAL FOREST	4.929±0.073	6.153±0.074
CEVAE	4.051±0.047	5.296±0.026
GANITE	4.139±0.071	4.997±0.148
TEDVAE	4.003±0.040	5.318±0.049
BNN	2.893±0.427	3.196±0.360
TARNET	2.023±0.223	2.830±0.261
CFR	2.035±0.054	2.923±0.077
DRCFR	2.107±0.307	2.850±0.428
ES-CFR	4.993±0.085	5.707±0.124
IPSW	2.055±0.242	2.840±0.378
AIPSW	1.939±0.296	2.669±0.422
CW	1.867±0.067	2.581±0.109
C^{2}GAM+BNN	**0.977±0.078**	1.201±0.111
C^{2}GAM+CFR	0.984±0.074	**1.083±0.086**
C^{2}GAM+DRCFR	1.175±0.072	1.285±0.120

4. Experiments

4.1. Baselines

To evaluate the effectiveness of the proposed method, we first utilized C^{2}GAM to generate $S=0$ samples, then employed the combination of the observational data and the generated samples to fit CATE estimators, including BNN (Johansson et al., 2016), CFR (Shalit et al., 2017), and DRCFR (Hassanpour & Greiner, 2020). We compare their estimation results against those of the following baselines without using samples generated by C^{2}GAM: (1) **Statistical methods**, including Doubly Robust (Bang & Robins, 2005) and Causal Forest (Wager & Athey, 2018); (2) **Generative methods**, including Causal Effect Variational Autoencoder (CEVAE) (Louizos et al., 2017), Generative Adversarial Nets for inference of Individualized Treatment Effects (GANITE) (Yoon et al., 2018), and TEDVAE (Zhang et al., 2021). (3) **Representation learning methods**, including BNN, TARNet, CFR, DRCFR, and Entire Space CounterFactual Regression (ES-CFR) (Wang et al., 2023). We also compare C^{2}GAM with several commonly used **data fusion methods**, including Inverse Probability of Sampling Weighting (IPSW) (Cole & Stuart, 2010), Augmented IPSW (AIPSW) (Dahabreh & Hernán, 2019), and Calibration Weighting (CW) (Lee et al., 2023).

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 above methods, where $\text{PEHE} = \frac{1}{N}\sum_{i=1}^{N}((\hat{y}_i(1)-\hat{y}_i(0))-(y_i(1)-y_i(0))^2$. Since the data fusion baselines, i.e., IPSW, AIPSW, and CW, merely adjust the distributions of the two datasets without directly estimating the CATE, we report the results of fitting DRCFR using the combined data adjusted by these methods to evaluate their CATE estimation performance. We used the Wasserstein distance as the Integral Probability Metric (IPM) to implement BNN, CFR, DRCFR, and ES-CFR. We implemented the estimators in the PyTorch environment with Python 3.9, with the CPU being 13th Gen Intel(R) Core(TM) i7-13700K and the GPU being NVIDIA GeForce RTX 3080 with CUDA 12.1. We split each dataset into 60/20/20 train/validation/test datasets. In each experimental setting, we performed 20 replications and recorded the mean and standard deviation (std) of $\sqrt{\text{PEHE}}$ on $S=1$ and $S=0$ data.

4.2. Experiments on Synthetic Data

4.2.1. DATASETS

In order to evaluate the effectiveness of our method against collider bias, in each experiment, we generated a large-scale collider-biased observational dataset and a small-scale representative dataset without collider bias using the same outcome model. Specifically, we first generated continuous pre-treatment variables $\mathbf{X} \in \mathbb{R}^{n\times d}$ with independent Gaussian distributions as $\mathbf{X} \sim \mathcal{N}(\mathbf{0},\mathbf{1})$, where $d=20$. Subsequently, We generated binary treatments $T \in \mathbb{R}^n$ from a logistic function as $T \sim \text{Bernoulli}(1/(1+e^{-t(\mathbf{X})}))$, where $\text{Bernoulli}(\cdot)$ denotes the Bernoulli distribution, $t(\mathbf{X}) = \sum_{i=1}^{d}(\mathbf{1}(\mathrm{mod}(i,2)\equiv 1)-\mathbf{1}(\mathrm{mod}(i,2)\neq 1))\cdot(\mathrm{mod}(i,2)+1)\cdot X_i/d)+\epsilon_{\mathrm{t}}$, $\mathbf{1}(\cdot)$ is 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 gen-

Table 2. The results (mean ± std of $\sqrt{\text{PEHE}}$) under different n_{rep}. The best results in each group are in bold.

