A Two-Stage Pretraining-Finetuning Framework for Treatment Effect Estimation with Unmeasured Confounding

Chuan Zhou*
zhouchuancn@pku.edu.cn
Peking University & MBZUAI
Beijing, China

Yaxuan Li*†
yaxuanli.cn@gmail.com
Peking University
Beijing, China

Chunyuan Zheng
cyzheng@stu.pku.edu.cn
Peking University & MBZUAI
Beijing, China

Haiteng Zhang
zhanghaiteng22@mails.ucas.ac.cn
Chinese Academy of Sciences & MBZUAI
Beijing, China

Min Zhang
mzhang@cs.ecnu.edu.cn
East China Normal University & MBZUAI
Shanghai, China

Haoxuan Li‡
hxli@stu.pku.edu.cn
Peking University & MBZUAI
Beijing, China

Mingming Gong‡
mingming.gong@unimelb.edu.au
The University of Melbourne & MBZUAI
Melbourne, Australia

Abstract

Estimating the conditional average treatment effect (CATE) from observational data plays a crucial role in areas such as e-commerce, healthcare, and economics. Existing studies mainly rely on the strong ignorability assumption that there are no unmeasured confounders, whose presence cannot be tested from observational data and can invalidate any causal conclusion. In contrast, data collected from randomized controlled trials (RCT) do not suffer from confounding, but are usually limited by a small sample size. In this paper, we propose a two-stage pretraining-finetuning (TSPF) framework using both large-scale observational data and small-scale RCT data to estimate the CATE in the presence of unmeasured confounding. In the first stage, a foundational representation of covariates is trained to estimate counterfactual outcomes through large-scale observational data. In the second stage, we propose to train an augmented representation of the covariates, which is concatenated to the foundational representation obtained in the first stage to adjust for the unmeasured confounding. To avoid overfitting caused by the small-scale RCT data in the second stage, we further propose a partial parameter initialization approach, rather than training a separate network. The superiority of our approach is validated on two public datasets with extensive experiments. The code is available at https://github.com/zhouchuanCN/KDD25-TSPF.

CCS Concepts

• **Computing methodologies → Machine learning approaches**.

Keywords

Causal Effect Estimation, Unmeasured Confounding, Data Fusion

ACM Reference Format:
Chuan Zhou, Yaxuan Li, Chunyuan Zheng, Haiteng Zhang, Min Zhang, Haoxuan Li, and Mingming Gong. 2025. A Two-Stage Pretraining-Finetuning Framework for Treatment Effect Estimation with Unmeasured Confounding. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1 (KDD '25), August 3–7, 2025, Toronto, ON, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3690624.3709161

1 Introduction

The conditional average treatment effect (CATE) is the average causal effect of a treatment or an intervention on the outcome of interest given the covariates [46], which plays an important role in diverse fields, such as e-commerce [62], healthcare [48], and economics [30]. In e-commerce, the platforms desire to predict how recommending a specific product to a particular user affects the probability of purchase [43], and thereby influence the total profit. In healthcare, doctors assess the potential outcome for different patient groups when administering a certain treatment [15] for precision medicine. Similarly in economics, the policymakers evaluate how much a job training program will raise employment opportunities for unemployed individuals [9].

To enhance the accuracy of CATE estimation, representation-based learning approaches have gathered increasing attention due to their impressive performance [34, 39, 59, 72]. These approaches focus on generating covariate representations, with the objective of

*Both authors contributed equally to this research.
†This work was done during the research internship of Yaxuan Li at Peking University.
‡Haoxuan Li and Mingming Gong are corresponding authors.

KDD '25, August 3–7, 2025, Toronto, ON, Canada

ACM ISBN 979-8-4007-1245-6/25/08
https://doi.org/10.1145/3690624.3709161

mitigating confounding bias by minimizing distributional discrepancies between the treatment group and control group. To obtain such representations, previous approaches have developed substantial theory and explored extensive practice. For instance, some of them use integral probability metric (IPM) for regularization [34], while a few approaches emphasize local similarity preservation [64], targeted learning [70], and optimal transport [59].

However, the aforementioned methods may ignore unmeasured confounding, which is very common in real-world scenarios. In our e-commerce example, the financial status of users should be sensitive and cannot be collected [42]. In the healthcare case, the personal lifestyles of patients are difficult to obtain [10]. For the example in economics, personal working status is difficult to measure [63]. These unmeasured variables can affect the treatment and outcome simultaneously, which causes confounding bias. Therefore, proposing methods to account for the confounding bias is crucial to accurately estimate CATE.

To address unmeasured confounding, one category of mainstream methods can only rely on large-scale observational (OBS) data, including sensitivity analysis [17, 31], front door adjustment methods [21, 73], and instrumental variables methods [2, 52]. These methods require additional strong assumptions that cannot be tested from the data and raise concerns if these assumptions are violated [23, 37]. Compared to the OBS data, randomized controlled trial (RCT) data are considered as the gold standard for causal effect estimation [47]. However, practical challenges such as high costs and ethical concerns may make the collection of RCT data difficult [6, 68], resulting in limited sample sizes. Due to the small size, it is impractical to directly train causal effect prediction models on RCT data alone [29]. Therefore, it is necessary to find an effective method to combine small-scale RCT data with large-scale OBS data. Some studies use methods of correcting for residuals, in which RCT data are utilized to correct for biased estimates from OBS data [14]. However, these methods still rely on strong assumptions, including linear and additive data generation processes [35] and a shared structure between the two data types (RCT and OBS) [26]. This can lead to poor model performance in the complex real world. In contrast, other studies without these assumptions generally use both types of data to estimate CATE and the residuals between the biased and unbiased estimates [61], yet these methods suffer from the risk of overfitting due to insufficient RCT sample sizes.

