PrivATE: Differentially Private Average Treatment Effect Estimation for Observational Data

Quan Yuan*†, Xiaochen Li§, Linkang Du¶, Min Chen‖, Mingyang Sun**,
Yunjun Gao*, Shibo He*, Jiming Chen*††, Zhikun Zhang*‡
*Zhejiang University, †University of Virginia, §University of North Carolina at Greensboro, ¶Xi'an Jiaotong University,
‖Vrije Universiteit Amsterdam, **Peking University, ††Hangzhou Dianzi University
Email: *{yq21, gaoyj, s18he, cjm, zhikun}@zju.edu.cn, §X_LI12@uncg.edu, ¶linkangd@gmail.com,
‖m.chen2@vu.nl, **smy@pku.edu.cn

***Abstract*—Causal inference plays a crucial role in scientific research across multiple disciplines. Estimating causal effects, particularly the average treatment effect (ATE), from observational data has garnered significant attention. However, computing the ATE from real-world observational data poses substantial privacy risks to users. Differential privacy, which offers strict theoretical guarantees, has emerged as a standard approach for privacy-preserving data analysis. However, existing differentially private ATE estimation works rely on specific assumptions, provide limited privacy protection, or fail to offer comprehensive information protection.**

To this end, we introduce PrivATE, a practical ATE estimation framework that ensures differential privacy. In fact, various scenarios require varying levels of privacy protection. For example, only test scores are generally sensitive information in education evaluation, while all types of medical record data are usually private. To accommodate different privacy requirements, we design two levels (*i.e.*, label-level and sample-level) of privacy protection in PrivATE. By deriving an adaptive matching limit, PrivATE effectively balances noise-induced error and matching error, leading to a more accurate estimate of ATE. Our evaluation validates the effectiveness of PrivATE. PrivATE outperforms the baselines on all datasets and privacy budgets.

I. Introduction

Causal inference, the process of determining a causal relationship by analyzing the conditions under which an effect occurs, has been a fundamental research focus for decades in various fields [1], including healthcare [2], economics [3], statistics [4], public policy [5], education [6], *etc.* A common example of causal inference is evaluating the impact of taking a drug by analyzing patient data, which can assist doctors in making informed decisions. There are two typical settings for causal inference: randomized controlled trials (RCTs) and observational studies. In RCTs, the treatment assignment is controlled by random assignment, *e.g.*, all patients are randomly assigned to two groups: One group takes the drug, while the other does not. However, randomized trials are frequently impractical due to ethical, technical, or economic limitations in contexts like studying smoking behavior or assessing economic policies. In contrast, observational studies (*i.e.*, not intervening in individual grouping, only observing and analyzing naturally occurring data) are more practical [7], *e.g.*, we can only analyze existing patient data but have no control over whether a patient takes the medicine. The setting of observational studies has gained increasing attention due to the abundance of available data and the low budget requirement [8].

A key task in observational studies is to estimate the average treatment effect (ATE), which quantifies the overall impact of treatment across all samples. Here, ATE is calculated as the mean of individual treatment effects, where an individual treatment effect is defined as the difference between a sample's potential outcome under treatment and its potential outcome without treatment. ATE estimation in observational studies often suffers from selection bias and missing information [9]. There may be significant differences in the characteristics of the treated and control groups. In addition, for people who take the drug, the effect of not taking it is unknown. To mitigate the impact of bias and missing information, a common solution is to estimate the counterfactual results of each sample and then calculate the causal effect, *e.g.*, sample matching [10].

However, directly computing ATE from real data using the above approach in observational studies poses significant risks of privacy leakage. Data utilized in causal inference often contain sensitive personal information [11]. Direct manipulation and analysis of true data are increasingly challenged by growing concerns over privacy and the emergence of regulations for safeguarding individual data. For example, releasing any statistical information derived from real medical data poses a risk of compromising patient privacy. For a strong threat model, the attacker could perform a differential attack to infer whether the specific sample's data is included in the dataset (*i.e.*, comparing query results that include and exclude the sample).

Differential privacy (DP) [12], a golden standard in the privacy community, has been widely applied for privacy-

†Quan Yuan's work on this paper was done while working as a visiting student at the University of Virginia.
‡Zhikun Zhang is the corresponding author.

Network and Distributed System Security (NDSS) Symposium 2026
23-27 February 2026, San Diego, CA, USA
ISBN 979-8-9919276-8-0
https://dx.doi.org/10.14722/ndss.2026.241350
www.ndss-symposium.org

preserving data analysis [13–16]. By injecting carefully designed noise into the aggregated statistical value, DP can ensure that a single user record has a limited impact on the final output. Due to the advantages of quantifiable privacy guarantees, high flexibility, and low cost, DP has been deployed by many companies and government agencies [13]. For instance, LinkedIn builds Pinot [14], a DP platform that enables analysts to gain insight from its members' content engagements. Although DP serves as an effective privacy protection strategy, within the context of ATE estimation in observational studies, only a small amount of literature has explored privacy-preserving solutions using DP [17–21].

Existing Solutions. Existing studies exhibit several limitations in terms of the assumptions of the problem, the scope of protection, and the methodological implementation. First, some approaches assume binary outcomes (*e.g.*, a patient's medication outcome is categorized as success or failure rather than represented by a continuous numerical indicator), which constrains their applicability. Second, many solutions offer protection only for partial data attributes (*e.g.*, safeguarding covariates such as patient age and height while leaving medication outcome unprotected). Third, most existing works address selection bias through sample reweighting [22], which aims to eliminate the distribution differences between the treated and control groups. To ensure a bounded sensitivity, these methods typically employ a fixed, pre-defined truncation threshold to limit individual sample weights. However, such a fixed configuration inherently lacks flexibility and interpretability. Therefore, it is challenging to design a more practical and highly flexible privacy-preserving ATE estimation framework in observational studies.

Our Proposal. To overcome the limitations of the existing literature and eliminate data bias while effectively protecting user privacy, we propose a matching-based framework PrivATE to estimate the ATE for observational data in a private manner. Compared to existing works, PrivATE provides a more practical and general solution. In particular, PrivATE does not rely on idealized assumptions such as binary outcomes, which expands its application scenarios. Furthermore, considering the different trade-offs between utility and privacy in various scenarios, PrivATE includes two levels of privacy protection: label-level and sample-level. Label-level protection only perturbs the outcome, which offers higher utility. Sample-level protection perturbs all attributes, including treatment, covariates, and outcome, which provides the strongest privacy protection. The two levels of protection provide solutions for different application scenarios: Label-level protection is suited for outcome-sensitive settings (*e.g.*, education evaluation), where other information is publicly available. Sample-level protection is for cases involving fully sensitive data (*e.g.*, medical study), where all attributes (including medical records, administered treatments, and outcomes) require protection.

In the causal effect estimation, it is crucial to restrict the maximum number of matches for each sample, otherwise high sensitivity will occur. In this way, there will be two types of error in the final ATE estimation: noise error and matching error. However, it is challenging to choose a suitable matching limit for vairous datasets and privacy budgets. Taking into account the combined influence of noise error and matching error, we propose an adaptive matching limit determination mechanism to strike a balance between reducing global sensitivity and improving matching accuracy. On this basis, we can calculate the counterfactual outcome for each sample in a more accurate manner. Furthermore, we choose to perturb the sum of aggregated outcomes rather than individual outcomes to reduce the error in the ATE estimation.

Evaluation. We compare PrivATE with the baseline methods on multiple typical datasets, including real, semi-real, and synthetic datasets. The experimental results show the superiority of PrivATE. For instance, for the sample-level privacy, PrivATE consistently outperforms other baseline methods across all datasets and budgets. In addition, for the label-level privacy, PrivATE can achieve a low relative error (*i.e.*, less than 0.2) on multiple datasets even when the privacy budget is 0.5. We further verify the effectiveness of our proposed adaptive matching limit determination mechanism with a comparison to the fixed matching limit methods. We also explore the impact of the hyperparameter for matching limit calculation. Moreover, we illustrate the influence of various privacy budget allocations on the ATE estimation.

Contributions. In summary, the main contributions of this paper are four-fold:

- We propose PrivATE, a more practical and effective privacy-preserving ATE estimation framework under differential privacy, outperforming existing works.
- In PrivATE, we provide two levels of privacy protection (*i.e.*, label-level and sample-level) to satisfy different trade-offs between utility and privacy in various scenarios.
- We further design an adaptive matching limit determination mechanism to strike a balance between reducing global sensitivity and improving matching accuracy.
- We conduct extensive empirical experiments on multiple datasets to illustrate the effectiveness of PrivATE. Under the same privacy settings, PrivATE achieves superior performance compared to the baseline methods. PrivATE is open-sourced at https://github.com/sec-priv/PrivATE.

II. Preliminaries

A. Causal Inference

In general, a causal inference task estimates how the outcome would change if another treatment had been applied. Due to the widespread availability of observational data, estimating treatment effects from such naturally occurring datasets has garnered increasing attention. Observational data typically includes a group of individuals who have received different treatments, their corresponding outcomes, and possibly additional information, but without direct access to the mechanism or reasons for taking the specific treatment [18].

Definition 1 (Treatment)**.** *Treatment T represents the action that applies to a sample. The group of samples with treatment*

$T = 1$ *is called the treated group, and the group of samples with* $T = 0$ *is called the control group.*

Definition 2 (Potential Outcome)**.** *For each unit-treatment pair, the outcome of that treatment when applied on that sample is the potential outcome. The potential outcome of treatment with value* t *is denoted as* $Y(T = t)$*.*

Definition 3 (Observed Outcome)**.** *The observed outcome is also called factual outcome (denoted as* Y^F*), which represents the outcome of the treatment that is actually applied.* $Y^F = Y(T = t)$ *where* t *is the treatment actually applied.*

Definition 4 (Counterfactual Outcome)**.** *The counterfactual outcome* Y^{CF} *is the outcome if the sample took another treatment.* $Y^{CF} = Y(T = 1 - t)$ *where* t *is the treatment actually applied.*

Definition 5 (Covariate)**.** *Covariate* X *is the variable that is not affected by the treatment but still affects the experimental results.*

Average Treatment Effect. Average treatment effect (ATE) is defined as follows:

$$\tau = \mathbb{E}[Y(T = 1) - Y(T = 0)], \tag{1}$$

where $Y(T = 1)$ and $Y(T = 0)$ are the potential treated and control outcomes of the whole population, respectively.

Mainstream Solutions for ATE Estimation. Currently, there are two main methods that can conduct the ATE estimation while mitigating the impacts of bias and missing information. One way is to eliminate the distribution differences between the treated and control groups, *e.g.*, sample reweighting [22]. By adjusting the weight of each sample, sample reweighting ensures a similar distribution between the treated and control groups. However, applying DP to this approach often requires predefined thresholds to limit global sensitivity, which lacks flexibility and interpretability.

The other is to estimate the counterfactual results of each sample and then calculate the causal effect, *e.g.*, sample matching [10]. This method pairs treated and control samples with similar characteristics. This approach can reduce selection bias by identifying individuals with similar characteristics in the treated and control groups, ensuring that the matched samples are as balanced as possible on the covariates. Given its intuitiveness and practicality, we choose to achieve a privacy-preserving framework based on matching in this work.

Propensity Score Matching. As a typical matching method, propensity score matching (PSM) is widely used in observational experiments due to its strong interpretability and low matching complexity [8].

Therefore, we utilize the PSM approach as a basis to estimate counterfactual results and eliminate the bias of causal effects caused by systematic differences between the treatment and control groups. The propensity score is defined as the conditional probability of treatment given related variables:

$$e(x) = \mathsf{Pr}\left[T = 1 | X = x\right]. \tag{2}$$

The propensity score reflects the probability of a sample being assigned to the treatment given a series of observed variables. However, in most observational studies, the treatment assignment mechanism is unknown. A common approach is to fit a propensity score function using a standard statistical model on the dataset D. In this paper, logistic regression is adopted since it is the most frequently used model in existing works. As a result, on the basis of the absolute value of the difference between various propensity scores, the similarity between any two samples can be calculated and utilized to match. The distance between the sample u_1 in the treated group and the sample u_2 in the control group is as follows:

$$dis(u_1, u_2) = |e(x_1) - e(x_2)|, \tag{3}$$

where $e(x_1)$ represents the propensity score of sample u_1, and $e(x_2)$ represents the propensity score of sample u_2.

The goal of matching is to identify several most similar samples from the opposite treatment group for each sample in the current treatment group. Then, the counterfactual outcome can be obtained based on the matched samples. In general, the counterfactual outcome of the i-th sample is as follows:

$$Y_i^{CF} = \frac{1}{|\mathcal{P}(i)|} \sum_{l \in \mathcal{P}(i)} Y_l^F, \tag{4}$$

where $\mathcal{P}(i)$ is the matched neighbors of sample i in the opposite treatment group. Based on the observed and counterfactual outcomes of each sample, ATE can be obtained by Equation 1.