	$n_{\text{obs}} : n_{\text{rep}} = 10000 : 500$		$n_{\text{obs}} : n_{\text{rep}} = 10000 : 200$		$n_{\text{obs}} : n_{\text{rep}} = 10000 : 100$	
DATA+ESTIMATOR	$S = 1$ SAMPLES	$S = 0$ SAMPLES	$S = 1$ SAMPLES	$S = 0$ SAMPLES	$S = 1$ SAMPLES	$S = 0$ SAMPLES
$\mathcal{D}_{\text{rep}}$+BNN	1.776±0.165	1.753±0.152	3.458±0.232	3.129±0.189	4.696±0.704	3.779±0.259
$\mathcal{D}_{\text{obs}}$+BNN	2.893±0.427	3.196±0.360	2.889±0.488	3.167±0.201	2.934±0.926	3.448±0.664
$\mathcal{D}_{\text{obs}}$&$\mathcal{D}_{\text{rep}}$+BNN	2.705±0.302	3.250±0.200	2.958±0.531	3.217±0.163	2.993±0.736	3.398±0.417
C^{2}GAM+BNN	**0.977±0.078**	**1.201±0.111**	**1.367±0.079**	**1.800±0.113**	**1.404±0.227**	**1.770±0.297**
$\mathcal{D}_{\text{rep}}$+CFR	1.551±0.079	1.584±0.068	2.422±0.170	2.265±0.174	4.022±0.182	3.618±0.256
$\mathcal{D}_{\text{obs}}$+CFR	2.023±0.223	2.830±0.261	1.903±0.237	2.751±0.382	1.813±0.221	2.580±0.397
$\mathcal{D}_{\text{obs}}$&$\mathcal{D}_{\text{rep}}$+CFR	2.110±0.266	2.962±0.434	2.785±0.586	3.334±0.456	2.169±0.250	3.045±0.473
C^{2}GAM+CFR	**0.984±0.074**	**1.083±0.086**	**1.090±0.101**	**1.263±0.119**	**1.444±0.184**	**1.794±0.377**
$\mathcal{D}_{\text{rep}}$+DRCFR	1.448±0.099	1.434±0.136	2.224±0.313	2.130±0.338	3.752±0.338	3.437±0.360
$\mathcal{D}_{\text{obs}}$+DRCFR	2.107±0.307	2.850±0.428	2.290±0.294	3.144±0.462	2.315±0.398	3.090±0.480
$\mathcal{D}_{\text{obs}}$&$\mathcal{D}_{\text{rep}}$+DRCFR	2.090±0.536	2.685±0.864	2.095±0.378	2.793±0.698	2.185±0.464	2.823±0.730
C^{2}GAM+DRCFR	**1.175±0.072**	**1.285±0.120**	**1.339±0.108**	**1.602±0.187**	**1.464±0.139**	**1.740±0.174**

erated continuous outcomes $Y \in \mathbb{R}^n$ from a non-linear function as $Y = T + \sum_{i=1}^{d}(T \cdot X_i + (\mathbf{1}(\text{mod}(i,2) \neq 1) - \mathbf{1}(\text{mod}(i,2) \equiv 1)) \cdot (\text{mod}(i,2)+1) \cdot (X_i + X_i^2)/d) + \epsilon_{\text{y}}$, where $\epsilon_{\text{y}} \sim \mathcal{N}(0,1)$. Following the above data generation process, we generated a target dataset. The representative dataset was then created by randomly selecting n_{rep} samples from the target dataset, where $n_{\text{rep}} \in \{100, 200, 500\}$. To further introduce collider bias for generating the observational data, we non-randomly sampled the target dataset by a binary selection variable $S \in \mathbb{R}^n$, which came from a logistic function as $S \sim \text{Bernoulli}(1/(1 + e^{-s(\mathbf{X},T)}))$, where $s(\mathbf{X}, T) = Y - 3 \cdot T + \sum_{i=1}^{d}(\mathbf{1}(\text{mod}(i,2) \equiv 1) - \mathbf{1}(\text{mod}(i,2) \neq 1)) \cdot X_i/d) + \epsilon_{\text{s}}$, $\epsilon_{\text{s}} \sim \mathcal{N}(0,1)$ and a unit was selected into the sample only when $S = 1$. The final observational dataset comprised $n_{\text{obs}} = 10000$ samples.

4.2.2. RESULTS

First, we aim to investigate the performance of the baselines facing collider bias and evaluate whether our proposed approach can work well against collider bias. We compare the CATE estimation results of the proposed method against those of the baselines as shown in Table 1. In this table, we only report the results under the setting of $n_{\text{rep}} = 500$ for clarity, and more results are shown in Table 2.