In this paper, we introduce a two-stage framework named TSPF for CATE estimation with unmeasured confounding based on the pretraining-finetuning principle. Our approach leverages large-scale OBS data to train a foundational representation of covariates and then uses relatively small-scale RCT data to adjust the representation learned from OBS data. We then train a more accurate prediction model using this adjusted representation. In the second stage, we introduce an additional module that ensures stronger representation ability compared to the methods that use RCT data to estimate the residuals between the biased and unbiased estimates.

The contributions of this paper are summarized as follows.

- We present a two-stage pretraining-finetuning framework for estimating the CATE. This framework tackles the issue of unmeasured confounding by using a small amount of unconfounded RCT data to calibrate the representations learned from observational (OBS) data.
- The proposed framework does not rely on the linear and additive generation assumptions, and can flexibly adjust its model structure according to the sample size of RCT data, thus mitigating the over-fitting problem.
- Extensive experiments conducted on the IHDP and Jobs datasets demonstrate the effectiveness of our approach.

2 Related Work

2.1 CATE Estimation

CATE also known as heterogeneous treatment effect (HTE), refers to the average causal effects of a treatment on an outcome for subgroups with different covariates. To accurately estimate CATE, early statistical methods include matching [16], stratification [45], reweighting [5, 50], and tree-based methods [13, 58]. With the impressive advances in deep learning, recent CATE estimation methods can be broadly categorized into representational learning-based and generative model-based methods. Specifically, the core idea of representation learning-based methods is to learn a balanced covariate representation that has similar distributions in the treatment and control groups to mitigate confounding bias [3, 22, 66]. Such balanced representations can be obtained by using integral probability metric (IPM) for regularization [34, 54], local similarity preservation [64, 65], targeted learning [55, 70], and optimal transport [56, 59]. An alternative category is generative model-based methods that estimate the counterfactual outcomes by assuming their data generation process and exploiting generative models [22, 69, 74]. Specifically, CEVAE uses variational autoencoder (VAE) to learn confounders from observed variables [44]. SCIGAN generates missing counterfactual results based on generative adversarial networks (GAN) and combines the facts and estimated counterfactual results to estimate CATE [8]. Unlike previous representation learning-based methods, we do not require no unmeasured confounding assumption. Instead, we learn an augmented representation to address the effects of unmeasured confounding for correcting biased estimates.

2.2 Unmeasured Confounding

Unmeasured confounding refers to a situation where there are unmeasured variables in the study that influence both the treatment and the outcome, which may lead to bias in CATE estimates [1]. To address this problem, previous studies can be broadly categorized into two types. The first only uses OBS data and mainly includes sensitivity analysis and auxiliary information methods. Sensitivity analysis aims to quantify the potential impact of unmeasured confounding on the treatment effect and to obtain a bound on the treatment effect [49, 51]. However, these approaches assume a determined confounding mechanism of the unmeasured variables [31]. This assumption is too strong and can easily be violated in practice [18, 57]. Auxiliary information methods in causal inference mainly include instrumental variable (IV) methods and front-door adjustment methods [41]. IV methods rely on external variables to address unmeasured confounding in observational studies [32, 52, 60]. However, these methods assume linearity and require unconfounded instruments, posing practical limitations [19]. Front door adjustment methods estimate the causal effect of a treatment on an outcome by leveraging a causal pathway, also known as the front-door criterion, which blocks all possible influences of unmeasured confounders [7, 21]. In general, front door adjustment may rely on knowing the true causal graph, which may not always be feasible in practice [40, 53]. Another type of approaches combine both OBS and RCT data to mitigate unmeasured confounding. Kallus et al. [35] propose to use RCT data to correct the bias in CATE estimates derived from OBS data under the assumption that the unmeasured confounding bias is a linear function of the covariates [35]. Hatt et al. [25] extend to the non-linear case by using shared model structure between OBS and RCT data. Rather than using OBS and RCT data in separate regressions, by joint modeling the unmeasured confounding bias and the CATE with OBS and RCT data simultaneously as input, Wu et al. [61] identify both the bias from OBS estimates and unbiased CATE estimates using the R-learner approach, and Cheng et al. [11] perform a weighted average of the estimates from the two data sources to mitigate the bias. Both the determination of weights and the estimation of CATE require a large sample size. However, this requirement is at odds with the small sample size of RCT. In this paper, we adopt a more flexible network architecture, with a deep neural network for predicting biased outcomes using the OBS data and extra neural networks for adjusting the representation as well as the prediction using the RCT data, which allows us to utilize limited RCT data to achieve better performance.