B. Differential Privacy

Differential Privacy (DP) [12] was designed for the data privacy-protection scenarios, where a trusted data curator collects data from individual users, applies perturbation to the aggregated results, and then publishes them. Intuitively, DP guarantees that any single sample from the dataset has only a limited impact on the output.

Definition 6 ((ε, δ)-Differential Privacy)**.** *An algorithm* $\mathcal{A}$ *satisfies* (ε, δ)*-differential privacy* ((ε, δ)*-DP), where* $\varepsilon > 0$*, if and only if for any two neighboring datasets* D *and* D'*, we have*

$$\forall O \subseteq Range(\mathcal{A}) : \mathsf{Pr}\left[\mathcal{A}(D) \in O\right] \leq e^{\varepsilon} \mathsf{Pr}\left[\mathcal{A}(D') \in O\right] + \delta,$$

where Range $(\mathcal{A})$ denotes the set of all possible outputs of the algorithm $\mathcal{A}$, and δ indicates the probability of $\mathcal{A}$ failing to satisfy DP. When $\delta = 0$, which is the case we consider in this work (*i.e.*, pure DP), we write ε-DP for convenience. Pure DP can provide strict theoretical guarantees, while approximate DP (*i.e.*, $\delta > 0$) has a certain probability of violating theoretical constraints. Approximate DP relaxes the privacy constraint to enable the use of a wider range of composition properties, which may be helpful to improve the utility. At the same time, according to the experimental results of Section IV, our method (satisfying pure DP) still shows significant advantages over the baselines (satisfying approximate DP).

In addition, the definition of ε-sample differential privacy (ε-Sample DP) in the paper is consistent with ε-DP. Here, we consider two datasets D and D' to be *neighbors*, denoted as $D \simeq D'$, if and only if $D = D' + r$ or $D' = D + r$, where $D + r$ is the dataset resulted from adding the record r to D.

Definition 7 ((ε, δ)-label Differential Privacy)**.** *An algorithm $\mathcal{A}$ satisfies (ε, δ)-label differential privacy ((ε, δ)-Label DP), where $\varepsilon > 0$, if and only if for any two datasets H and H' that differ in the label (observed outcome) of a single sample, we have*

$$\forall O \subseteq Range(\mathcal{A}) : \mathsf{Pr}\left[\mathcal{A}(H) \in O\right] \leq e^{\varepsilon} \mathsf{Pr}\left[\mathcal{A}(H') \in O\right] + \delta.$$

Similar to Definition 6, we consider $\delta = 0$ in this paper, and we write ε-Label DP instead of (ε, δ)-Label DP.

Laplace Mechanism. Laplace mechanism (LM) satisfies the DP requirements by adding a random Laplace noise to the aggregated results [23]. The magnitude of the noise depends on GS_f, *i.e.*, *global sensitivity*,

$$GS_f = \max_{D \simeq D'} \parallel f(D) - f(D') \parallel_1,$$

where f represents the aggregation function and D (or D') is the users' data. When f outputs a scalar, the Laplace mechanism $\mathcal{A}$ is given below:

$$\mathcal{A}_f(D) = f(D) + \mathcal{L}\left(\frac{GS_f}{\varepsilon}\right),$$

where $\mathcal{L}(\beta)$ stands for a random variable sampled from the Laplace distribution $\mathsf{Pr}\left[\mathcal{L}(\beta) = x\right] = \frac{1}{2\beta} e^{-|x|/\beta}$. When f outputs a vector, $\mathcal{A}$ adds independent samples of $\mathcal{L}(\beta)$ to each element of the vector. The global sensitivity of all elements is the same value.

Random Response Mechanism. The random response (RR) mechanism can be applied to protect the privacy of binary variables [24–26]. Given a specific privacy budget ε, the probability of outputting a true binary variable p is as follows:

$$p = \frac{e^{\varepsilon}}{e^{\varepsilon} + 1}.$$

Composition Properties of DP. The following composition properties of DP are commonly utilized to construct complex differentially private algorithms from simpler subroutines [12].

- **Sequential Composition.** Combining multiple subroutines that satisfy differential privacy for $\{\varepsilon_1, \cdots, \varepsilon_k\}$ results in a mechanism which satisfies ε-differential privacy for $\varepsilon = \sum_i \varepsilon_i$.
- **Parallel Composition.** Given k algorithms working on disjoint subsets of the dataset, each satisfying DP for $\{\varepsilon_1, \cdots, \varepsilon_k\}$, the result satisfies ε-differential privacy for $\varepsilon = \max\{\varepsilon_i\}$.
- **Post-processing.** Given an ε-DP algorithm $\mathcal{A}$, releasing $z(\mathcal{A}(D))$ for any z still satisfies ε-DP, *i.e.*, post-processing the output of a differentially private algorithm does not incur any additional loss of privacy.

TABLE I: Summary of mathematical notations.

Notation	Description
D	Dataset
T	The treatment
X	The covariates
Y	The observed outcome
Y_1	The potential treated outcome of the whole population
Y_0	The potential control outcome of the whole population
ε	Privacy budget
n	The number of all samples
n_t	The number of samples in the treated group
n_c	The number of samples in the control group
d	The dimensions of covariates
k	The value of matching limit
N	The number of neighbors for each sample in the matching
B	The maximum variation range of outcome
w	The weights of the regression model
e	The propensity score
τ	The average treatment effect estimate

III. Methodology

A. Problem Definition

In this paper, we consider a dataset $D = (T, X, Y)$ that contains multiple dimensions, where T stands for the treatment, X stands for the related covariates, and Y stands for the observed outcome. Note that both T and Y contain only one column, while X can contain multiple columns. Without loss of generality, we assume that $T = 0/1$, $X \in [0, 1]^d$ (d is the dimension of covariates), and the maximum variation range of outcome is B. Our goal is to estimate an average treatment effect that closely aligns with the result obtained through propensity score matching while ensuring strict differential privacy. Specifically, we aim to achieve two levels of privacy protection, *i.e.*, *label-level* and *sample-level*. For label-level privacy, only the observed outcome Y is private. For sample-level privacy, all types of data are private. For ease of reading, we summarize the frequently used notations in Table I.

B. Motivation

When implementing propensity score matching under DP, a common idea is to directly add noise to the ATE estimate to achieve DP. However, the impact of adding or deleting any sample on the propensity score matching is difficult to evaluate. Therefore, we choose to apply DP to each phase of propensity score matching, thus ensuring the privacy of the entire process. If ATE is estimated completely according to the matching results, it could introduce excessive noise. This will make some samples match too many times, making the sensitivity too high. If the number of matches is too low, the estimation of ATE will be inaccurate. Thus, we determine the maximum number of matches for each sample by estimating the combined impact of noise injection and matching accuracy, thereby achieving a great trade-off between these two aspects. Considering the privacy requirements in various scenarios in practice, we designed two different levels of privacy protection approaches to estimate ATE, *i.e.*, label-level privacy and sample-level privacy.

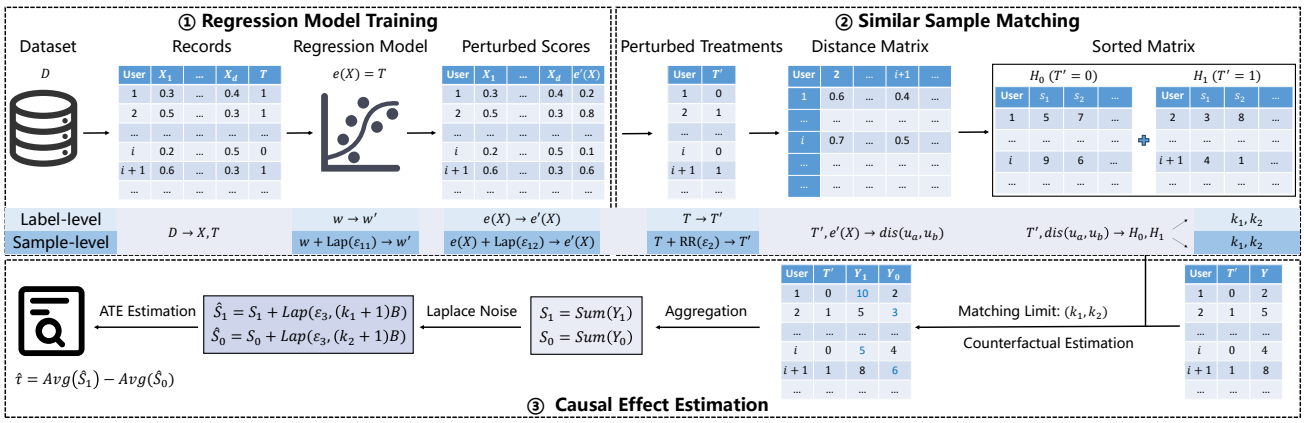

Fig. 1: PrivATE overview. PrivATE consists of three phases: Regression model training, similar sample matching, and causal effect estimation. First, a regression model for calculating propensity scores can be obtained in the regression model training phase. Then, PrivATE finds the closest neighbors in the opposite group for each sample in the similar sample matching phase. In the causal effect estimation phase, PrivATE calculates each sample's counterfactual outcome based on the matching results and the matching limit, *i.e.*, the maximum number of matched of each sample. After that, the potential outcomes for the control and treated groups are aggregated and perturbed. Finally, the average treatment effect can be estimated by the perturbed outcomes.

Here, we summarize the key challenges of ATE estimation under the two privacy settings as follows: For label-level privacy, directly applying a standard DP mechanism to existing ATE estimation methods often introduces excessive noise. Furthermore, using a fixed matching upper limit is unsuitable across different privacy budgets and data distributions, making it difficult to adaptively determine a limit that balances matching accuracy and privacy protection. For sample-level privacy, which protects the entire dataset rather than just labels, similar issues arise. Under stricter privacy requirements, additional challenges include allocating the overall privacy budget and determining an appropriate matching limit while ensuring DP guarantees at each step.

C. Overview

As shown in Figure 1, the framework of PrivATE contains three phases, *i.e.*, regression model training, similar sample matching, and causal effect estimation.

Phase 1: Regression Model Training (RMT). We train a logistic regression model to estimate the propensity scores of all samples. In the label-level setting, the model training and the propensity score calculation do not need to consume the privacy budget. The reason is that this process does not require access to the observed outcomes of the samples. In the sample-level setting, the training of the regression model and the estimation of the propensity score need to be perturbed to meet DP.

Phase 2: Similar Sample Matching (SSM). In this phase, the distance between the propensity score of any sample and the propensity scores of all samples in the opposite treatment group can be calculated. By sorting these scores, we can obtain the most similar neighbors of each sample in the opposite group. Note that two sorted matrices are calculated here, one for the control group and the other for the treated group. In the label-level setting, this procedure still does not visit the observed outcome, thus consuming no privacy budget. In the sample-level setting, the true treatment T is perturbed to satisfy DP in this phase.

Phase 3: Causal Effect Estimation (CEE). Based on the sorted matrices in the last phase, we can find the closest neighbors in the opposite group for each sample. Then, the counterfactual outcomes of all samples can be estimated. Here, we limit the maximum number of times each sample can be used for matching. The matching limit can be adaptively adjusted based on the privacy budget and the characteristics of the dataset. After calculating the counterfactual outcomes, we aggregate and perturb the sum of potential outcomes of all samples. Then, the ATE can be obtained by Equation 1.

D. Regression Model Training

In the first phase, we train a logistic regression model based on the original dataset. The covariates X is the independent variable of the regression model, while the treatment T is the dependent variable. Algorithm 1 illustrates the basic process of the first phase.

Label-level Privacy. In the label-level privacy, only the outcomes need to be protected. The regression model training phase does not require access to the true outcomes. Therefore, we can utilize the true parameters of the trained model w to predict the propensity scores of all samples without consuming the privacy budget. The predicted propensity scores $e(X)$ also do not require to be perturbed in the label-level privacy setting.