From the results, we have the following observations: (1) For all the estimators, the overall performance on $S = 0$ samples is worse than that on $S = 1$ samples because of the distribution shift problem caused by collider bias. (2) Doubly Robust and Causal Forest show the worst performance among all estimators because they employ multiple models susceptible to collider bias, resulting in poor cumulative performance of the final results. (3) The performance of generative methods, including CEVAE, GANITE, and TEDVAE, is poor because their generated results are based on the collider-biased dataset, which reduces the confounding bias but amplifies the collider bias. (4) Representation learning methods, including BNN, TARNet, CFR, and DR-CFR, perform better than the above baselines since they neither accumulate nor amplify the collider bias. However, the estimation errors are still significant because they can only address confounding bias. (5) The data fusion methods outperformed all baselines, as they address selection bias directly. However, their performance still falls short compared to our approach. It demonstrates that introducing a generative model to learn the target population's distribution and generate unbiased data is important and helpful. (6) Applying C^{2}GAM to the estimators achieves noticeable performance improvement compared to all the baselines. It demonstrates that our approach can practically address collider bias and achieve more accurate CATE estimation.

Second, we aim to further evaluate the effectiveness of our method under different n_{rep}. Meanwhile, we are also interested in comparing the performance of the CATE estimators across four scenarios: (1) using only $\mathcal{D}_{\text{rep}}$ for fitting, (2) using only $\mathcal{D}_{\text{obs}}$ for fitting, (3) combining $\mathcal{D}_{\text{rep}}$ and $\mathcal{D}_{\text{obs}}$ for fitting, and (4) combining $\mathcal{D}_{\text{obs}}$ and samples generated by C^{2}GAM for fitting. We report the CATE estimation results of BNN, CFR, and DRCFR under the above scenarios with $n_{\text{rep}} \in \{100, 200, 500\}$ in Table 2.

From the results, we have the following observations: (1) The performance of training solely on $\mathcal{D}_{\text{obs}}$ is the poorest. The reason is that while using collider-biased observational data alone may exhibit superior performance in predicting factual outcomes due to overfitting, our evaluation focuses on estimating the CATE, which necessitates accurate predictions of both factual and counterfactual outcomes. Thus, because of the spurious correlation introduced by collider bias, relying solely on biased observational data for CATE estimation would result in inaccurate counterfactual outcome predictions and consequently harm the CATE estimation. (2) Using only $\mathcal{D}_{\text{rep}}$ achieve better performance than other scenarios except for using C^{2}GAM when $n_{\text{rep}} = 500$ be-

Table 3. The results (mean ± std of $\sqrt{\text{PEHE}}$) of treatment effect estimation on real-world datasets. The best results are in bold and the second best ones are underlined.

	IHDP		TWINS		ACIC	
METHOD	$S=1$ SAMPLES	$S=0$ SAMPLES	$S=1$ SAMPLES	$S=0$ SAMPLES	$S=1$ SAMPLES	$S=0$ SAMPLES
USING $\mathcal{D}_{\text{rep}}$	2.035±0.478	2.120±0.720	0.327±0.030	0.330±0.022	3.933±0.407	3.655±0.232
DOUBLY ROBUST	1.391±0.288	1.630±0.342	0.485±0.052	0.529±0.031	2.540±0.141	2.931±0.359
CAUSAL FOREST	1.305±0.095	1.490±0.114	0.378±0.021	0.421±0.010	4.178±0.141	4.388±0.121
CEVAE	3.078±0.129	4.397±0.140	0.512±0.031	0.537±0.043	4.328±0.261	5.442±0.202
GANITE	3.063±0.158	3.160±0.382	0.329±0.022	0.331±0.061	3.782±0.072	4.908±0.079
TEDVAE	4.143±0.022	4.154±0.037	0.435±0.038	0.438±0.031	5.004±0.166	6.130±0.186
BNN	1.970±0.465	2.086±0.441	0.332±0.014	0.384±0.054	2.320±0.724	2.773±0.556
TARNET	2.124±0.260	2.147±0.225	0.532±0.074	0.534±0.086	2.127±0.201	2.938±0.526
CFR	2.278±0.306	2.405±0.345	0.435±0.038	0.438±0.031	2.324±0.525	2.754±0.325
DRCFR	2.267±0.339	2.413±0.370	0.388±0.028	0.396±0.033	2.146±0.203	2.567±0.268
ES-CFR	4.058±0.025	4.248±0.035	0.391±0.027	0.660±0.100	3.826±0.169	4.627±0.062
IPSW	1.563±0.195	1.703±0.217	0.456±0.123	0.452±0.129	3.332±0.382	3.508±0.246
AIPSW	1.501± 0.195	1.714±0.354	0.413±0.043	0.441±0.106	3.343±0.236	3.345±0.347
CW	1.427±0.109	1.496±0.115	0.370±0.039	0.381±0.069	1.958±0.155	2.376±0.175
C^{2}GAM+BNN	**1.042±0.113**	1.276±0.184	0.303±0.014	0.310±0.022	1.938±0.322	2.441±0.209
C^{2}GAM+CFR	1.203±0.148	1.235±0.165	0.311±0.050	0.328±0.022	2.015±0.550	1.917±0.583
C^{2}GAM+DRCFR	1.064±0.120	**0.979±0.282**	**0.296±0.021**	**0.301±0.030**	**1.675±0.299**	**1.876±0.410**