3 Preliminaries

3.1 Problem Setup

We consider two independent data sources taken from the same target population: one from OBS and the other from RCT. Each individual in the OBS or RCT study is an observation of (X, Y, T, G), a random tuple with distribution P. For the i-th individual, the observation comprises d-dimensional covariates $X_i \in \mathcal{X} \subseteq \mathbb{R}^d$, the observed outcome Y_i, the assigned binary treatments $T_i \in \{0, 1\}$ ($T_i = 0$ for the controlled and $T_i = 1$ for the treated individuals) and G_i denoting participation in the OBS ($G_i = 0$) or RCT ($G_i = 1$) study. Using the Neyman-Rubin potential outcome framework [33], we let Y_i^1, Y_i^0 be the potential outcomes. We denote the OBS data as $\mathcal{D}^{OBS} = \{(X_i, T_i, Y_i, G_i = 0) : i \in \mathcal{O}\}$ with sample size n, and the RCT data as $\mathcal{D}^{RCT} = \{(X_i, T_i, Y_i, G_i = 1) : i \in \mathcal{R}\}$ with sample size m, where $\mathcal{O} = \{1, \dots, n\}$ and $\mathcal{R} = \{n+1, \dots, n+m\}$ are sample index sets for the OBS and RCT data, respectively. The total sample size is $N = n+m$. We define the propensity score as $e(x, G) = P(T = 1 \mid X = x, G)$. The CATE is defined as the conditional expectation of difference between potential outcomes under the treatment group and the control group as follows:

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

3.2 Identification of CATE

To identify the CATE from observed data, in addition to the Stable Unit Treatment Value Assumption (SUTVA) that there are no interference between units and there are no different forms of each treatment level, the following three assumptions are required:

Assumption 1. *(Ignorability)* $(Y^1, Y^0) \perp\!\!\!\perp T \mid X$,

Assumption 2. *(Consistency)* $Y = TY^1 + (1 - T)Y^0$,

Assumption 3. *(Positivity)* $0 < e(x, G) < 1$ *for all* $x \in \mathcal{X}$.

Assumption 1 is also known as no unmeasured confounding, which holds in the RCT by default due to the randomized treatment assignment. We can identify CATE based on RCT:

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

Compared to RCT data, unmeasured confounding may exist in OBS data. Unconfoundedness assumption is not assumed to hold for the observational data, i.e., $(Y^1, Y^0) \not\perp\!\!\!\perp T \mid (X, G = 0)$. We may not identify $\tau(x)$ only based on OBS data. Let us denote the difference in conditional average outcomes in the observational data by:

$$\omega(x) = \mathbb{E}\left[Y \mid T = 1, X = x, G = 0\right] - \mathbb{E}\left[Y \mid T = 0, X = x, G = 0\right].$$

Note that due to unmeasured confounding, $\omega(x) \neq \tau(x)$ for any x. The difference between these two quantities is precisely the confounding effect, which we denote the residual function as:

$$\eta(x) = \tau(x) - \omega(x).$$

In each treatment group, we denote the residual functions:

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

3.3 Previous Work

To address the unmeasured confounding, previous work proposes many methods to estimate the residual function $\eta(x)$ and obtain an accurate estimation of CATE $\tau(x)$ with extra small-scale RCT data.

3.3.1 Residual Correction. This type of approaches first only uses the OBS data for initial estimation and then uses RCT data for correcting the biased estimation with a residual function [35].

Step 1: Estimate $\omega(x)$ using the OBS data, denoted as $\hat{\omega}(x)$.

Step 2: Estimate $\eta(x)$ using the RCT data by minimizing:

$$\min_{\eta} \sum_{i \in \mathcal{R}} (Y_i^* - \hat{\omega}(X_i) - \eta(X_i))^2,$$

where $Y_i^* = \frac{T_i Y_i}{e(X_i, G_i)} - \frac{(1-T_i) Y_i}{1-e(X_i, G_i)}$ is the pseudo-outcome.

The final estimate of $\tau(x)$ is $\hat{\omega}(x) + \hat{\eta}(x)$. To ensure a consistent estimate of $\tau(x)$, the residual $\eta(x)$ is assumed to be linear and additive. Moreover, the small-scale RCT data used in the second phase may cause overfitting of the estimated residual $\hat{\eta}(x)$.

3.3.2 Joint Learning of Residual and CATE Models. This type of approaches involves training the models for estimating $\eta(x)$ and $\tau(x)$ simultaneously using both OBS and RCT data [61].

Step 1: Estimate the conditional average outcome $\mu(X, G)$ and propensity score $e(X, G)$ by using the combined OBS and RCT data, denoted as $\hat{\mu}(X, G)$ and $\hat{e}(X, G)$.

Step 2: Estimate $\{\eta(\cdot), \tau(\cdot)\}$ by the optimization problem:

$$\arg\min_{\eta,\tau} \sum_{i \in \mathcal{O} \cup \mathcal{R}} \{Y_i - \hat{\mu}\left(X_i, G_i\right) - [\tau(X_i) + (1 - G_i)\eta(X_i)] \cdot [T_i - \hat{e}(X_i, G_i)]\}^2 + \Lambda(\tau) + \Lambda(\eta),$$

where $\Lambda(\tau)$ and $\Lambda(\eta)$ are regularization terms on the complexity of the $\tau(\cdot)$ and $\eta(\cdot)$ functions.

Since RCT generally involves much smaller sample sizes compared to OBS studies, the data available for explicitly learning $\eta(x)$ is limited. Insufficient RCT data can lead to overfitting and a high variance of $\hat{\eta}(x)$, thus reducing the accuracy of CATE estimation.