Sample-level Privacy. In contrast, sample-level privacy requires to protect the privacy of all types of variables. If we directly adopt the weight w of the unprotected regression model, there will be a risk of privacy leakage [27]. Therefore, we choose to perturb the training of the regression model to satisfy DP. The training of the logistic regression model can be regarded as a specific case for regularized empirical risk minimization. For the logistic regression with ℓ_2 regularization

Algorithm 1: Regression Model Training (Phase 1)

Input: Original dataset D, privacy level l, privacy budget $\varepsilon_{11}, \varepsilon_{12}$ (if l is sample-level)
Output: Propensity Score $e'(X)$

Train a logistic regression model based on the covariates X and treatment T.
if *l is label-level* **then**
  $w' \leftarrow w$
else
  `// Private model training (sample-level)`
  $w' \leftarrow w + Lap(\varepsilon_{11}, \Delta f_w)$
Calculate the propensity score of each sample $e(X)$ based on the model parameter w'.
if *l is label-level* **then**
  $e'(X) \leftarrow e(X)$
else
  `// Private score calculation (sample-level)`
  $e'(X) \leftarrow e(X) + Lap(\varepsilon_{12}, \Delta f_e)$

Algorithm 2: Similar Sample Matching (Phase 2)

Input: Original dataset D, propensity score $e'(X)$, privacy level l, privacy budget ε_2 (if l is sample-level)
Output: Treatment T', Sorted Matrices H

if *l is label-level* **then**
  $T' \leftarrow T$
else
  `// Treatment perturbation (sample-level)`
  $T' \leftarrow RR(T; \varepsilon_2)$
Obtain the division of treated and control groups by T'
`// Distance sorting`
for *each sample j in the control group* **do**
  Calculate the distance vector dis_{j,s_t} between j and the samples s_t in the treated group based on Equation 3
  $H_0^j \leftarrow argsort(dis_{j,s_t})$
for *each sample j in the treated group* **do**
  Calculate the distance vector dis_{j,s_c} between j and the samples s_c in the control group based on Equation 3
  $H_1^j \leftarrow argsort(dis_{j,s_c})$

penalty, the regularized empirical loss can be written as follows:

$$J(w) = \frac{1}{n}\sum_{i=1}^{n}\log(1 + e^{-X_i^T w t_i}) + \frac{\lambda}{2}||w||_2^2, \tag{5}$$

where X is the training feature (covariates) containing d-dim and t_i is the i-th sample's treatment. The weight w can be perturbed to satisfy DP. The L_1-sensitivity of w is $\frac{2d}{n\lambda}$, and the detailed derivation is given in Proof 1 of Appendix C. By injecting Laplace noise into the true weight w with the privacy budget of ε_{11}, we can generate a privacy-preserving regression model.

After finishing the private model training, we can utilize the model to calculate the propensity score of each sample. Since this step requires accessing the true covariate again, we need to add Laplace noise to the relevant query results to meet differential privacy. The output range of logistic regression model is $[0, 1]$, thus the sensitivity of propensity score is $\Delta f_e = 1$.

In this phase, we inject noise into w and $e(X)$, respectively. Note that both parts of noise are indispensable. The purpose of adding noise to w is to protect the privacy of training data (*i.e.*, X and T). If the DP regression model is queried using public or non-sensitive data, no additional privacy budget is consumed due to the post-processing property. However, $e(X)$ is computed using the actual data X, which constitutes additional access to private information. To preserve the privacy of X, we still need to inject noise into $e(X)$. In addition, if w is not perturbed and only $e(X)$ is perturbed, we cannot apply the parallel composition to perturb each sample in X because w contains sensitive information. Therefore, it is necessary to inject noise into w and $e(X)$.

E. Similar Sample Matching

In the second phase, we try to calculate the similarity of each sample with all samples in the other group and rank them. Algorithm 2 provides the specific procedures of the similar sample matching.

Label-level Privacy. In this setting, the treatment T is accessible, which does not need to be perturbed. Then, we can calculate the distance between each sample and all samples in the opposite treatment group based on the propensity score $e'(X)$ obtained from the first phase. The specific calculation formula is shown in Equation 3. Then, we traverse each sample and sort all candidate samples in the opposite treatment group in ascending order based on the distance vectors. From this, we can obtain two sorted index matrices, one for the control group (*i.e.*, H_0) and the other one for the treated group (*i.e.*, H_1). The sorted matrices can be utilized for counterfactual estimation in the next phase.

Sample-level Privacy. Unlike label-level privacy, the treatment T is sensitive information in the sample-level setting. Considering that T is a binary variable, it is not appropriate to inject Laplace noise, which is used for continuous variables. Here, based on the privacy budget of ε_2, we adopt the random response mechanism to perturb T, which effectively protects privacy and improves data utility. The next steps are similar to the label-level settings. Using the perturbed treatment T' and perturbed propensity scores $e'(X)$, we can calculate the distance between each sample and other samples in the opposite perturbed treatment group. Then, we traverse each sample and sort other samples according to the corresponding distance values. After that, we obtain two sorted matching

Algorithm 3: Causal Effect Estimation (Phase 3)

Input: Original dataset D, sorted matrices H, the number of neighbors in matching N, treatments T', privacy level l, the range of outcome B, privacy budget ε_3, the error coefficient c (label-level) or h (sample-level)

Output: Average treatment effect estimate $\hat{\tau}$

Count the maximum number of times that all samples appear in the first N neighbors of the sorted index matrices M (label-level) or M' (sample-level).

`// Obtain the number of samples in the treated and control groups, and the average maximum number of matches.`

if *l is label-level* **then**

  $n_t, n_c \leftarrow T, M_1 = \frac{M}{N}$

else

  $n'_t, n'_c \leftarrow T', M'_1 = \frac{M'}{N}$ // Sample-level

$r_1 = \frac{n_t}{n_c}$ (or $r_1 = \frac{n'_t}{n'_c}$)

`// Matching limit calculation`

if $r_1 \leq 1$ **then**

  Calculate the matching limit of treated group k_1 based on Equations 9-10 (or Equations 11-12)

  $k_2 = \max(1, round(k_1 \cdot r_1))$

else

  Calculate the matching limit of control group k_2 based on Equations 9-10 (or Equations 11-12)

  $k_1 = \max(1, round(k_2/r_1))$

`// Counterfactual estimation`

for *i-th sample in D* **do**

  Remove candidate samples that have reached the upper matching limit from the original sorted list.

  Select the nearest N neighbors from the remaining sorted list, and add 1 to their matching counts.

  Calculate the counterfactual outcome based on Equation 4.

`// Outcome aggregation`

$S_1 = Sum(Y_1)$, $S_0 = Sum(Y_0)$

`// Noise perturbtion`

$\hat{S}_1 = S_1 + Lap(\varepsilon_3, (k_1+1)B)$, $\hat{S}_0 = S_0 + Lap(\varepsilon_3, (k_2+1)B)$

`// ATE estimation`

$\hat{\tau} = \frac{1}{n} \cdot \hat{S}_1 - \frac{1}{n} \cdot \hat{S}_0$

matrices.

F. Causal Effect Estimation

In the third phase, we need to calculate each sample's counterfactual estimate and obtain the final ATE estimate.

Label-level Privacy. If we apply the original non-private approach to select N neighbors for computing counterfactual results, the number of matches for some samples may be extremely high, which will make the global sensitivity large. On the other hand, if we set the upper limit of the number of times each sample is matched very small, the disturbance intensity of the noise will be reduced, but the error of the counterfactual estimate will tend to increase. In addition, the noise intensity under various privacy budgets is different, making matching selection more challenging.

To address the above difficulties, we design an adaptive matching upper limit determination mechanism by considering the combined impact of noise injection and matching error, which can provide different matching upper limits for various privacy budgets.

We first consider the expected squared error of estimating an aggregated potential outcome $S = Sum(Y)$. Assuming that $\hat{S}$ is the estimation of S, then the expected squared error can be written as the summation of variance and the squared bias of $\hat{S}$:

$$\mathbb{E}[(\hat{S} - S)^2] = \mathsf{Var}[\hat{S}] + \mathsf{Bias}[\hat{S}]^2 \tag{6}$$

Given the maximum variation range of the outcome B, the matching upper limit for each sample k and privacy budget ε, we can obtain $\mathsf{Var}[\hat{S}] \approx \frac{2k^2B^2}{\varepsilon^2}$.

For $\hat{\mathsf{S}}$, its value is related to the number of samples, the true maximum number of matches, and the set matching upper limit. We can count the maximum number of times that all samples appear in the first N neighbors of the sorted index matrices, denoted as M. Here, we let $M_1 = \frac{M}{N}$, which represents the average maximum number of matches. Intuitively, the smaller the matching upper bound, the greater the bias. The larger the number of samples and the average maximum number of matches, the larger the bias. Therefore, we estimate $\hat{\mathsf{S}}$ as follows:

$$\mathsf{Bias}[\hat{S}] \approx c \cdot B \cdot n_1 \cdot \frac{M_1}{k}, \tag{7}$$

where $n_1 = \max(n_t, n_c)$ is the number of samples in the treated group or control group and c is the error coefficient, which is a hyperparameter. Therefore, the combined error of noise perturbation and matching limit can be approximately estimated as follows:

$$\mathbb{E}[(\hat{S} - S)^2] \approx \frac{2k^2B^2}{\varepsilon^2} + c^2 \cdot B^2 \cdot n_1^2 \cdot \frac{{M_1}^2}{k^2}. \tag{8}$$

By calculating the minimum value of Equation 8, we can obtain the optimal value k^* as follows:

$$k^* = \sqrt{\frac{\varepsilon \cdot c \cdot n_1 \cdot M_1}{2}}. \tag{9}$$

Since the matching limit is a positive integer and does not require to exceed the true average maximum number of matches M_1, we can obtain the final optimal matching limit k_f as follows:

$$k_f = \min(\max(round(k^*), 1), M_1). \tag{10}$$

Here, k_f represents the upper bound of the number of matches for the larger number of treatment groups. The matching upper bound for the other group can be calculated using k_f and the number of the two treatment groups. For instance, if the number of control group n_c is larger than the

number of treated group n_t (*i.e.*, $r_1 = \frac{n_t}{n_c} \leq 1$), the number of matches for the treated group will usually be higher than the number of matches for the control group. In this case, we obtain $k_1 = k_f$ and $k_2 = \max(1, round(k_1 \cdot r_1))$. On the contrary, if the number of control group n_c is smaller, we let $k_2 = k_f$ and $k_1 = \max(1, round(k_2/r_1))$. Since the upper limit is calculated based on the average maximum number of matches, the final matching limit needs to be multiplied by the preset number of neighbors N. The matching limit for treated group is $k_1 \cdot N$, and the matching limit for control group is $k_2 \cdot N$.

After determining the upper limit of the matching, we traverse all samples and find the closest N neighbors for each sample. At the beginning, the records of the number of matches for all samples are initialized to 0. When N neighbors are selected in each traversal, these N samples' matching counts are increased by 1. Samples that reach the upper limit will not be selected in subsequent matches. For each sample, the counterfactual outcome can be computed by the matched neighbors' observed outcome, as shown in Equation 4.

After calculating the counterfactual outcomes of all samples, we can estimate the treatment effect of each sample. However, calculating and perturbing the treatment effect of each individual will introduce a lot of noise, making the average treatment effect estimate inaccurate. Therefore, we choose to aggregate the potential outcomes of all samples and add Laplace noise, which can effectively reduce the impact of noise.

As shown in Figure 1, Y_1 is composed of the outcomes of the treated group, while Y_0 is composed of the outcomes of the control group. Laplace mechanism is applied to protect the samples' privacy. Note that the samples of the treated group and the control group are non-overlapping, thus these two parts can share the same privacy budget. Regarding the global sensitivity, the sensitivity of treated group is $(k_1+1)B$, which is determined by two factors: the matching upper limit of the counterfactual estimation and its own observed outcome. Similarly, we can obtain the sensitivity of control group is $(k_2+1)B$. Based on the perturbed aggregated outcomes, we can obtain the final ATE estimate $\hat{\tau}$.

Sample-level Privacy. Considering the impact of noise injection and matching error, sample-level also needs to determine a suitable maximum number of matches for each sample to achieve a promising estimation result, which is similar to label-level privacy. However, since the treatment of each sample and the matching results are perturbed, it is difficult to accurately estimate the errors caused by noise and matching. According to the source of the error, the setting of an ideal match upper limit is related to the privacy budget and the true maximum number of matches, as well as the characteristics of the dataset. Unfortunately, most of this information cannot be obtained in the sample-level setting. Inspired by Equation 9 in the label-level privacy, we set the value of matching limit in the sample-level privacy as follows:

$$k^* = \sqrt{\frac{\varepsilon_3 \cdot h \cdot n_1' \cdot M_1'}{2}}, \tag{11}$$

where ε_3 is the privacy budget used for perturbing the aggregated outcomes, h is the error coefficient, and $n_1' = \max(n_t', n_c')$ is the number of samples in the perturbed treated group or control group. M_1' is the average maximum number of matches. The calculation of M_1' is similar to M_1 in the label-level privacy. The only difference is that M_1' is calculated based on the perturbation information rather than true information. When the privacy budget is high, the noise intensity is low, and increasing k^* helps reduce the matching error. If n is high, the true matching upper limit is usually higher, which requires a higher k^*. Since the calculation result of Equation 11 may not be an integer, we further process k^* as follows:

$$k_f = \max(round(k^*), 1) \tag{12}$$

After calculating the matching limit of one group, the maximum matching upper bound of the other group can be obtained based on the number of samples in the perturbed groups. Next, we can calculate the counterfactual estimate for each sample. We then aggregate the potential outcomes of $T' = 1$ and $T' = 0$ for all samples. To satisfy DP, we utilize Laplace noise to perturb the aggregated outcomes, with the privacy budget of ε_3. Finally, we can compute the ATE estimate according to the perturbed aggregated outcomes, which is similar to the calculation at label-level.