cause $\mathcal{D}_{\text{rep}}$ has no collider bias. However, as n_{rep} decreases, the performance gets worse due to severe overfitting. (3) Combining $\mathcal{D}_{\text{rep}}$ and $\mathcal{D}_{\text{obs}}$ does not show performance improvement and even hurts the performance of CFR compared to simply using $\mathcal{D}_{\text{obs}}$, mainly for two reasons. The first reason is that n_{rep} is much smaller than n_{obs}, so the contribution of n_{rep} is minimal. The second reason is that though $\mathcal{D}_{\text{rep}}$ is the same as the target population, the combination of $\mathcal{D}_{\text{rep}}$ and $\mathcal{D}_{\text{obs}}$, on the contrary, is no longer the target population. Therefore, such a combination only brings a few $S=0$ samples for fitting but cannot address collider bias. (4) The performance of employing C^{2}GAM achieves the best performance under all settings. Moreover, the performance improvement is still significant even under $n_{\text{rep}}=100$. It proves the robustness of our method in scenarios where the representative data is hard to obtain. We provide more ablation studies that demonstrate the necessity of each module in C^{2}GAM in Appendix C.

Finally, we want to verify whether C^{2}GAM achieves its objective, i.e., to generate $S=0$ samples $\mathcal{D}_{\text{gen}}$ such that combining with the observational data $\mathcal{D}_{\text{obs}}$, the distribution of which is the same as that of $\mathcal{D}$. Therefore, we visualize the distribution of $\mathcal{D}_{\text{gen}}$, $\mathcal{D}_{\text{obs}}$, and $\mathcal{D}_{\text{rep}}$, as shown in Figure 3. Since $\mathbf{X}$ is high-dimensional, we employed t-SNE (Van der Maaten & Hinton, 2008) to reduce the dimension of $\mathbf{X}$. From the visualization result, we have the following observations: (1) The $T=1$ data in the three datasets is concentrated around $Y=0$, with no significant differences. It is reasonable since the coefficient of T in the sample selection function is -3, which means that $T=1$ units are less affected by the collider bias. (2) the $T=0$ data in $\mathcal{D}_{\text{rep}}$ is scattered across the entire $\mathbf{X}$ and Y space. On the contrary, the majority of $T=0$ data in $\mathcal{D}_{\text{obs}}$ is distributed above $Y=0$. It shows that the collider bias causes a severe distribution shift problem. (3) The $T=0$ data in $\mathcal{D}_{\text{gen}}$ compensates for the lack of data below $Y=0$ in $\mathcal{D}_{\text{obs}}$, resulting in the distribution of the combined dataset similar to that of $\mathcal{D}_{\text{rep}}$. It demonstrates that C^{2}GAM can achieve the objective of recovering the target distribution.

4.3. Experiments on Real-World Data

4.3.1. DATASETS

We conducted experiments on three benchmark datasets obtained from real-world applications, i.e., **IHDP**[2] (Shalit et al., 2017), **Twins**[3] (Louizos et al., 2017), and **ACIC**[4] (Dorie et al., 2019) datasets. The ground truth CATE is known in the IHDP and ACIC datasets because the outcomes in these datasets are simulated. In the Twins dataset, we can also obtain the ground truth CATE because the treatment in this dataset is being the heavier twin, and the outcome of the one twin can be regarded as the counterfactual outcome of the other twin. All three datasets are only confounding biased. Therefore, we manually introduce collider bias into each dataset. The detailed information on the three datasets and how the collider bias is introduced are in Appendix B.

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

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

[4] https://users.nber.org/˜rdehejia/nswdata2.html.

4.3.2. Results

We compare the results of the proposed method against those of the baselines and those of using only $\mathcal{D}_{\text{rep}}$ as shown in Table 3. For clarity, we only report the best results of using $\mathcal{D}_{\text{rep}}$ to fit different estimators. The results show that using a combination of the observational data and samples generated by C^{2}GAM for training achieves better performance than the baselines on both $S = 0$ and $S = 1$ data. Note that C^{2}GAM achieves performance improvement regardless of specific estimators. It demonstrates that the proposed method can address collider bias in real-world scenarios and achieve a more precise treatment effect estimation.