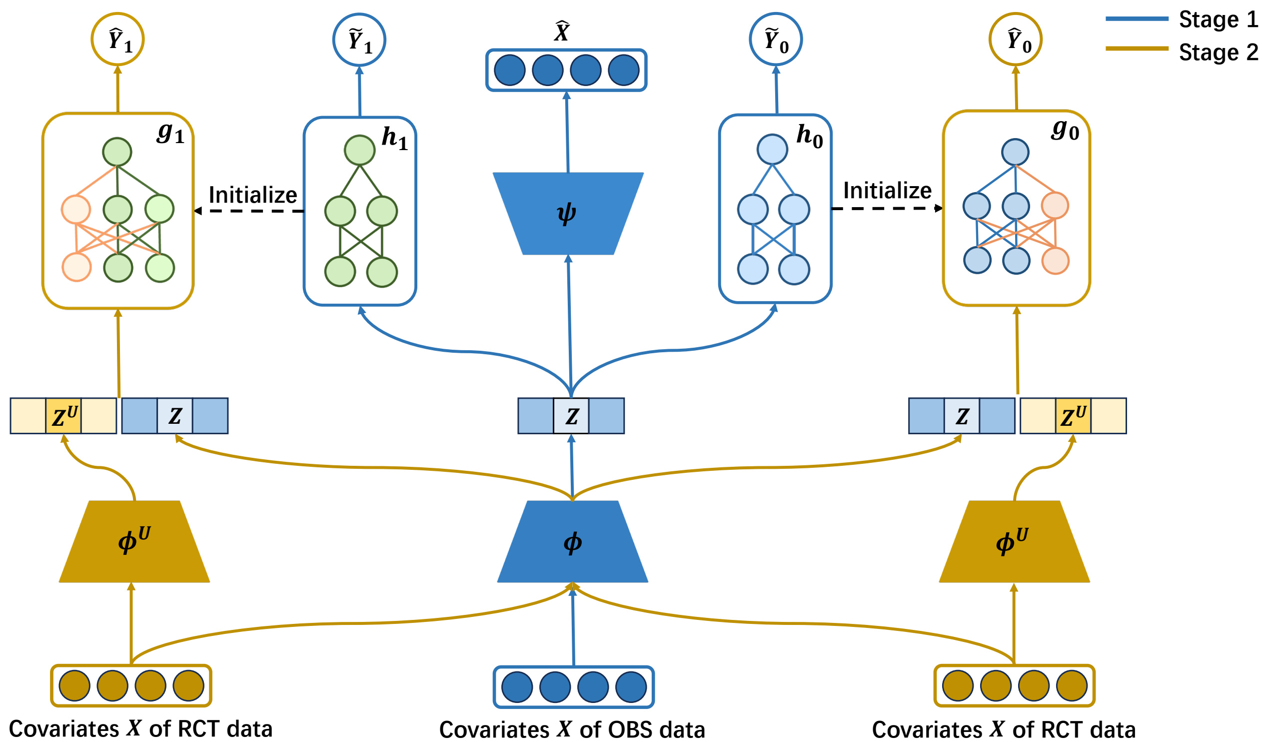

Figure 1: The framework of our proposed method, which is composed of the modules for the first stage (blue) and second stage (yellow). Note that the two ϕ^U shown in the figure represent the same module.

4 Methodology

In this section, we present a two-stage framework for CATE estimation based on the pretraining-finetuning principle, as shown in Figure 1. The motivation is to use large-scale OBS data to train a base representation of covariates, then use relatively small-scale unbiased RCT data to calibrate the bias in the representation learned from OBS data for training the unbiased prediction model. Specifically, in the first stage, only the OBS data is used. We start with a representation module ϕ, followed by one reconstruction module ψ and two prediction heads h_0 and h_1 to ensure the learned covariate representation can have enough information and predict the outcome for control group and treatment group simultaneously. While in the second stage, we use only the RCT data. The representation module ϕ learned in the first stage is frozen, with a learnable representation adapter module ϕ^U to calibrate the bias of ϕ. Then the concatenated representation is fed to two prediction heads g_0 and g_1 to obtain the unbiased predicted potential outcomes under control and treatment groups respectively. We carefully design an initialization strategy to ensure that the initialized second-stage model produces the same predictions as the converged model in the first stage. Our approach distinguishes from the one proposed by Kallus et al. [35], which only uses linear regression to estimate the residual function $\eta(x)$ in the second stage. We can regard the first stage as pretraining on the large OBS data and the second stage as finetuning on the small-scale unbiased RCT data.

4.1 First Stage: Pretraining Stage

The goal of our first-stage training is to obtain a representation module as well as prediction heads that can accurately estimate the potential outcomes of OBS data. These modules offer high-quality initialization for second-stage training, allowing fine-tuning on the RCT data to avoid the overfitting problem. A three-headed architecture and a multi-task training framework are employed to achieve this goal. Next, we will look into the details of each module.

4.1.1 Representation. We design a multi-layer feed-forward neural network ϕ to obtain a representation Z for the covariates X for both treatment and control groups. In other words, for an individual sample $(X_i, T_i, Y_i, G_i = 0)$, the representation $Z_i = \phi(X_i)$ remains the same whether $T_i = 0$ or $T_i = 1$. To better estimate the causal effect, we adopt the idea of covariate balancing on the representation space as Shalit et al. [54] Specifically, the representations for treatment group $\{Z_i = \phi(X_i) : G_i = 0, T_i = 1\}$ are regarded as i.i.d samples drawn randomly from a distribution $P_\phi^{t=1}$ and similarly $P_\phi^{t=0}$ for the control group. We anticipate the distributions of representations to be similar between the treatment and control groups. An integral probability metric (IPM) is employed to measure the distance between the two distributions. Thus the covariate imbalancing loss is defined as:

$$\mathcal{L}_{imb} = \text{IPM}_{\mathcal{G}}(\hat{P}_\phi^{t=0}, \hat{P}_\phi^{t=1}),$$

where $\mathrm{IPM}_{\mathcal{G}}(\cdot)$ is the empirical IPM defined by the function family $\mathcal{G}$, and $\hat{P}_{\phi}^{t=1}$ and $\hat{P}_{\phi}^{t=0}$ are empirical distributions of $P_{\phi}^{t=1}$ and $P_{\phi}^{t=0}$ respectively. In the implementation, we adopt Wasserstein distance as a showcase, which can be consistently estimated from finite samples within a mini-batch [20].