G. Putting Things Together

The above three phases constitute the overall process of PrivATE. Due to space limitations, we defer the pseudo-code to Appendix A.

H. Algorithm Analysis

Privacy Analysis. Recalling Figure 2, PrivATE mainly consists of three phases: regression model training, similar sample sampling, and causal effect estimation. For the label-level privacy, the outcome is visited and perturbed in the causal effect estimation phase, with the privacy budget of ε.

For the sample-level privacy of PrivATE, the total privacy budget is divided into all phases. In the phase of regression model training, PrivATE needs to protect the true regression model weights and the true propensity score, which consume privacy budget of ε_{11} and ε_{12} respectively. In the second phase, the treatment is perturbed based on the privacy budget ε_2. In the causal effect estimation phase, the aggregated outcome consumes the privacy budget ε_3. Therefore, the total privacy budget is $\varepsilon = \varepsilon_{11} + \varepsilon_{12} + \varepsilon_2 + \varepsilon_3$. We obtain the following theorems, and the detailed proofs are deferred to Appendix C.

Theorem 1. *If l is label-level, Algorithm 4 satisfies ε-Label DP.*

Proof. (Sketch) In the "regression model training" and "similar sample matching" phases, no privacy budget needs to be consumed since the true observed outcome is not visited. In the "causal effect estimation" phase, the aggregated outcome is perturbed by Laplace mechanism with the privacy budget ε. Therefore, Algorithm 4 satisfies ε-Label DP. □

Theorem 2. *If l is sample-level, Algorithm 4 satisfies ε-Sample DP, where $\varepsilon = \varepsilon_{11} + \varepsilon_{12} + \varepsilon_2 + \varepsilon_3$.*

Proof. (Sketch) In the sample-level setting, all types of data are sensitive. In the first phase, both the model training and propensity score calculation use real information. Therefore, the model weights (consuming the privacy budget ε_{11}) and the propensity scores (consuming ε_{12}) are injected with Laplace noise to achieve DP. In the second phase, the true treatment is utilized to guide the sample matching. To ensure DP, the treatment is perturbed with the privacy budget ε_2. The distance matrix calculation and sorting are finished based on the perturbed information, which is regarded as post-processing. In the third phase, the outcome is perturbed based on the privacy budget ε_3, which is similar to the label-level setting. According to the sequential composition, Algorithm 4 satisfies ε-Sample DP, where $\varepsilon = \varepsilon_{11} + \varepsilon_{12} + \varepsilon_2 + \varepsilon_3$. □

Error Analysis. For label-level privacy, we theoretically analyze the error bound of the aggregated potential outcome in Theorem 3. The detailed proof is in Appendix D.

Theorem 3. *For the label-level privacy, the expected squared error of aggregated potential outcome S is bounded by $2(\frac{(k+1)B}{\varepsilon})^2 + (\frac{R}{N}B)^2$, where R is the total number of times that neighbor samples are replaced when matching without matching upper limit and matching with matching upper limit.*

Proof. (Sketch) Let $\hat{S}$ denotes the estimation of S, the expected squared error can be written as the summation of variance and the squared bias of $\hat{S}$ according to Equation 6. The variance part comes from Laplace noise, and the expected value is $2(\frac{(k+1)B}{\varepsilon})^2$. The bias part comes from the matching difference caused by whether the matching upper limit is applied, and the upper bound is $(\frac{R}{N}B)^2$. Combining the above results, we can obtain the final error bound. □

For sample-level privacy, the treatment of each sample is perturbed to satisfy DP, making the sample grouping of the original and privacy-preserving data inconsistent. Furthermore, the regression model is also perturbed, resulting in different matching results for original and privacy-preserving settings. Therefore, it is difficult to directly derive the error bound.

Complexity Analysis. We compare the time complexity and the space complexity of various methods [17, 21, 28–30]. The running time of PrivATE is significantly lower than AIM and PrivSyn. The space consumption of PrivATE is the lowest, and the space consumption of AIM is higher than that of other methods. The detailed analysis can be found in Appendix E.

IV. Evaluation

In this section, we first conduct an end-to-end experiment to illustrate the effectiveness of PrivATE in Section IV-B. Then, we conduct a hyperparameter study for label-level privacy in Section IV-C. Furthermore, we explore the impact of hyperparameter for sample-level privacy in Section IV-D.

TABLE II: Dataset Statistics.

Datasets	Treated	Control	Total	Type
IHDP [31]	139	608	747	Semi-real
Lalonde [32]	185	260	445	Real
ACIC [33]	858	3944	4802	Semi-real
Synth [18]	489	511	1000	Synthetic

A. Experimental Setup

Datasets. We run experiments on the four typical datasets, including real, semi-real, and synthetic datasets. These datasets are classic benchmarks in causal inference and are widely adopted in existing studies [17, 21]. The basic information of these four datasets are shown in Table II, and the details of these datasets are deferred to Appendix F.

Metric. To evaluate the quality of various methods [17, 21, 28, 29], we utilize the metric of relative error (RE) to show their performance. The related formula is as follows:

$$RE_{ATE} = \frac{|\hat{\tau} - \tau|}{\tau},$$

where τ is the true ATE estimate based on the non-private PSM method and $\hat{\tau}$ is the perturbed ATE estimate based on the privacy-preserving mechanism. The RE of the non-private ATE estimate is 0. A value of RE closer to 0 indicates a more accurate ATE estimate.

Competitors. In this work, we compare PrivATE with two representative approaches. The first is the existing differential private ATE estimation methods (*i.e.*, IPW-PP [17], SmoothDPM [21], and DPCI [30]), which are most comparable to our work in terms of problem assumptions and privacy protection scope. IPW-PP first uses a subset of the original dataset to learn a perturbed propensity score function, and then estimates causal effect on the remaining samples using privacy-preserving inverse probability weighting. SmoothDPM employs a smooth sensitivity-based mechanism combined with an exact matching estimator, where the matching variables are required to be discrete, to achieve privacy-preserving ATE estimation. DPCI estimates the ATE through doubly robust estimation, and guarantee DP by output perturbation. This method further estimates the differentially private variance and constructs the confidence intervals (CIs), which is beyond the focus of this paper. To ensure a fair comparison, we allocate all privacy budgets to the ATE estimation.

The second is the advanced differentially private data synthesis methods (*i.e.*, PrivSyn [29] and AIM [28]), which are another potential solution to the ATE estimation problem. Including this type of comparison helps better understand the performance of general DP synthesis schemes on this problem. By effectively capturing the correlation between various attributes of the original dataset and restoring the original distribution as much as possible, PrivSyn achieves great data synthesis performance. By following the select-measure-generate paradigm, combined with an iterative and greedy approach to select the most useful queries, AIM can achieve

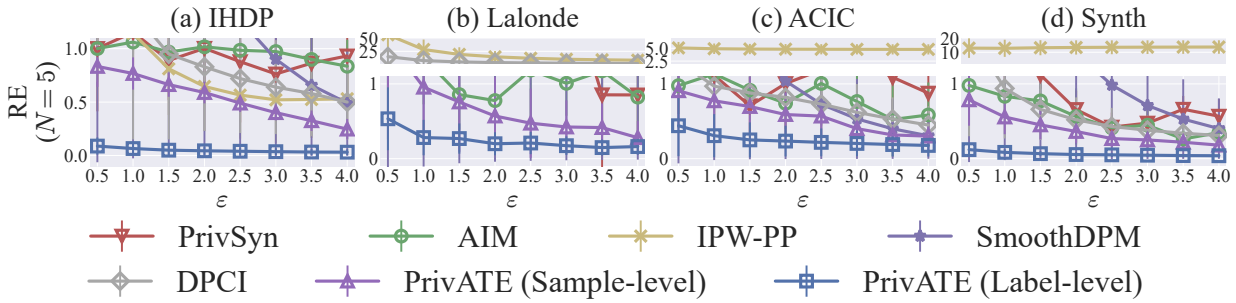

Fig. 2: End-to-end comparison of different methods when the number of matched neighbors N is 5. In each plot, the x-axis denotes the privacy budget ε, and the y-axis denotes the relative error.

low errors across a range of experimental settings compared to existing privacy-preserving data synthesis mechanisms.

Note that the above baselines guarantee (ε, δ)-Sample DP, while our proposed PrivATE can provide stricter ε-DP.

Experimental Settings. For the label-level setting of PrivATE, we set the error coefficient $c = 0.01$ in the matching limit calculation. Moreover, we have a further discussion about the choice of c in Section IV-C. For the sample-level setting of PrivATE, we set the coefficient $h = 0.001$ in the matching limit calculation. We also provide the impact of various h in Section IV-D. For the allocation of privacy budget in sample-level setting, we set $\varepsilon_{11} = \varepsilon_{12} = 0.5\varepsilon_1$, and $\varepsilon_1 : \varepsilon_2 : \varepsilon_3 = 0.1 : 0.7 : 0.2$. We also explore the impact of different privacy budget allocation in Appendix I of [34].

Implementation. We set the total privacy budget ε ranges from 0.5 to 4.0. Regarding the number of neighbors N in the counterfactual estimation, we set $N = 5$ in the experiments. In addition, we also provide the results of $N = \{1, 3, 7\}$ in Appendix G of [34]. We implement PrivATE with Python 3.8, and all experiments are conducted on a server with Intel(R) Core(TM) i7-11700K @ 3.60GHz and 128GB memory. We repeat experiment 10 times for each settings, and provide the mean and the standard variance.

B. End-to-End Evaluation

In this section, we perform an end-to-end evaluation of the two levels of PrivATE and the two types of competitors. Figure 2 illustrates the experimental results on four datasets.

In Figure 2, different columns represent various datasets. We have the following observations. First, as the privacy budget ε increases, the REs of all approaches show a downward trend. The reason is that the increase in the privacy budget reduces the noise intensity, allowing these methods to capture the feature of the original dataset in a more accurate manner. On this basis, a lower RE of ATE estimate can be obtained.

Second, the two levels of privacy protection schemes of PrivATE significantly outperform baselines on all datasets. Label-level privacy of PrivATE performs the best, followed by sample-level privacy of PrivATE. For the real Lalonde dataset, the REs of most baselines are larger than 1 even when the privacy budget is 3, while PrivATE achieves a low RE of less than 0.2. This emphasizes the superiority of PrivATE in ATE estimation. By carefully selecting the matching limit, PrivATE effectively strikes a balance between the noise perturbation and estimation error, thus obtaining a low ATE estimate error. For label-level setting of PrivATE, the privacy requirements are lower than those of sample-level, thus the regression model training and similar sample matching in this setting are better, making the REs small even when the privacy budget is small. Sample-level privacy of PrivATE needs to protect all types of variables, thus the REs are higher than label-level when ε is low. As the privacy budget increases to a certain extent (*i.e.*, $\varepsilon = 4$), the performance of the two levels is similar. Moreover, in Appendix G of [34], we find that when the number of matched neighbors N takes different values, PrivATE shows consistent and similar performance, which illustrates that PrivATE is robust to the variation of N.

IPW-PP demonstrates the poorest performance across most datasets, particularly under small privacy budgets. On the one hand, IPW-PP requires a fixed threshold to constrain the sample weights in order to achieve bounded sensitivity. However, this approach lacks flexibility and interpretability. When the privacy budget is small, the injected noise becomes excessively large, leading to highly inaccurate ATE estimates. On the other hand, to ensure privacy protection, IPW-PP splits the original dataset, using one subset to learn the propensity score function and the other to estimate the ATE. This partitioning further compromises estimation accuracy. According to the results in Figure 2, IPW-PP fails to effectively handle varying datasets and realistic privacy constraints.