5. Conclusion

In this paper, we focus on the collider bias problem in treatment effect estimation, which previous works failed to address. We rethink collider bias from an out-of-distribution perspective and propose a novel Coupled Counterfactual Generative Adversarial Model (C^{2}GAM) that leverages small-scale representative data without collider bias to address the collider bias problem in large-scale observation data. C^{2}GAM generates unselected samples using two generators and two discriminators, which can be jointly optimized. Combining the samples generated by C^{2}GAM with the observational data, we can fit any CATE estimators to improve performance. One main limitation is that C^{2}GAM relies on generative models to generate unselected samples in observations. Training generative models can be challenging, especially when dealing with complex and high-dimensional data with limited samples.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (2021YFC3340300), 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.

Angrist, J. D. and Pischke, J.-S. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2009.

Askalidis, G., Kim, S. J., and Malthouse, E. C. Understanding and overcoming biases in online review systems. *Decision Support Systems*, 97:23–30, 2017.

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 *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 433–450. ACM, 2022.

Blank, T. and Schmidt, P. National identity in a united germany: Nationalism or patriotism? an empirical test with representative data. *Political psychology*, 24(2):289–312, 2003.

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.

Buchanan, A. L., Hudgens, M. G., Cole, S. R., Mollan, K. R., Sax, P. E., Daar, E. S., Adimora, A. A., Eron, J. J., and Mugavero, M. J. Generalizing evidence from randomized trials using inverse probability of sampling weights. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 181(4):1193–1209, 2018.

Cole, S. R. and Stuart, E. A. Generalizing evidence from randomized clinical trials to target populations: the actg 320 trial. *American journal of epidemiology*, 172(1): 107–115, 2010.

Correa, J. and Bareinboim, E. Causal effect identification by adjustment under confounding and selection biases. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Correa, J., Tian, J., and Bareinboim, E. Generalized adjustment under confounding and selection biases. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

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

Dahabreh, I. J. and Hernán, M. A. Extending inferences from a randomized trial to a target population. *European journal of epidemiology*, 34:719–722, 2019.

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

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.

Emdin, C. A., Khera, A. V., and Kathiresan, S. Mendelian randomization. *Jama*, 318(19):1925–1926, 2017.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. Generative adversarial nets. *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

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

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.

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

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.

Hirano, K., Imbens, G. W., and Ridder, G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica: Journal of the econometric society*, 71(4):1161–1189, 2003.

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

Imbens, G. W. and Wooldridge, J. M. Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86, 2009.

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.

Kellerman, S. E. and Herold, J. Physician response to surveys: a review of the literature. *American journal of preventive medicine*, 20(1):61–67, 2001.

Lee, D., Yang, S., Dong, L., Wang, X., Zeng, D., and Cai, J. Improving trial generalizability using observational studies. *Biometrics*, 79(2):1213–1225, 2023.

Lesko, C. R., Buchanan, A. L., Westreich, D., Edwards, J. K., Hudgens, M. G., and Cole, S. R. Generalizing study results: a potential outcomes perspective. *Epidemiology*, 28(4):553–561, 2017.

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.

O'Muircheartaigh, C. and Hedges, L. V. Generalizing from unrepresentative experiments: a stratified propensity score approach. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 63(2):195–210, 2014.

Pearl, J. *Causality*. Cambridge university press, 2009.

Robins, J. M., Hernan, M. A., and Brumback, B. Marginal structural models and causal inference in epidemiology. *Epidemiology*, pp. 550–560, 2000.

Rosenbaum, P. R. and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

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.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., and Leaf, P. J. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 174(2): 369–386, 2011.

Tay, L., Ng, V., Kuykendall, L., and Diener, E. Demographic factors and worker well-being: An empirical review using representative data from the united states and across the world. *The role of demographics in occupational stress and well being*, pp. 235–283, 2014.

Tipton, E. Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics*, 38(3):239–266, 2013.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

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.

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*, 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*, 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.

A. Pseudo-Code

As stated in Section 3, we propose a novel C^{2}GAM method, which consists of two generators that respectively generate the missing $S=0$ samples and the missing S labels, as well as two discriminators that align the distribution of the combined generated $S=0$ samples with the observational dataset to match that of the entire data space. By jointly optimizing the generators and discriminators, C^{2}GAM can effectively generate missing data following the original distribution. Specifically, the pseudo-code of C^{2}GAM is detailed in Algorithm 1, where $\mathrm{Wass}(\cdot,\cdot)$. denotes the Wasserstein distance.