4.1.2 Reconstruction and Prediction. To ensure that Z retains as much information about the original covariates as possible, we introduce the decoder network ψ to reconstruct the original covariates: $\hat{X} = \psi(Z) = \psi(\phi(X))$. The reconstruction loss is computed by the mean squared error (MSE):

$$\mathcal{L}_{rec} = \frac{1}{|O|} \sum_{i \in O} ||\hat{X}_i - X_i||_2^2,$$

where $\hat{X}_i = \psi(\phi(X_i))$ is reconstructed covariate for the i-th sample in the OBS data. The reconstruction design resembling an autoencoder allows the learned representations to encompass nearly complete information in the covariates, rather than only the information necessary for fitting the training set, thereby enhancing the generalization of our representation module.

We then use the representations Z to estimate the potential outcomes with two l_p−layer prediction heads h_0 and h_1, which are the predictors for control and treatment outcomes, respectively. Note that unmeasured confounding in the observational data can lead to biased estimations of potential outcomes, we refer to the prediction result $\tilde{Y}^0 = h_0(Z) = h_0(\phi(X))$ as the pseudo control outcome and $\tilde{Y}^1 = h_1(Z) = h_1(\phi(X))$ as the pseudo treatment outcome. To enhance comparability between the treatment group and control group, we employ a reweighting technique to balance the two groups. Formally, let $f_h(x, t) = h_t(\phi(x))$ with $t \in \{0, 1\}$ be the predicted potential outcomes by via the two heads h_0 and h_1, the loss for outcome prediction is as follows:

$$\mathcal{L}_f = \frac{1}{|O|} \sum_{i \in O} w_i \cdot l(Y_i, f_h(X_i, T_i)),$$

with $w_i = \frac{T_i}{2u} + \frac{1-T_i}{2(1-u)}$, where $u = \frac{1}{n} \sum_{i=1}^{n} T_i$. The loss function $l(\cdot, \cdot)$ in $\mathcal{L}_f$ is flexible and can be determined based on the value range of potential outcomes. If the potential outcomes are binary, a cross-entropy loss is appropriate, whereas for continuous potential outcomes, an MSE loss is preferable.

In summary, in the first-stage training, we use the following training objective:

$$\min_{\phi, \psi, h_0, h_1} \mathcal{L}_f + \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{imb},$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are tunable hyperparameters.

4.2 Second Stage: Finetuning Stage

In the second stage of training, we exploit the small-scale unconfounded RCT data to remove the hidden confounding by concatenating the learned covariate representation in the first stage with a newly learned augmented covariate representation, then finetuning the prediction heads to obtain an unbiased CATE estimation. To achieve this, we keep the biased representation Z produced by ϕ unchanged but only treat it as a part of the representation, together with an additional Z^U generated by another representation module ϕ^U. We call ϕ^U a representation adapter, as it helps to adapt the final representation to account for the hidden confounding in the observed data. In addition, a large proportion of the parameters of the prediction heads g_0 and g_1 are initialized by h_0 and h_1, respectively. With the above steps, the second stage aims at adjusting the hidden confounding through the augmented covariate representation and the finetuned prediction heads with partial parameter initialization. Below we explain the modules in detail.

4.2.1 Representation Adapter. We employ a shallower feed-forward network ϕ^U as the representation adapter. It is worth noting that the width and depth of ϕ^U can be adjusted based on the scale of RCT data size. If the size of RCT data was comparable to that of OBS data, we can use the same architecture as ϕ. However, in real-world cases, the RCT data is rare compared to the OBS data, so the size of ϕ^U should be smaller. We denote the representation generated by ϕ^U as Z^U. To make sure that Z^U captures different features of covariates from Z, we employ mutual information to control the overlap between the two covariate representations:

$$\mathcal{L}_{MI} = CLUB(Z, Z^U),$$

$$\begin{aligned} CLUB(Z, Z^U) &= \frac{1}{m^2} \sum_{i \in \mathcal{R}} \sum_{j \in \mathcal{R}} \Big[\log q_\theta(Z_i^U \mid Z_i) - \log q_\theta(Z_j^U \mid Z_i) \Big] \\ &= \frac{1}{m} \sum_{i \in \mathcal{R}} \Big[\log q_\theta(Z_i^U \mid Z_i) - \frac{1}{m} \sum_{j \in \mathcal{R}} \log q_\theta(Z_j^U \mid Z_i) \Big], \end{aligned}$$

where $CLUB(Z, Z^U)$ is the empirical Contrastive Log-ratio Upper Bound (CLUB) of mutual information [12] between two covariate representations Z and Z^U, and $q_\theta(Z^U \mid Z)$ is the variational approximation of $P(Z^U \mid Z)$. With good variational approximation $q_\theta(Z^U \mid Z)$, it can be shown that the empirical CLUB is still a valid upper bound of the ground-truth mutual information.