SmoothDPM and DPCI fail to deliver satisfactory performance under small privacy budgets. Although these methods generally outperform IPW-PP, the estimation accuracy remains limited when strict privacy guarantees are required. When $\varepsilon \leq 1.5$, the REs of SmoothDPM and DPCI are larger than or close to 1 on most datasets, which indicates poor performance. This is primarily because they directly inject noise to the final ATE estimate, and smaller privacy budgets lead to higher noise intensity, thereby degrading accuracy. As ε increases, the estimation errors of DPCI and SmoothDPM gradually decrease. Overall, these two method struggle to achieve low error under strong privacy requirements, which highlights the necessity of PrivATE.

The relative error of PrivSyn is high, especially when the privacy budget ε is small. For instance, when the privacy budget is 1, the REs of PrivSyn exceed 1 on the four datasets,

and the RE on the Lalonde dataset even reaches 4. In contrast, the RE of PrivATE is less than 1 in all cases. This is because PrivSyn is designed to generate a new dataset that closes to the original dataset, rather than specially designed for ATE estimation. Due to the small number of samples in the Lalonde dataset and a large income gap between various individuals, PrivSyn performs poorly in capturing the true data distribution when the privacy budget is low, resulting in a high RE. We also notice that for the ACIC dataset, the REs of PrivSyn are around 1 at various privacy budgets, rather than decreasing as the budget increases. The main reason is that the dimensionality of ACIC is high, which makes it challenging for PrivSyn to capture data characteristics. In addition, the error variance of PrivSyn under various datasets and privacy settings is significantly higher than that of PrivATE, which shows that PrivSyn is not as stable as PrivATE in this task.

The performance of AIM is worse than PrivATE, but better than PrivSyn. On the one hand, AIM selects the key queries through adaptive and iterative mechanisms, improving the quality of synthetic data. On the other hand, the goal of PrivSyn and AIM is to generate a dataset similar to the original dataset. This type of method aims to achieve promising results on a variety of tasks, but not to achieve SOTA results on a specific task, such as ATE estimation. According to the experimental results, AIM and PrivSyn cannot achieve promising performance under low privacy budgets. They experience some fluctuation in RE with increasing ε, but overall show a downward trend. In contrast, PrivATE is carefully designed to balance noise-induced error and matching error in ATE estimation under DP protection. Therefore, the performance of PrivATE is better, and the trend in RE becomes more pronounced as ε increases.

C. Parameter Variation for Label-level Privacy

Choice of Matching Limit. Recalling Equation 9 and Equation 10 in Section III-F, we approximately estimate the total error caused by noise and matching, and further calculate an optimized matching limit for each sample, which can adaptively vary with the privacy budget and the dataset.

In this section, we verify the rationality of our framework by comparing this adaptive calculation with the fixed value method. In particular, we only modify the calculation of matching limit k^* in the label-level setting of PrivATE to fixed values (*i.e.*, 1, 10 and 50), and keep the other parts unchanged. Figure 3 illustrates the performance of various matching limit determination mechanisms.

From Figure 3, we observe that the fixed value method is difficult to achieve great performance across all datasets. Since the data characteristics and matching situations of different datasets are various, it is not suitable to set the same matching limit on all datasets and privacy budgets. When the matching limit is set to a small value, a low RE usually be achieved when the privacy budget ε is small. As ε increases, the impact of noise decreases, and a small matching limit cannot achieve a great result. On the other hand, larger k can perform better in high privacy budgets, but this approach will introduce significant errors when the noise intensity is high. Moreover, for different datasets, the optimal matching limits under the same privacy budget are various, making the determination of the matching upper limit more challenging.

According to the results in Figure 3, PrivATE almost achieves great performance under various settings. We find that the RE of the method with fixed small values k does not decrease significantly with the growth of the privacy budget. The reason is that a small matching limit makes the variation of noise intensity small, and the matching error does not change under a fixed k. Unlike the above, as the privacy budget ε increases, the matching limit calculated by PrivATE gradually increases, making the RE decreases. Even with a small privacy budget, PrivATE still shows promising performance on these datasets, reflecting the superiority of adaptive calculation. For the ACIC dataset, the sample size and dimensionality are both high, which makes the calculated matching limit large. When ε is small, the noise error plays a dominant role, making the RE of PrivATE high. With the increase of ε, PrivATE still shows competitive performance.

Impact of Error Coefficient in Matching Limit Calculation. In Equation 7, we utilize an error coefficient c to assist the bias estimate caused by matching in the label-level setting of PrivATE. In this section, we explore the impact of various c on final ATE estimation. A suitable c is crucial to reduce the estimation error. If the value of c is reasonable, the bias caused by matching can be estimated accurately, which helps to select an ideal matching limit. However, if c is too large or too small, the estimation of matching error will be severely distorted, resulting in an inappropriate setting of the matching limit and inaccurate estimation of ATE.

Figure 4 illustrates the RE of ATE estimate when the number of matched neighbors N is 5. We obtain the following observations. First, the impact of various c on the final RE is significant. If c is too small, it means that the bias from matching is overestimated, which will make the matching limit too low. If the value of c is too high, the influence of matching will be underestimated, making the matching limit too large. Second, for the same dataset, the optimal value c under various privacy budgets may be different. The reason is that the noise perturbation and matching error under various privacy budgets are changing. In addition, the optimal c for various datasets is also different.

In general, we find that PrivATE achieves a great performance under various privacy budgets and datasets when $c = 0.01$. We believe that this setting can accurately characterize the matching error, thus we adopt $c = 0.01$ in our experiments.

D. Parameter Variation for Sample-level Privacy

Choice of Matching Limit. Recalling Equation 11 and Equation 12 in Section III-F, we calculate the matching limit for sample-level privacy similar to the computation of label-level privacy. In this section, we evaluate the performance of the adaptive calculation of PrivATE and fixed matching limit (*i.e.*, 1, 10, and 50) methods, which is similar to the comparison in label-level privacy. Figure 5 illustrates the related results.

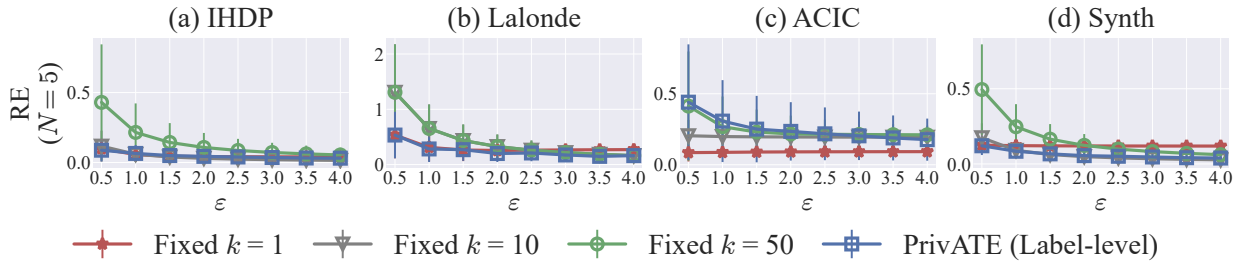

Fig. 3: Impact of different matching limit determination mechanisms in the label-level privacy of PrivATE when the number of matched neighbors N is 5. The columns represent the used datasets. In each plot, the x-axis denotes the privacy budget ε, and the y-axis denotes relative error.

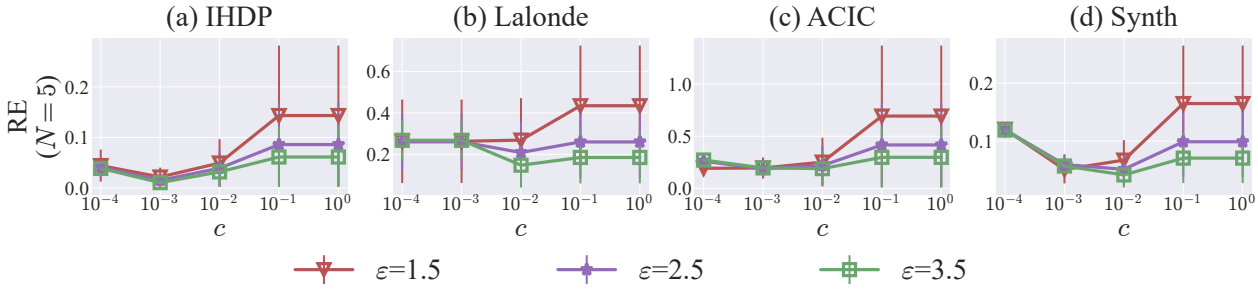

Fig. 4: Impact of different error coefficients c in the label-level privacy when the number of matched neighbors N is 5. The columns denote the used datasets. In each plot, the x-axis denotes the error coefficient c, and the y-axis denotes relative error.

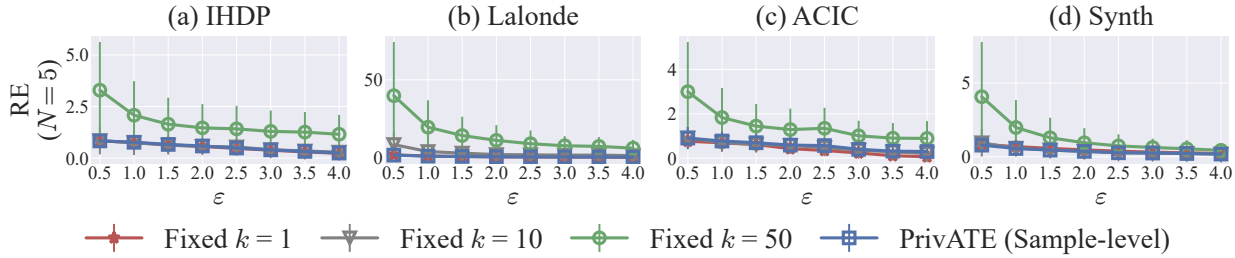

Fig. 5: Impact of different matching limit determination mechanisms in the sample-level privacy of PrivATE when the number of matched neighbors N is 5. The columns represent the used datasets. In each plot, the x-axis denotes the privacy budget ε, and the y-axis denotes relative error.

We observe that the relative error for a large fixed matching limit is significantly higher than that for a small fixed matching limit. The reason is that sample-level privacy allocates the total privacy budget across multiple phases to satisfy DP. As a result, the privacy budget available in the final phase is smaller. In addition, due to noise perturbation in both the regression model and the propensity score, the matching results are not entirely accurate. Consequently, increasing the matching limit has less impact compared to label-level privacy.

At the same time, setting the matching limit to a very small fixed value is not optimal. When noise has low interference in matching or the privacy budget is large enough, moderately increasing the matching limit can help reduce estimation error. According to the results in Figure 5, our method achieves promising performance in various settings.

Impact of Error Coefficient in Matching Limit Calculation. Recalling Equation 11 in Section III-F, an error coefficient h is utilized to help determine the value of matching limit in the sample-level privacy of PrivATE. In the section, we explore the influence of different h on the ATE estimate. Figure 6 illustrates the relative errors of various h.

We observe that a larger h tends to produce higher relative errors. The reason is that the noise is injected into each phase in the sample-level setting, thus the fidelity of regression model and matching results is significantly lower than that of the label-level. In this case, increasing h cannot effectively reduce the matching error. On the other hand, the sensitivity will increase as h grows, making the noise error higher. At the same time, a relatively small value of h may not be a great choice since this cannot reduce the matching error. Furthermore, it is impossible to find an h that achieves the lowest RE for various privacy budgets. Taking into account the impact of different budgets and datasets, we choose to set $h = 0.001$ in the experiments.

V. Discussion

Generalization. In this paper, our proposed PrivATE mainly focuses on the binary treatments. For more complex causal inference setups (*e.g.*, multi-valued treatments), the potential schemes are as follows: A multinomial logit regression model

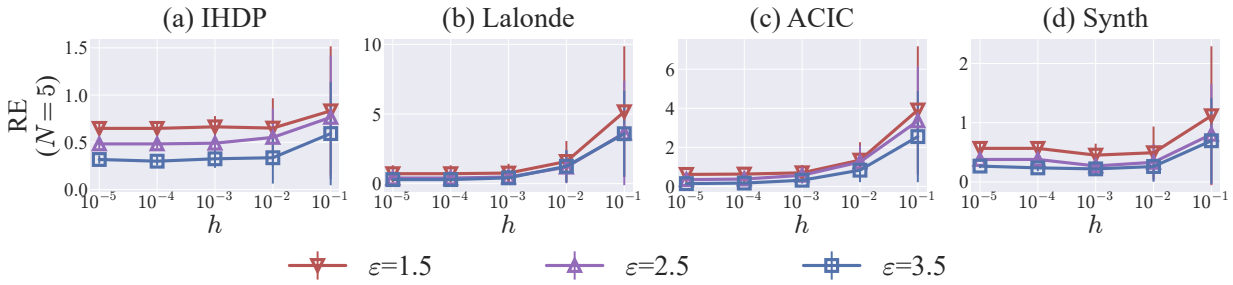

Fig. 6: Impact of different error coefficients h in the sample-level privacy when the number of matched neighbors N is 5. The columns denote the used datasets. In each plot, the x-axis denotes the error coefficient h, and the y-axis denotes relative error.

can be applied for fitting. Then, the noise can be injected into the true probability vector. The multi-dimensional treatments can be perturbed by Generalized Randomized Response (GRR). Moreover, a distance metric function can be utilized to calculate the distance between different categories and achieve pairwise privacy-preserving ATE estimation.