Algorithm 1 Coupled Counterfactual Generative Adversarial Model

```
Input: the observational dataset $\mathcal{D}_{\text{obs}}$, the representative dataset $\mathcal{D}_{\text{rep}}$, distance threshold $\alpha$.
Output: generated samples $\mathcal{D}_{\text{gen}}$.
  $n_{\text{obs}} \leftarrow$ the sample size of $\mathcal{D}_{\text{obs}}$.
  $n_{\text{gen}} \leftarrow n_{\text{obs}}$.
  $d \leftarrow +\infty$.
  initialization of parameters in $\mathrm{G_s}$, $\mathrm{G_d}$, $\mathrm{D_o}$, $\mathrm{D_u}$.
  while $d \geq \alpha$ and training losses of $\mathrm{G_s}$, $\mathrm{G_d}$, $\mathrm{D_o}$, $\mathrm{D_u}$ do not converge do
    use mini-batch gradient descent to optimize $\mathrm{G_s}$, $\mathrm{G_d}$, $\mathrm{D_o}$, $\mathrm{D_u}$ by

      $\min_{\mathrm{D_o},\mathrm{D_u}} -\mathbb{E}_{\{\mathbf{x},t,y\}\sim\mathcal{D}_{\text{obs}}}[\log(\mathrm{D_o}(\mathbf{x},t,y))] - \mathbb{E}_{\{\mathbf{x},t,y\}\sim\mathcal{D}_{\text{rep}}}[\mathrm{G_s}(\mathbf{x},t,y)\cdot\log(1-\mathrm{D_o}(\mathbf{x},t,y))]$
      $-\mathbb{E}_{z\sim\mathcal{N}(0,1)^{n_{\text{gen}}}}[\mathrm{G_s}(\mathrm{G_d}(z))\cdot\log(1-\mathrm{D_o}(\mathrm{G_d}(z)))] - \mathbb{E}_{z\sim\mathcal{N}(0,1)^{n_{\text{gen}}}}[\log(1-\mathrm{D_u}(\mathrm{G_d}(z))]$
      $-\mathbb{E}_{z\sim\mathcal{N}(0,1)^{n_{\text{gen}}}}[(1-\mathrm{G_s}(\mathrm{G_d}(z)))\cdot\log(\mathrm{D_u}(\mathrm{G_d}(z)))] - \mathbb{E}_{\{\mathbf{x},t,y\}\sim\mathcal{D}_{\text{rep}}}[(1-\mathrm{G_s}(\mathbf{x},t,y))\cdot\log(\mathrm{D_u}(\mathbf{x},t,y))]$

      $\min_{\mathrm{G_s},\mathrm{G_d}} \mathbb{E}_{z\sim\mathcal{N}(0,1)^{n_{\text{gen}}}}[(1-\mathrm{G_s}(\mathrm{G_d}(z)))\cdot\log(\mathrm{D_u}(\mathrm{G_d}(z)))] + \mathbb{E}_{\{\mathbf{x},t,y\}\sim\mathcal{D}_{\text{rep}}}[\mathrm{G_s}(\mathbf{x},t,y)\cdot\log(1-\mathrm{D_o}(\mathbf{x},t,y))]$
      $+\mathbb{E}_{\{\mathbf{x},t,y\}\sim\mathcal{D}_{\text{rep}}}[(1-\mathrm{G_s}(\mathbf{x},t,y))\cdot\log(\mathrm{D_u}(\mathbf{x},t,y))] + \mathbb{E}_{z\sim\mathcal{N}(0,1)^{n_{\text{gen}}}}[\mathrm{G_s}(\mathrm{G_d}(z))\cdot\log(1-\mathrm{D_o}(\mathrm{G_d}(z)))]$
      $+\mathbb{E}_{z\sim\mathcal{N}(0,1)^{n_{\text{gen}}}}[\mathrm{G_s}(\mathrm{G_d}(z))\cdot\log(1-\mathrm{D_o}(\mathrm{G_d}(z)))] + \mathbb{E}_{z\sim\mathcal{N}(0,1)^{n_{\text{gen}}}}[\log(1-\mathrm{D_u}(\mathrm{G_d}(z))]$
      $+\mathbb{E}_{z\sim\mathcal{N}(0,1)^{n_{\text{gen}}}}[(1-\mathrm{G_s}(\mathrm{G_d}(z)))\cdot\log(\mathrm{D_u}(\mathrm{G_d}(z)))]$.

    $n_0 \leftarrow$ the count of units in $\mathcal{D}_{\text{rep}}$ with $\mathrm{G_s}(\mathbf{x},t,y)=0$.
    $n_1 \leftarrow$ the count of units in $\mathcal{D}_{\text{rep}}$ with $\mathrm{G_s}(\mathbf{x},t,y)=1$.
    $n_{\text{gen}} \leftarrow \frac{n_{\text{obs}}\cdot n_0}{n_1}$.
    $\mathcal{D}_{\text{gen}} \leftarrow \{\mathrm{G_d}(z_i\sim\mathcal{N}(0,1))\}_{i=1}^{n_{\text{gen}}}$.
    $d \leftarrow \mathrm{Wass}(\mathbb{P}_{\{\mathbf{x},t,y\}\sim\mathcal{D}_{\text{rep}}}(\mathbf{x},t,y), \mathbb{P}_{\{\mathbf{x},t,y\}\sim\mathcal{D}_{\text{obs}}\cup\mathcal{D}_{\text{gen}}}(\mathbf{x},t,y))$.
    $\min_{\mathrm{G_d}} d$.
  end while
  $n_0 \leftarrow$ the count of units in $\mathcal{D}_{\text{rep}}$ with $\mathrm{G_s}(\mathbf{x},t,y)=0$.
  $n_1 \leftarrow$ the count of units in $\mathcal{D}_{\text{rep}}$ with $\mathrm{G_s}(\mathbf{x},t,y)=1$.
  $n_{\text{gen}} \leftarrow \frac{n_{\text{obs}}\cdot n_0}{n_1}$.
  $\mathcal{D}_{\text{gen}} \leftarrow \{\mathrm{G_d}(z_i\sim\mathcal{N}(0,1))\}_{i=1}^{n_{\text{gen}}}$.
  return $\mathcal{D}_{\text{gen}}$
```