4.2.2 Prediction. Similarly to the first stage, we design two l_p−layer prediction heads g_0 and g_1 to estimate the potential outcomes under control and treatment groups with unmeasured confoundings, respectively. Notice that g_0, g_1 and h_0, h_1 have the same depth l_p, yet every layer of g_t has a larger or equal width than h_t. For $t \in \{0, 1\}$, we define the header h_t as:

$$\begin{aligned} &a_{h_t}^{(0)} = Z,\ a_{h_t}^{(l)} = \sigma(W_{h_t}^{(l)} a_{h_t}^{(l-1)} + b_{h_t}^{(l)}),\ \text{for } l = 1, 2, \ldots, l_p - 1, \\ &a_{h_t}^{(l_p)} = \tilde{Y}_t = W_{h_t}^{(l_p)} a_{h_t}^{(l_p-1)} + b_{h_t}^{(l_p)}, \end{aligned}$$

where $W_{h_t}^{(l)}$ is the weight matrix from layer $l-1$ to layer l, $b_{h_t}^{(l)}$ is the bias vector of layer l, $a_{h_t}^{(l)}$ is the output of layer l for $l \in \{1, 2, \ldots, l_p\}$ and σ is the activation function. The definition of g_t is:

$$\begin{aligned} &a_{g_t}^{(0)} = \begin{bmatrix} Z \\ Z^U \end{bmatrix},\ a_{g_t}^{(l)} = \sigma(W_{g_t}^{(l)} a_{g_t}^{(l-1)} + b_{g_t}^{(l)}),\ \text{for } l = 1, 2, \ldots, l_p - 1, \\ &a_{g_t}^{(l_p)} = \hat{Y}_t = W_{g_t}^{(l_p)} a_{g_t}^{(l_p-1)} + b_{g_t}^{(l_p)}, \end{aligned}$$

with $W_{g_t}^{(l)}, b_{g_t}^{(l)}, a_{g_t}^{(l)}$ having similar meanings as $W_{h_t}^{(l)}, b_{h_t}^{(l)}, a_{h_t}^{(l)}$ but for g_t, $\hat{Y}_t$ is the final prediction for potential outcome under treatment t. In our design, every layer of g_t has a larger or equal width than h_t, thus the dimension of $W_{g_t}^{(l)}, b_{g_t}^{(l)}, a_{g_t}^{(l)}$ is no less than that

Algorithm 1: Learning algorithm of the TSPF framework.

Input: OBS data $\mathcal{D}^{OBS} = \{(X_i, T_i, Y_i, G_i = 0)\}_{i=1}^{n}$, RCT data $\mathcal{D}^{RCT} = \{(X_i, T_i, Y_i, G_i = 1)\}_{i=n+1}^{n+m}$ and four hyperparameters $\lambda_k > 0, k = 1, \ldots, 4$.

Compute $w_i = \frac{T_i}{2u} + \frac{1-T_i}{2(1-u)}$ with $u = \frac{1}{n}\sum_{i=1}^{n} T_i$ for $i = 1, ..., n$;

for *number of steps for training the first-stage model* **do**

  Sample a batch $\{(X_i, T_i, Y_i)\}_{i \in B}$ from $\mathcal{D}^{OBS}$;

  Update $\theta_1 = (\theta_\phi, \theta_\psi, \theta_{h_0}, \theta_{h_1})$ by descending along the gradient $\nabla_{\theta_1}(\mathcal{L}_f + \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{unb})$;

end

Initialize $({}^1W_{g_t}^{(l)}, {}^2W_{g_t}^{(l)}, {}^3W_{g_t}^{(l)}, {}^4W_{g_t}^{(l)}, {}^1b_{g_t}^{(l)}, {}^2b_{g_t}^{(l)})$ by $(W_{h_t}^{(l)}, 0, 0, 0, b_{h_t}^{(l)}, 0)$ for $l = 1, 2, \ldots, l_p$ and $t = 0, 1$;

Compute $w_i = \frac{T_i}{2u} + \frac{1-T_i}{2(1-u)}$ with $u = \frac{1}{m}\sum_{i=n+1}^{n+m} T_i$ for $i = n+1, ..., n+m$;

for *number of steps for training the second-stage model* **do**

  Sample a batch $\{(X_i, T_i, Y_i)\}_{i \in B}$ from $\mathcal{D}^{RCT}$;

  Update $\theta_2 = (\theta_{\phi^U}, \theta_{g_0}, \theta_{g_1})$ by descending along the gradient $\nabla_{\theta_2}(\mathcal{L}_{pred} + \lambda_3 \mathcal{L}_{MI} + \lambda_4 \mathcal{L}_{shift})$;

end

of $W_{h_t}^{(l)}, b_{h_t}^{(l)}, a_{h_t}^{(l)}$ respectively. We divide the parameters of g_t as:

$$W_{g_t}^{(l)} = \begin{bmatrix} {}^1W_{g_t}^{(l)}, {}^2W_{g_t}^{(l)} \\ {}^3W_{g_t}^{(l)}, {}^4W_{g_t}^{(l)} \end{bmatrix}, b_{g_t}^{(l)} = \begin{bmatrix} {}^1b_{g_t}^{(l)} \\ {}^2b_{g_t}^{(l)} \end{bmatrix},$$

where ${}^1W_{g_t}^{(l)}, {}^1b_{g_t}^{(l)}$ have the same shapes as $W_{h_t}^{(l)}, b_{h_t}^{(l)}$ respectively, for $l = 1, 2, \ldots, l_p$, with the detailed initialization strategy as follows.