Scalability. For PrivATE, propensity score matching and adaptive matching algorithms involve non-trivial computation, which may be computationally slow on large-scale datasets. Here, we clarify that PrivATE achieves a low computational complexity compared to existing solutions according to the complexity analysis in Appendix E. This suggests that PrivATE has the potential to run on larger datasets. Furthermore, to address the scalability bottlenecks that matching algorithms may encounter on large datasets, the efficiency of PrivATE can be further improved by employing approaches such as approximate nearest neighbor (ANN) matching, hierarchical reduction matching, and GPU acceleration.

VI. Related Work

There are several literatures explore differentially private ATE estimation in observational studies [17–21, 30, 35]. Specifically, Guha *et al.* [19] design a differentially private weighted average treatment effect estimator for binary outcomes by splitting the data into several disjoint groups. Similarly, Lebeda *et al.* [35] also propose a data splitting-based framework to estimate the average treatment effect. Ohnishi *et al.* [20] present a differentially private covariate balancing weighting estimator to infer causal effects while protecting the privacy of covariates. Schröder *et al.* [30] further propose a framework to estimate the ATE by doubly robust estimation and construct the confidence intervals. In this setting, it is challenging to achieve promising performance under small privacy budgets, since the noise is directly injected to the ATE estimate. In addition, Lee *et al.* [17] propose a privacy-preserving inverse probability weighting (IPW) method [18] to estimate the causal effect. However, this approach relies on a pre-defined truncation threshold to bound the sample weights, which lacks flexibility and interpretability. Koga *et al.* [21] introduce a smooth-sensitivity-based DP algorithm to perturb the true average treatment effect. Nevertheless, it requires that the matching variables are discrete, and performs poorly under strong privacy constraints.

Moreover, some research incorporates differential privacy to protect real data in randomized experiments [11, 36–40]. Kancharla *et al.* [36] investigate the problem of ATE estimation in randomized controlled trials. They assume a binary outcome space and propose two consistent estimators for estimating the ATE. Betlei *et al.* [37] focus on privacy-preserving individual treatment effect (ITE) estimation and introduce a differentially private method, ADUM, which learns uplift models from data aggregated according to a given partition of the feature space. Javanmard *et al.* [41] propose a differential privacy mechanism, CLUSTER-DP, which leverages the inherent cluster structure of the data to estimate causal effects, while perturbing the outcomes to preserve individual privacy. Furthermore, Ohnishi *et al.* [38] develop a method for inferring causal effects from locally privatized data in randomized experiments. In addition, Niu *et al.* [42] introduce a meta-algorithm for estimating conditional average treatment effects using DP-EBMs [43] as the base learner. Schröder *et al.* [44] further propose a framework for conditional average treatment effects estimation that is Neyman-orthogonal.

In addition, differentially private data synthesis can also be used for differentially private ATE estimation [28, 29, 45–48]. In this way, the ATE estimate can be calculated based on a synthetic dataset that satisfies DP. Zhang *et al.* [29] design a new method to automatically and privately identify correlations in the data, and then generate sample data from a dense graphic model. McKenna *et al.* [28] propose a workload-adaptive algorithm that first selects a set of queries, then privately measures those queries, and finally generates synthetic data from the noisy measurements. However, these approaches are essentially different from our work: Their goal is to generate a synthetic dataset that closely resembles the original one under DP [49–53], while our focus is on accurately estimating ATE while satisfying DP. The experimental results also demonstrate the superiority of our method.

Moreover, there are also some other privacy-preserving solutions (*e.g.*, k-anonymity, secure multi-party computation) that can be used for data protection. For instance, Abadi *et al.* [54] propose DP-SGD, which is a classic method to train a model while ensuring the privacy of training samples. Davidson *et al.* [55] design a practical mechanism named STAR for providing cryptographically-enforced k-anonymity protections. Furthermore, Shamsabadi *et al.* [56] present Nebula, a

system for differentially private histogram estimation on data distributed among clients. Although these methods cannot be directly applied to differentially private ATE estimation, their underlying ideas offer valuable insights for future research.

VII. Conclusion

In this paper, we propose a practical framework PrivATE for estimating the average treatment effect (ATE) for observational data under differential privacy (DP). Based on propensity score matching, two different levels (*i.e.*, label-level and sample-level) of privacy protection approaches in PrivATE are proposed to accommodate various privacy requirements. To strike a great trade-off between noise and matching errors, PrivATE achieves an adaptive matching limit determination by considering the joint influence caused by noise perturbation and matching inaccuracy. Extensive experiments on four datasets demonstrate the superiority of our proposed PrivATE. We further verify the effectiveness of matching limit determination. We also analyze the impact of hyper-parameters of PrivATE and provide the guideline for their selection.

Ethics Considerations

This paper focuses on differentially private average treatment effect estimation. We strictly followed ethical guidelines by using synthetic datasets or publicly available, open-source datasets, under licenses that permit research and educational use. As these open-source datasets were curated and released by third parties, direct informed consent was not applicable. However, we are committed to ethical data use and will comply with all licensing terms for any future modifications or redistribution.

Acknowledgment

We would like to thank the anonymous reviewers for their insightful comments. This work is supported in part by the National Natural Science Foundation of China under Grants No. (62402431, 62441618, 62025206, U23A20296, U24A20237, 62402379, 72594583011, 7257010373), Key R&D Program of Zhejiang Province under Grants No. (2024C01259, 2025C01061, 2024C01065, 2024C01012, 2025C01089), the project CiCS of the research programme Gravitation which is (partly) financed by the Dutch Research Council (NWO) under Grant 024.006.037, the China Postdoctoral Science Foundation under Grants No. (2025M771501, BX20250380), and Zhejiang University.

References

[1] J. E. Brand, X. Zhou, and Y. Xie, "Recent Developments in Causal Inference and Machine Learning," *Annual Review of Sociology*, vol. 49, no. 1, pp. 81–110, 2023.

[2] T. A. Glass, S. N. Goodman, M. A. Hernán, and J. M. Samet, "Causal Inference in Public Health," *Annual Review of Public Health*, vol. 34, no. 1, pp. 61–75, 2013.

[3] H. R. Varian, "Causal Inference in Economics and Marketing," *Proceedings of the National Academy of Sciences*, vol. 113, no. 27, pp. 7310–7315, 2016.

[4] J. Pearl, "Statistics and Causal Inference: A Review," *Test*, vol. 12, pp. 281–345, 2003.

[5] E. C. Matthay and M. M. Glymour, "Causal Inference Challenges and New Directions for Epidemiologic Research on the Health Effects of Social Policies," *Current Epidemiology Reports*, vol. 9, no. 1, pp. 22–37, 2022.

[6] J. M. Cordero, V. Cristóbal, and D. Santín, "Causal Inference on Education Policies: A Survey of Empirical Studies using PISA, TIMSS and PIRLS," *Journal of Economic Surveys*, 2018.

[7] U. Shalit, F. D. Johansson, and D. Sontag, "Estimating Individual Treatment Effect: Generalization Bounds and Algorithms," in *International Conference on Machine Learning*. PMLR, 2017, pp. 3076–3085.

[8] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang, "A Survey on Causal Inference," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 15, no. 5, pp. 1–46, 2021.

[9] J. Berrevoets, F. Imrie, T. Kyono, J. Jordon, and M. Van der Schaar, "To Impute or Not to Impute? Missing Data in Treatment Effect Estimation," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2023, pp. 3568–3590.

[10] P. R. Rosenbaum and D. B. Rubin, "The Central Role of The Propensity Score in Observational Studies for Causal Effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[11] M. J. Kusner, Y. Sun, K. Sridharan, and K. Q. Weinberger, "Private Causal Inference," in *Artificial Intelligence and Statistics*. PMLR, 2016, pp. 1308–1317.

[12] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating Noise to Sensitivity in Private Data Analysis," in *Theory of Cryptography Conference*. Springer, 2006, pp. 265–284.

[13] A. Bittau, Ú. Erlingsson, P. Maniatis, I. Mironov, A. Raghunathan, D. Lie, M. Rudominer, U. Kode, J. Tinnes, and B. Seefeld, "Prochlo: Strong Privacy for Analytics in the Crowd," in *SOSP*, 2017.

[14] R. Rogers, S. Subramaniam, S. Peng, D. Durfee, S. Lee, S. K. Kancha, S. Sahay, and P. Ahammad, "Linkedin's Audience Engagements Api: A Privacy Preserving Data Analytics System at Scale," *CoRR abs/2002.05839*, 2020.

[15] Q. Yuan, M. Sun, Y. Sheng, and Q. Guo, "PrivCPM: Privacy-Preserving Cooperative Pricing Mechanism in Coupled Power-Traffic Networks," *IEEE Transactions on Smart Grid*, vol. 16, no. 1, pp. 612–626, 2025.

[16] T. Wang, J. Q. Chen, Z. Zhang, D. Su, Y. Cheng, Z. Li, N. Li, and S. Jha, "Continuous Release of Data Streams under both Centralized and Local Differential Privacy," in *ACM CCS*, 2021, pp. 1237–1253.

[17] S. K. Lee, L. Gresele, M. Park, and K. Muandet, "Privacy-preserving Causal Inference via Inverse Probability Weighting," *CoRR abs/1905.12592*, 2019.

[18] G. W. Imbens and D. B. Rubin, *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

[19] S. Guha and J. P. Reiter, "Differentially Private Estimation of Weighted Average Treatment Effects for Binary Outcomes," *Computational Statistics & Data Analysis*, p. 108145, 2025.

[20] Y. Ohnishi and J. Awan, "Differentially Private Covariate Balancing Causal Inference," *CoRR abs/2410.14789*, 2024.

[21] T. Koga, K. Chaudhuri, and D. Page, "Differentially Private Multi-Site Treatment Effect Estimation," in *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE, 2024, pp. 472–489.

[22] P. R. Rosenbaum, "Model-based Direct Adjustment," *Journal of the American statistical Association*, vol. 82, no. 398, pp. 387–394, 1987.

[23] C. Dwork, A. Roth *et al.*, "The Algorithmic Foundations of Differential Privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[24] Y. Wang, X. Wu, and D. Hu, "Using Randomized Response for Differential Privacy Preserving Data Collection," in *EDBT/ICDT Workshops*, vol. 1558, 2016, pp. 0090–6778.

[25] Z. Zhang, T. Wang, N. Li, S. He, and J. Chen, "CALM: Consistent Adaptive Local Marginal for Marginal Release under Local Differential Privacy," in *ACM CCS*, 2018, pp. 212–229.

[26] L. Du, Z. Zhang, S. Bai, C. Liu, S. Ji, P. Cheng, and J. Chen, "AHEAD: Adaptive Hierarchical Decomposition for Range Query under Local Differential Privacy," in *ACM CCS*, 2021, pp. 1266–1288.

[27] C. Wei, M. Zhao, Z. Zhang, M. Chen, W. Meng, B. Liu, Y. Fan, and W. Chen, "DPMLBench: Holistic Evaluation of Differentially Private Machine Learning," in *ACM CCS*, 2023, pp. 2621–2635.

[28] R. McKenna, B. Mullins, D. Sheldon, and G. Miklau, "AIM: An Adaptive and Iterative Mechanism for Differentially Private Synthetic Data," *Proceedings of the VLDB Endowment*, vol. 15, no. 11, pp. 2599–2612, 2022.

[29] Z. Zhang, T. Wang, J. Honorio, N. Li, M. Backes, S. He, J. Chen, and Y. Zhang, "PrivSyn: Differentially Private Data Synthesis," in *USENIX Security Symposium*, 2021.