B. Introduction to the Real-World Datasets

IHDP dataset: The original representative data of the Infant Health and Development Program (IHDP) aims to evaluate the effect of specialist home visits on the future cognitive test scores of premature infants (Brooksgunn et al., 1992). Following previous studies (Hill, 2011; Shalit et al., 2017), we removed a non-random subset of the treated group and used simulated outcomes to introduce confounding bias. To obtain the representative data, we randomly selected 60 samples from the dataset. To 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. We sampled 557 units from the $S=1$ data as the observational dataset. 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

Table 4. The results (mean ± std of $\sqrt{\text{PEHE}}$) of different generative approaches. The best results are in bold.

METHOD	$S = 1$ SAMPLES	$S = 0$ SAMPLES
C^2GAM	**0.984±0.074**	**1.083±0.086**
W/O G_s	1.453±0.649	1.449±0.789
W/O G_d	1.649±0.369	1.638±0.384
W/O D_o	1.205±0.434	1.506±0.680
W/O D_u	1.595±0.410	1.569±0.565

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 observational dataset comprises 557 units (139 treated, 418 control), and the representative dataset comprises 60 units (8 treated, 52 control) with 25 pre-treatment variables related to the infants and their families.

Twins dataset: The original data of twins birth in the USA between 1989-1991 aims at evaluating the effect of low birth weight on the mortality of infants in their first year of life (Almond et al., 2005). Following Louizos et al. (2017), we selected the twins whose gender is the same and weight is less than 2000kg into records. The treatment is being the heavier one in the twins, and the outcome is the one-year mortality. Because both treated (the heavier one in the twin) and control (the lighter one in the twin) outcomes are observed, the ground truth CATE is available in the Twins dataset. We used the same simulation as previous works to introduce confounding bias (Louizos et al., 2017), and we randomly selected 180 samples as our representative data. To introduce collider bias into the dataset, we set $S = 0$ for $T = 1$ units that both the mother uses tobacco and the twin is alive. We sampled 3000 units from the $S = 1$ data as our observational dataset. 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 observational dataset comprises 3000 units (1348 treated, 1652 control), and the representative dataset comprises 180 units (94 treated, 86 control) with 48 pre-treatment variables related to the twins and their families.

ACIC datasets: The 2016 Atlantic Causal Inference Challenge (ACIC) (Dorie et al., 2019) contains a series of causal benchmark datasets with diverse data generation processes. We randomly selected 20 settings from all the data generation processes and conducted experiments on them. These datasets are confounding biased but have no collider bias. Therefore, we used the same way as stated in Section 4.2.1 to introduce collider bias into them where $n_{\text{rep}} = 500$ and $n_{\text{obs}} = 10000$.

C. Ablation Studies of Each Module in C^2GAM

To further demonstrate the necessity of each module in the dual framework of C^2GAM, we compare our **C^2GAM** with the following ablation version of C^2GAM:

- **C^2GAM w/o** G_d, uses only G_d to generate samples from $\mathcal{D}_{\text{rep}}$ to estimate CATE,
- **C^2GAM w/o** G_d, uses only G_s to generate missing S labels for IPSW (Cole & Stuart, 2010),
- **C^2GAM w/o** D_o, uses only D_u to optimize the generators,
- **C^2GAM w/o** D_o, uses only D_o to optimize the generators,

Note that we use the generated S labels by C^2GAM w/o G_d to estimate sample selection probability for estimating CATE using IPSW (Cole & Stuart, 2010), which reweights each observational sample with its inverse probability of sample selection. Then, we conduct the experiments on the synthetic datasets as stated in Section 4.2.1 and compare the performance among the above different generative approaches. For simplicity, we only report the results of plugging the above approaches into CFR under $n_{\text{rep}} = 500$. The experimental results shown in Table 4 show that each module is essential for achieving high performance in C^2GAM. Removing any of the modules leads to a noticeable decrease in performance. The observations and detailed analysis of each module are presented below.