Initialization. The initialization of the model parameters is crucial for the preservation of covariate information from the first stage as well as the effectiveness of the finetuning stage. The goal of initialization is to make sure the model initially produces the same prediction as the trained first-stage model. Nonetheless, a challenge of parameter initialization is that the model architecture of the second stage differs from that of the first stage, because of the augmented covariate representation. Based on the division of the parameters, we propose the following initialization strategy:

$$({}^1W_{g_t}^{(l)}, {}^2W_{g_t}^{(l)}, {}^3W_{g_t}^{(l)}, {}^4W_{g_t}^{(l)}, {}^1b_{g_t}^{(l)}, {}^2b_{g_t}^{(l)}) \leftarrow (W_{h_t}^{(l)}, 0, 0, 0, b_{h_t}^{(l)}, 0)$$

for $l = 1, 2, \ldots, l_p$. That is, the shared parameters between the prediction heads g_t and h_t are initialized to be the same, and the rest parameters of the prediction head g_t are initialized to be zero.

As in the first stage, we denote $f_g(x, t) = g_t([\phi(x)^\top | \phi^U(x)^\top]^\top)$ for $t \in \{0, 1\}$, where $[\phi(x)^\top | \phi^U(x)^\top]$ is the concatenated covariate representation. Given the RCT data $\{(X_i, T_i, Y_i, G_i = 1) : i \in \mathcal{A}\}$, the prediction loss is computed as:

$$\mathcal{L}_{pred} = \frac{1}{|\mathcal{R}|}\sum_{i \in \mathcal{R}} w_i \cdot l(Y_i, f_g(X_i, T_i)),$$

similarly with $w_i = \frac{T_i}{2u} + \frac{1-T_i}{2(1-u)}$ and $u = \frac{1}{m}\sum_{i=n+1}^{n+m} T_i$.

Regularization. The presence of unmeasured confounding may cause a slight shift in the distribution of OBS data from the RCT data. Therefore the second-stage fine-tuned model should not deviate significantly from the first-stage model. We denote the initial value of θ_{g_t} as $\theta_{g_t}^0$. In order to constrain the deviation from the initial value, we include an l_2−norm in the loss function:

$$\mathcal{L}_{shift} = ||\theta_{g_0} - \theta_{g_0}^0||_2^2 + ||\theta_{g_1} - \theta_{g_1}^0||_2^2.$$

Overall, the training objective of the second stage is given by:

$$\min_{\phi^U,\, g_0,\, g_1} \mathcal{L}_{pred} + \lambda_3 \mathcal{L}_{MI} + \lambda_4 \mathcal{L}_{shift},$$

where $\lambda_3 > 0$ and $\lambda_4 > 0$ are tunable hyperparameters. Note that during the second-phase training, we froze the parameters of the representation module ϕ and the decoder network ψ, while train the representation adapter module ϕ^U and the two prediction heads g_0, g_1. We summarize the whole learning algorithm in Alg. 1.

Compared to residual correction methods as in Kallus et al. [35], our representation adapter module guarantees a stronger representation ability, relaxing the linearly additive assumption. While compared to methods that jointly learn residual and CATE, one advantage is that when RCT data are limited, our proposed partial initialization strategy in the TSPF framework can avoid overfitting.

5 Experiment

5.1 Datasets

Following previous studies [44, 54, 67], we conduct experiments on two publicly available datasets, namely **IHDP** [28] and **Jobs** [54]. The **IHDP** is a semi-synthetic dataset for causal effect estimation. The dataset is based on the Infant Health and Development Program, where the covariates are obtained by a randomized experiment investigating the effect of home visits by specialists on future cognitive scores. It consists of 747 units (19% treated, 81% control) and 25 covariates measuring the children and their mothers. The **Jobs** is a common benchmark dataset developed by LaLonde in 1986, studying the change of income and employment status after job training. We use an extended version of **Jobs** that comprises about 3,000 units (10% treated, 90% control) with 17 covariates.

5.2 Data Preprocessing

For both **IHDP** and **Jobs**, we simulate unmeasured confounding by generating a c-dimensional confounder $U_i \in \mathbb{R}^c$. To make sure the U_i has a non-zero effect on Y_i and T_i, we generate the data below:

$$\begin{aligned} &W_1 \sim \mathcal{N}(0, 0.1)^d,\ W_2 \sim \mathcal{N}(0.02, 0.1)^c,\ W_3 \sim \mathcal{N}(0.1, 1)^d, \\ &W_4 \sim \mathcal{N}(0.1, 1)^c,\ W_5 \sim \mathcal{U}(0, 0.2)^d,\ W_6 \sim \mathcal{U}(0, 0.2)^c, \\ &U_i \sim \mathcal{U}(0, 0.2)^c,\ T_i \sim \text{Bern}(\sigma(W_1 \cdot X_i + W_2 \cdot U_i)), \\ &\mu_i^0 = W_3 \cdot X_i + W_4 \cdot U_i,\ \mu_i^1 = W_5 \cdot X_i + W_6 \cdot U_i + 4, \\ &Y_i^0 \sim \mathcal{N}(\mu_i^0, 0.1),\ Y_i^1 \sim \mathcal{N}(\mu_i^1, 0.1), \end{aligned}$$

where $\mathcal{N}(\mu, D)$ denotes the normal distribution with mean μ and variance D, $\mathcal{U}(a, b)$ is the uniform distribution on interval (a, b), Bern(p) means the Bernoulli distribution with probability p, $\sigma(x) = 1/(1+\exp(-x))$ is the sigmoid function. Note that we keep $E[Y_i^1] = E[Y_i^0] + 4$ as the same as the **IHDP** dataset and we let $E[W_i] >$

Table 1: The experiment results on the IHDP dataset and Jobs dataset. The best result is bolded and the second best is underlined.