[30] M. Schröder, J. Hartenstein, and S. Feuerriegel, "PrivATE: Differentially Private Confidence Intervals for Average Treatment Effects," *CoRR abs/2505.21641*, 2025.
[31] J. L. Hill, "Bayesian Nonparametric Modeling for Causal Inference," *Journal of Computational and Graphical Statistics*, vol. 20, no. 1, pp. 217–240, 2011.
[32] R. H. Dehejia and S. Wahba, "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 1999.
[33] V. Dorie, J. Hill, U. Shalit, M. Scott, and D. Cervone, "Automated Versus Do-it-yourself Methods for Causal Inference: Lessons Learned from a Data Analysis Competition," *Statistical Science*, 2019.
[34] Q. Yuan, X. Li, L. Du, M. Chen, M. Sun, Y. Gao, S. He, J. Chen, and Z. Zhang, "PrivATE: Differentially Private Average Treatment Effect Estimation for Observational Data," *CoRR abs/2512.14557*, 2025.
[35] C. Lebeda, M. Even, A. Bellet, and J. Josse, "Model Agnostic Differentially Private Causal Inference," *CoRR abs/2505.19589*, 2025.
[36] M. Kancharla and H. Kang, "A Robust, Differentially Private Randomized Experiment for Evaluating Online Educational Programs with Sensitive Student Data," *CoRR abs/2112.02452*, 2021.
[37] A. Betlei, T. Gregoir, T. Rahier, A. Bissuel, E. Diemert, and M.-R. Amini, "Differentially Private Individual Treatment Effect Estimation from Aggregated Data," in *FOCS Workshop*, 2021.
[38] Y. Ohnishi and J. Awan, "Locally Private Causal Inference for Randomized Experiments," *Journal of Machine Learning Research*, vol. 26, no. 14, pp. 1–40, 2025.
[39] A. Farzam and G. Sapiro, "Causal Inference under Differential Privacy: Challenges and Mitigation Strategies," in *NeurIPS 2024 Causal Representation Learning Workshop*, 2024.
[40] S. Mukherjee, A. Mustafi, A. Slavkovic, and L. Vilhuber, "Improving Privacy for Respondents in Randomized Controlled Trials: A Differential Privacy Approach," National Bureau of Economic Research, Tech. Rep., 2024.
[41] A. Javanmard, V. Mirrokni, and J. Pouget-Abadie, "Causal Inference with Differentially Private (Clustered) Outcomes," *CoRR abs/2308.00957*, 2023.
[42] F. Niu, H. Nori, B. Quistorff, R. Caruana, D. Ngwe, and A. Kannan, "Differentially Private Estimation of Heterogeneous Causal Effects," in *Conference on Causal Learning and Reasoning*, 2022, pp. 618–633.
[43] H. Nori, R. Caruana, Z. Bu, J. H. Shen, and J. Kulkarni, "Accuracy, Interpretability, and Differential Privacy via Explainable Boosting," in *International Conference on Machine Learning*, 2021, pp. 8227–8237.
[44] M. Schröder, V. Melnychuk, and S. Feuerriegel, "Differentially Private Learners for Heterogeneous Treatment Effects," in *ICLR*, 2025.
[45] R. McKenna, D. Sheldon, and G. Miklau, "Graphical-Model Based Estimation and Inference for Differential Privacy," in *International Conference on Machine Learning*. PMLR, 2019, pp. 4435–4444.
[46] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao, "PrivBayes: Private Data Release via Bayesian Networks," *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 4, pp. 1–41, 2017.
[47] G. Vietri, G. Tian, M. Bun, T. Steinke, and S. Wu, "New Oracle-Efficient Algorithms for Private Synthetic Data Release," in *International Conference on Machine Learning*. PMLR, 2020, pp. 9765–9774.
[48] K. Cai, X. Lei, J. Wei, and X. Xiao, "Data Synthesis via Differentially Private Markov Random Fields," *Proceedings of the VLDB Endowment*, vol. 14, no. 11, pp. 2190–2202, 2021.
[49] Q. Yuan, Z. Zhang, L. Du, M. Chen, P. Cheng, and M. Sun, "PrivGraph: Differentially Private Graph Data Publication by Exploiting Community Information," in *USENIX Security Symposium*, 2023.
[50] Q. Yuan, H. Wu, S. He, and M. Sun, "PrivLoad: Privacy-preserving Load Profiles Synthesis Based on Diffusion Models," *IEEE Transactions on Smart Grid*, vol. 16, no. 6, pp. 5628–5640, 2025.
[51] Y. Du, Y. Hu, Z. Zhang, Z. Fang, L. Chen, B. Zheng, and Y. Gao, "LDPTrace: Locally Differentially Private Trajectory Synthesis," *Proceedings of the VLDB Endowment*, 2023.
[52] H. Wang, Z. Zhang, T. Wang, S. He, M. Backes, J. Chen, and Y. Zhang, "PrivTrace: Differentially Private Trajectory Synthesis by Adaptive Markov Model," in *USENIX Security Symposium*, 2023.
[53] Q. Yuan, Z. Zhang, L. Du, M. Chen, M. Sun, Y. Gao, M. Backes, S. He, and J. Chen, "PSGraph: Differentially Private Streaming Graph Synthesis by Considering Temporal Dynamics," *CoRR abs/2412.11369*, 2024.
[54] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep Learning with Differential Privacy," in *ACM CCS*, 2016, pp. 308–318.
[55] A. Davidson, P. Snyder, E. Quirk, J. Genereux, B. Livshits, and H. Haddadi, "STAR: Secret Sharing for Private Threshold Aggregation Reporting," in *ACM CCS*, 2022, pp. 697–710.
[56] A. S. Shamsabadi, P. Snyder, R. Giles, A. Bellet, and H. Haddadi, "Nebula: Efficient, Private and Accurate Histogram Estimation," in *ACM CCS*, 2025.

APPENDIX

A. *Workflow of* PrivATE

Algorithm 4 illustrates the overall process of PrivATE. In the label-level privacy, only the third phase needs to consume privacy budget to perturb the outcomes. While in the sample-level privacy, all three phases need to allocate the privacy budget to protect the privacy of all types of data.

Algorithm 4: PrivATE

Input: Original dataset D, privacy level l, privacy budget ε (label-level) or $\varepsilon = \varepsilon_{11} + \varepsilon_{12} + \varepsilon_2 + \varepsilon_3$ (sample-level), the number of matched neighbors N, the maximum variation range of outcome B, the error coefficient c (label-level) or h (sample-level)

Output: Propensity score $e'(X)$, treatment T', sorted matrices H, average treatment effect estimate $\hat{\tau}$

if *l is label-level* **then**
  `// Phase 1: Regression model training`
  $e'(X) \leftarrow$ Algorithm 1(D, l)
  `// Phase 2: Similar sample matching`
  $T', H \leftarrow$Algorithm 2$(D, e'(X), l)$
  `// Phase 3: Causal effect estimation`
  $\hat{\tau} \leftarrow$Algorithm 3$(D, H, N, T', l, B, \varepsilon, c)$
else
  `// Phase 1: Regression model training`
  $e'(X) \leftarrow$ Algorithm 1$(D, l, \varepsilon_{11}, \varepsilon_{12})$
  `// Phase 2: Similar sample matching`
  $T', H \leftarrow$Algorithm 2$(D, e'(X), l, \varepsilon_2)$
  `// Phase 3: Causal effect estimation`
  $\hat{\tau} \leftarrow$Algorithm 3$(D, H, N, T', l, B, \varepsilon_3, h)$

B. *Proof of Theorem 1*

Proof. In the label-level setting, only the outcome is sensitive information. In the first two phases, the outcome does not need to be accessed, thus does not consume the privacy budget. In the third phase, the original potential outcomes of all samples are aggregated to calculate the potential outcome sum of the treated and control groups. We utilize Laplace mechanism to perturb the original aggregated outcome, thus this step satisfies ε-DP. Note that the samples of the treated and control groups are non-overlapping, thus these two parts can share the same privacy budget according to the parallel composition. Finally,

the final ATE estimate is computed based on the perturbed aggregated outcome, which can be considered post-processing and does not consume the privacy budget. Therefore, if the privacy level l is label-level, Algorithm 4 satisfies ε-Label DP. □

Lemma 4. *Let $G(w)$ and $g(w)$ be two vector-based functions, which are continuous, and differentiable at all points. Moreover, let $G(w)$ and $G(w)+g(w)$ be λ_1-strongly convex in L_1-norm. If $w_1 = \arg\min_w G(w)$ and $w_2 = \arg\min_w G(w) + g(w)$, then*

$$||w_1 - w_2||_1 \leq \frac{1}{\lambda_1} \max_w ||\nabla g(w)||_\infty$$

Proof. Using the definition of w_1 and w_2, and the fact that G and g are continuous and differentiable everywhere,

$$\nabla G(w_1) - \nabla G(w_2) = \nabla g(w_2)$$

As $G(w)$ is λ_1-strongly convex, then

$$\begin{aligned}(\nabla G(w_1) - \nabla G(w_2))^T(w_1 - w_2) &= (\nabla g(w_2))^T(w_1 - w_2) \\ &\geq \lambda_1 ||w_1 - w_2||_1^2\end{aligned}$$

Based on Hölder inequality, we have

$$(\nabla g(w_2))^T(w_1 - w_2) \leq ||\nabla g(w_2)||_\infty ||w_1 - w_2||_1$$

Then, we obtain

$$\lambda_1 ||w_1 - w_2||_1^2 \leq ||\nabla g(w_2)||_\infty ||w_1 - w_2||_1$$

Further, we have

$$\lambda_1 ||w_1 - w_2||_1 \leq ||\nabla g(w_2)||_\infty$$

Finally, we can obtain

$$||w_1 - w_2||_1 \leq \frac{1}{\lambda_1} \max_w ||\nabla g(w)||_\infty$$

□

C. Proof of Theorem 2

Proof 1: Private model training satisfies ε_{11}-DP.

Proof. Assuming that the dataset D and the adjacent dataset D' differ in the i-th sample, the difference between their corresponding loss functions is as follows:

$$\begin{aligned}g(w) &= J(w, D) - J(w, D') \\ &= \frac{1}{n}[\log(1 + e^{-X_i^T w t_i}) - \log(1 + e^{-X'^T_i w t'_i})]\end{aligned}$$

We observe that the logarithmic loss function $l = \log(1 + e^{-X^T wt})$ is convex and differentiable, and $J(w)$ is λ-strongly convex. Then, we can obtain the derivative of $g(w)$ as follows:

$$\nabla g(w) = \frac{1}{n}[X_i^T l'(X_i^T w t_i) t_i - X'^T_i l'(X'^T_i w t'_i) t'_i],$$

where $l'(\cdot)$ stands for the derivative of the logarithmic loss function, and its range is $[0, 1]$. Then, we obtain

$$||\nabla g(w)||_\infty = \frac{1}{n}||X_i^T l'(X_i^T w t_i) t_i - X'^T_i l'(X'^T_i w t'_i) t'_i||_\infty$$

For vector v, we have $||v||_\infty = \max_j |v_j|$. Since $|l'(\cdot)t| \leq 1$, we further obtain

$$|\nabla g(w)_j| \leq \frac{1}{n}|X_{i,j} - X'_{i,j}| \leq \frac{1}{n}(|X_{i,j}| + |X'_{i,j}|) \leq \frac{2}{n}$$

Therefore, we can obtain

$$\max_w ||\nabla g(w)||_\infty \leq \frac{2}{n}$$

It is obvious that $\frac{\lambda}{2}||w||_2^2$ is λ-convex in L_2-norm. Next, we need to convert it to L_1-norm strongly convex. For vector v, $||v||_1 \leq \sqrt{d}||v||_2$, then

$$\lambda ||w_1 - w_2||_2^2 \geq \lambda \frac{1}{d} ||w_1 - w_2||_1^2$$

Therefore, the regularization term $\frac{\lambda}{2}||w||^2$ is $\frac{\lambda}{d}$-strongly convex in L_1-norm (*i.e.*, $\lambda_1 = \frac{\lambda}{d}$). Based on Lemma 4, we can obtain

$$||w_1 - w_2||_1 \leq \frac{1}{\lambda_1} \max_w ||\nabla g(w)||_\infty \leq \frac{d}{\lambda} \cdot \frac{2}{n} = \frac{2d}{n\lambda}$$

Therefore, the L_1-sensitivity of w is $\frac{2d}{n\lambda}$. The Laplace noise with the privacy budget of ε_{11} is adopted to perturb the true weights, thus the privacy model training satisfies ε_{11}-DP. □

Proof 2: Private score calculation satisfies ε_{12}-DP.

Proof. Different samples are independent of each other, thus the same privacy budget ε_{12} can be used to add Laplace noise to their propensity scores. According to the parallel composition introduced in Section II-B, the step of private score calculation satisfies ε_{12}-DP. □

Proof 3: Similar sample matching satisfies ε_2-DP.

Proof. For treatment T, the random response mechanism is applied to protect the sample's privacy. According to Section II-B, the random response mechanism meets the requirements of DP. The distance calculation between various samples and distance sorting are based on the perturbed T' and $e'(X)$. According to the post-processing property of DP, these steps do not incur additional privacy loss. Following the sequential composition of DP, the similar sample matching phase satisfies ε_2-DP. □

Proof 4: Causal effect estimation satisfies ε_3-DP.