Representative data is limited, and observational data can provide more information. Compared the results of C^2GAM with C^2GAM without G_s, we can find that using only G_s to generate samples from $\mathcal{D}_{\text{rep}}$ suffers from limited information and cannot provide significant performance improvement compared to using the original data in $\mathcal{D}_{\text{rep}}$ as shown in Table 2. Thus, we must use observational data to supplement more observations for generating unselected samples.

Both generators are necessary, and the absence of either generator will result in the other generator not working. C^2GAM without G_d and C^2GAM without G_s both result in a performance decrease. The reason that G_d is necessary for G_s

is that since we only have the information of $S = 1$ data, we need $S = 0$ data generated by $\mathrm{G_d}$ to help recover $\mathbb{P}(S)$ of the target population. The reason that $\mathrm{G_s}$ is necessary for $\mathrm{G_d}$ is that because the ground truth $S = 0$ data is unavailable, we need $\mathrm{G_s}$ to generate $S = 0$ labels such that we can use the labeled $S = 0$ samples in $\mathcal{D}_{\mathrm{rep}}$ as approximations. Therefore, the results prove the necessity of using both generators.

Both discriminators are necessary for matching the distributions of $\mathcal{D}$ and $\mathcal{D}_{\mathrm{obs}} \cup \mathcal{D}_{\mathrm{gen}}$. The performance of C^{2}GAM without either discriminator decreases for the following reasons. If we only use $\mathrm{D_u}$ to match the distribution of the generated samples and that of data from $\mathcal{D}_{\mathrm{rep}}$ with generated $S = 0$ labels, since the two distributions are both from the generators but not already known in $\mathcal{D}_{\mathrm{rep}}$ and $\mathcal{D}_{\mathrm{obs}}$, the performance of $\mathrm{D_u}$ depends entirely on the performance of the generators, which itself depends on the performance of $\mathrm{D_u}$. Therefore, neither the generators nor $\mathrm{D_u}$ can be well-optimized. If we only use $\mathrm{D_o}$ to match the distribution of $\mathcal{D}_{\mathrm{obs}}$ and that of data from $\mathcal{D}_{\mathrm{rep}}$ with generated $S = 1$ labels, there are not any constraints on samples generated by $\mathrm{G_d}$, resulting in the performance of $\mathrm{G_d}$ becoming unpredictable. What is worse, since $\mathrm{G_d}$ highly possibly generate incorrect $S = 0$ samples, $\mathrm{G_s}$ also cannot work well because it needs the generated samples to recover $\mathbb{P}(S)$. Therefore, the results prove the necessity of using both discriminators.

The above observations and analysis demonstrate that the design of C^{2}GAM is reasonable, and each module in C^{2}GAM is practical and necessary.

D. Discussion on the Differences between C^{2}GAM and Previous Generative Model Based Causal Inference Methods

Our method is different from previous generative-model-based causal methods (Louizos et al., 2017; Yoon et al., 2018; Zhang et al., 2021) in the following aspects:

1) **The solved problems are different.** Previous methods use generative models to solve confounding bias, while our work focuses on collider bias, which was overlooked in previous works.
2) **The targets generated by the generated models are different.** Previous methods use generative models to generate counterfactual outcomes ($Y(1-t)$) to address confounding bias, while our work aims to use generative models to generate the $S = 0$ data ($\mathbf{X}, T, Y$) that were not selected into the observational samples. To achieve this goal, we introduce an additional representative dataset and use generative models to generate missing S labels in representative data to generate $S = 0$ samples better.
3) **The termination conditions are different.** Previous methods only use a single generative model to generate counterfactual samples, and their termination conditions mainly are the same as those of the based generative models, such as GAN. However, our proposed model consists of two coupled generators and discriminators. Therefore, in addition to meeting the primary constraints of GAN, our model also needs to ensure that the distribution of the final generated $S = 0$ samples plus the original observational samples is the same as the target population distribution. Therefore, the termination condition also involves constraints of the distance between the above distributions.

To the best of our knowledge, we are the first to introduce GANs to solve collider bias, which is entirely different from the previous works using GANs to solve confounding bias.