Proof. The matching limit is calculated without touching the true data, thus it meets DP. The counterfactual outcome is perturbed by Laplace noise with the privacy budget of ε_3. The ATE estimation is finished based on the noisy outcomes. According to the sequential composition and post-processing of DP, causal effect estimation satisfies ε_3-DP. □

Overall Privacy Budget. According to the above proofs, in the first phase, the private regression model training satisfies

ε_{11}-DP, and the private propensity score calculation satisfies ε_{12}-DP. In the second phase, the similar sample matching satisfies ε_2-DP. In the third phase, the causal effect estimation satisfies ε_3-DP. Based on the sequential composition of DP, we obtain that if l is set to sample-level, Algorithm 4 satisfies ε-Sample DP, where $\varepsilon = \varepsilon_{11} + \varepsilon_{12} + \varepsilon_2 + \varepsilon_3$.

D. Proof of Theorem 3

Proof. Here, we provide the derivation for the outcome sum of the treated group S_1 (the derivation of the control group S_0 is similar). According to Equation 6, let $\hat{S}$ denote the estimation of S, the expected square error $\mathbb{E}[(\hat{S} - S)^2]$ can be written as the summation of variance and the squared bias of $\hat{S}$, *i.e.*, $\mathbb{E}[(\hat{S} - S)^2] = \mathsf{Var}[\hat{S}] + \mathsf{Bias}[\hat{S}]^2$.

The variance of $\hat{S}$ comes from Laplace noise, given the maximum variation range of the outcome B, the matching upper limit for each sample k and the privacy budget ε, we obtain the expected error of variance part is $\mathsf{Var}[\hat{S}] = 2(\frac{(k+1)B}{\varepsilon})^2$.

The error of bias part comes from the matching difference caused by whether the matching upper limit is applied. For each sample $j \in \{1, 2, ..., n_c\}$ in the control group, let u_j represents the number of times sample j is selected as a neighbor in the original nearest neighbor matching. Then, we can obtain the total number of times $R = \sum_{j=1}^{n_c} \max(0, u_j - k)$ that neighbor samples are replaced when matching without the matching upper limit and when matching with the matching upper limit. Further, we obtain $\mathsf{Bias}[\hat{S}]^2 = |\mathbb{E}(\hat{S}) - S|^2 \leq (\frac{R}{N}B)^2$, where N is the number of neighbors for each sample in the matching.

Based on the above derivation, we can obtain:

$$\mathbb{E}[(\hat{S} - S)^2] \leq 2(\frac{(k+1)B}{\varepsilon})^2 + (\frac{R}{N}B)^2$$

□

E. Complexity Analysis

In this section, we analyze the computational complexity of various methods, and quantitatively evaluate their running time and memory consumption.

Time Complexity. We provide the time complexity by analyzing each phase of the algorithms. The number of samples is n, and the number of covariates is d.

For label-level privacy of PrivATE, the goal of the first phase is to train a logistic regression model. The time complexity of model training is $\mathcal{O}(nd)$. In the second phase, we need to calculate and sort the distance between the sample and other samples in the opposite treatment group, $\mathcal{O}(n \log n)$. In the third phase, the counterfactual outcome of each sample is estimated and the potential outcomes are aggregated to compute the final ATE. The time complexity is $\mathcal{O}(nN)$, where N is the number of neighbors in the counterfactual estimation. Above all, we obtain the total time complexity is $\mathcal{O}(nd + n \log n + nN)$. For sample-level privacy of PrivATE, additional noise injection will incur a time complexity of $\mathcal{O}(n) < \mathcal{O}(n \log n)$. Therefore, the time complexity of sample-level privacy is also $\mathcal{O}(nd + n \log n + nN)$.

For IPW-PP, the time complexity of propensity score model training is $\mathcal{O}(n_1 d)$, where n_1 is the number of samples used to train the model. The time complexity of differentially private ATE estimation phase is $\mathcal{O}(n_2 d)$, where $n_2 = n - n_1$ is the number of samples used to estimate ATE. Since $\mathcal{O}(n_1 d) + \mathcal{O}(n_2 d) < \mathcal{O}(nd)$, the total time complexity of IPW-PP is $\mathcal{O}(nd)$.

For SmoothDPM, the time complexity of matching is $\mathcal{O}(n)$, and the time complexity of smooth sensitivity calculation is $\mathcal{O}(ng)$, where g is the number of different discrete covariate combinations. Therefore, the time complexity of SmoothDPM is $\mathcal{O}(ng)$.

For DPCI, the time complexity of nuisance model fitting is typically $\mathcal{O}(nd)$. The time complexity of ATE calculation is $\mathcal{O}(n)$. Therefore, the total time complexity of DPCI is $\mathcal{O}(nd)$.

For PrivSyn, in the marginal selection step, there are $p = \frac{d(d-1)}{2}$ possible pairwise marginals. In the i-th iteration of marginal selection algorithm, $(p - i)$ pairwise marginals need to be checked; thus the time complexity is $\sum_{i=1}^{m}(p - i) = mp - \frac{m(m+1)}{2} = \mathcal{O}(md^2)$, where m is the number of marginals. In the dataset generation step, PrivSyn should go through all marginals r times to ensure consistency. Thus, the time complexity is mr, and the overall time complexity is $\mathcal{O}(md^2 + mr)$.

For AIM, all 1-way marginals are measured in the beginning, and the time complexity is $\mathcal{O}(nd)$. In the iterative stage, the noise is injected into the selected v-way marginals in each round. The corresponding time complexity is $\mathcal{O}(Tqnv)$, where T is the number of iteration, and q is the number of candidate queries in each round. Therefore, the total time complexity of AIM is $\mathcal{O}(nd + Tqnv)$.

Space Complexity. For PrivATE, it requires storing the covariate matrix and the propensity score, thus the space complexity of PrivATE is $\mathcal{O}(nd)$. For IPW-PP, the memory consumption mainly comes from the original dataset and model training, and the corresponding space complexity is $\mathcal{O}(nd)$. For SmoothDPM, it requires storing the original dataset and the grouping information, thus the space time complexity is $\mathcal{O}(nd+g)$. For DPCI, it mainly requires storing the dataset and nuisance models, thus the space complexity is $\mathcal{O}(nd)$. For PrivSyn, the memory consumption consist of two parts, *i.e.*, marginal tables and synthetic dataset. The memory comsumption of marginal tables is the product of the number of marginals m and the average number of cells for each marginal C, and the memory consumption of synthetic dataset is $\mathcal{O}(nd)$. Therefore, the space complexity of PrivSyn is $\mathcal{O}(mC + nd)$. For AIM, it requires to store the original dataset with the space complexity of $\mathcal{O}(nd)$. In the iteration process, AIM needs to store the noise measurement values. The space complexity is $\mathcal{O}(Th^v)$, where h^v is the domain size for any v-way marginal. In addition, AIM also needs to storage the Private-PGM model parameters, with the space complexity of $\mathcal{O}(S)$, where S is the junction tree size. Therefore, the total space complexity of AIM is $\mathcal{O}(nd + Th^v + S)$.

Empirical Evaluation. Table IV and Table V show the running time and the memory consumption for all methods

TABLE III: Comparison of computational complexity.

Methods	Time	Space
IPW-PP	$\mathcal{O}(nd)$	$\mathcal{O}(nd)$
SmoothDPM	$\mathcal{O}(ng)$	$\mathcal{O}(nd+g)$
DPCI	$\mathcal{O}(nd)$	$\mathcal{O}(nd)$
PrivSyn	$\mathcal{O}(md^2+mr)$	$\mathcal{O}(mC+nd)$
AIM	$\mathcal{O}(nd+Tqnv)$	$\mathcal{O}(nd+Th^v+S)$
PrivATE (sample)	$\mathcal{O}(nd+n\log n+nN)$	$\mathcal{O}(nd)$
PrivATE (label)	$\mathcal{O}(nd+n\log n+nN)$	$\mathcal{O}(nd)$

TABLE IV: Comparison of running time (measured by seconds).

Methods	Datasets			
	IHDP	Lalonde	ACIC	Synth
IPW-PP	0.29s	0.16s	0.32s	0.06s
SmoothDPM	0.03s	0.04s	0.06s	0.02s
DPCI	0.09s	0.04s	0.94s	0.08s
PrivSyn	67.02s	5.18s	4861.72s	8.11s
AIM	210.35s	22.24s	12569.72s	104.63s
PrivATE (sample)	0.08s	0.04s	0.68s	0.05s
PrivATE (label)	0.06s	0.02s	0.55s	0.04s

TABLE V: Comparison of memory consumption (measured by Megabytes).

Methods	Datasets			
	IHDP	Lalonde	ACIC	Synth
IPW-PP	471.13	470.92	493.67	472.70
SmoothDPM	469.13	468.92	477.31	470.81
DPCI	479.84	476.26	483.60	476.05
PrivSyn	540.44	534.98	572.21	537.60
AIM	1712.74	933.76	4177.00	1043.12
PrivATE (sample)	470.60	469.97	484.55	470.92
PrivATE (label)	468.19	467.46	479.83	468.29

on the four datasets (see their details in Table II).

The empirical running time in Table IV illustrates that the performance of differentially private ATE estimation methods is better than synthesis-based methods. The running time of IPW-PP, SmoothDPM and PrivATE on the four datasets is smaller than 1s, which reflects their high efficiency. PrivATE incurs longer runtime on the ACIC dataset due to its larger size, which increases the computational cost of both model training and distance sorting. The running time of sample-level of PrivATE is slightly higher than label-level since sample-level requires additional noise injection. In addition, the running time of PrivSyn is significantly longer than PrivATE since it requires capturing the features of many marginals. AIM exhibits the longest time due to a large number of iterations.

Table V shows the memory consumption. The consumption of PrivATE is the lowest. We find that the consumption of PrivATE, IPW-PP and SmoothDPM is close because most of the same memory is used to store the original dataset. The consumption of AIM is significantly higher than that of other methods since it requires to storage the selected v-way marginals and the junction tree.

F. Dataset Description

The details of the four datasets are as follows.

- **IHDP [31].** The Infant Health and Development Program (IHDP) dataset is a semi-real dataset, where only the outcome value is simulated. This dataset includes $n = 747$ individuals comprised of $n_t = 139$ treated and $n_c = 608$ control individuals. The treatment involves specialist home visits for children, while the outcome is their future cognitive test scores. The dimension of the covariates is $d = 25$, including demographic information, infant health, socioeconomic status, *etc.*
- **Lalonde [32].** The Lalonde dataset is a real dataset that comes from an evaluation study of the national supported work (NSW) program. In particular, the experimental treated group from the NSW study (which received job training) is combined with a non-experimental control group drawn from observational surveys. This dataset is composed of $n = 445$ individuals, where $n_t = 185$ individuals belong to the treated group and $n_c = 260$ individuals belong to the control group. The treatment refers to whether or not to participate in the NSW program, and the outcome is earning in 1978. This dataset includes $d = 8$ dimensions of covariates, such as the age, years of education, *etc.*
- **ACIC [33].** ACIC comes from the Atlantic Causal Inference Conference competition in 2016 for causal challenges. The dataset used in this competition is semi-real, *i.e.*, the covariates are real, while the treatment and outcome are synthetic. In our experiment, there are a total of $n = 4802$ samples, of which $n_t = 858$ samples are in the treated group and $n_c = 3944$ samples are in the control group.
- **Synth [18].** This dataset is a synthetic dataset. We adopt the simulated method introduced in [21] to generate this dataset. First, we simulate a covariate matrix. A total of $n = 1000$ samples are generated, each characterized by $d = 20$ covariates. These covariates are independently sampled from a uniform distribution. To mimic the selection bias inherent in observational data, the treatment assignment T is made dependent on the covariates. We employ a logistic model where the propensity score for each sample is given by $e(X_i) = sigmoid(a \cdot (2X_i - 1))$. The parameter a, which controls the extent of selection bias, is drawn from the uniform distribution. Next, the observed outcome Y for each sample is synthesized using a linear response function: $Y_i = b \cdot X_i + \tau \cdot T_i + q_i$. Here, b is a vector of coefficients sampled from a uniform distribution, representing the heterogeneous influence of each covariate on the outcome. The scalar τ is set to 0.5, which stands for the ATE estimate. q_i is an independent noise term, also sampled from a uniform distribution, which introduces random variability into the outcome. After the above process, the treatment assignment yields a naturally imbalanced split, resulting in a final dataset with $n_t = 489$ samples in the treated group and $n_c = 511$ samples in the control group.