	IHDP				Jobs			
	In-sample		Out-sample		In-sample		Out-sample	
Methods	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}
T-learner	0.44 ± 0.03	0.04 ± 0.02	0.52 ± 0.05	**0.02 ± 0.01**	0.66 ± 0.27	0.02 ± 0.02	0.60 ± 0.20	0.02 ± 0.01
S-learner	0.98 ± 0.18	0.04 ± 0.03	1.37 ± 0.34	0.15 ± 0.10	0.70 ± 0.30	0.02 ± 0.02	0.67 ± 0.39	0.03 ± 0.02
DR-learner	0.71 ± 0.19	0.06 ± 0.04	0.81 ± 0.25	0.06 ± 0.03	0.50 ± 0.07	0.07 ± 0.02	0.48 ± 0.09	0.07 ± 0.03
Causal Forest	1.90 ± 0.29	0.07 ± 0.05	2.04 ± 0.42	0.19 ± 0.11	1.38 ± 0.40	0.13 ± 0.08	1.23 ± 0.35	0.14 ± 0.08
TARNet	0.42 ± 0.09	0.05 ± 0.04	0.44 ± 0.12	0.05 ± 0.05	0.19 ± 0.17	0.02 ± 0.01	0.13 ± 0.02	0.02 ± 0.01
CEVAE	2.89 ± 0.72	0.15 ± 0.12	2.87 ± 0.80	0.22 ± 0.16	2.86 ± 0.80	0.35 ± 0.24	2.82 ± 0.74	0.46 ± 0.38
SCIGAN	2.53 ± 0.47	0.63 ± 0.29	2.58 ± 0.57	0.55 ± 0.48	2.28 ± 0.75	0.56 ± 0.15	2.15 ± 0.80	0.47 ± 0.21
DragonNet	0.19 ± 0.04	0.03 ± 0.02	0.26 ± 0.08	0.04 ± 0.02	0.15 ± 0.11	**0.01 ± 0.01**	0.11 ± 0.03	**0.01 ± 0.01**
DESCN	0.28 ± 0.06	0.05 ± 0.05	0.41 ± 0.11	0.07 ± 0.06	0.44 ± 0.09	0.28 ± 0.13	0.44 ± 0.08	0.27 ± 0.13
DRCFR	0.74 ± 0.32	0.15 ± 0.10	0.90 ± 0.52	0.18 ± 0.17	0.91 ± 0.49	0.08 ± 0.08	0.71 ± 0.43	0.09 ± 0.07
Twostep linear	0.64 ± 0.13	0.31 ± 0.23	0.87 ± 0.18	0.56 ± 0.24	0.69 ± 0.35	0.20 ± 0.12	0.56 ± 0.27	0.25 ± 0.16
CorNet	0.34 ± 0.12	0.05 ± 0.03	0.32 ± 0.09	0.04 ± 0.04	0.21 ± 0.09	0.04 ± 0.04	0.22 ± 0.08	0.05 ± 0.03
TSFP (ours)	**0.13 ± 0.02**	**0.02 ± 0.02**	**0.16 ± 0.04**	0.04 ± 0.02	**0.09 ± 0.03**	**0.01 ± 0.01**	**0.06 ± 0.01**	**0.01 ± 0.01**

0 for $i \in \{2, 4, 6\}$ to ensure the non-zero effect of U_i. The unmeasured confounding strength parameter c is set to 30. Then we slice the training, validation, and test sets in the ratio of 63/27/10. In addition, to obtain a separate RCT training dataset for data fusion, we first randomly split $r\%$ of the training samples, and then assign treatments T_i^{new} according to the following formula and replace the factual treatment T_i and outcome Y_i^f to obtain a RCT dataset:

$$T_i^{new} = \text{Bern}(0.5),\ \ Y_i^{new} = \mathbb{I}(T_i^{new} = T_i)(Y_i^f - Y_i^{cf}) + Y_i^{cf},$$

where $\mathbb{I}(\cdot)$ is the indicator function, $Y_i^f = T_i Y_i^1 + (1 - T_i)Y_i^0$ is the factual outcome, and $Y_i^{cf} = T_i Y_i^0 + (1 - T_i)Y_i^1$ is the counterfactual outcome. The RCT data ratio $r\%$ is set to 10% unless otherwise stated. Finally, we replace the T_i and Y_i using the above formula for all samples in the validation set.

5.3 Baselines and Evaluation Metrics

5.3.1 Baselines.

- T-learner [38]: T-learner utilizes two separate regressors for each treatment group.
- S-learner [4]: S-learner treats the indicator of treatment as features, utilizing a single model to estimate the potential outcome for both treatment and control groups.
- DR-learner [36]: DR-learner estimates the CATE via cross-fitting a doubly robust score function in two stages.
- SCIGAN [67]: SCIGAN utilizes a generative adversarial network to model treatment effect.
- Causal Forest [58]: Causal Forest is a random forest-based model that directly estimates the treatment effect.
- TARNet [54]: TARNet applies a shared representation layer and a two-head network inference layer.
- DragonNet [55]: DragonNet designs an adaptive neural network to learn propensities and counterfactual outcomes.
- DESCN [71]: DESCN uses deep networks to model treatment effects in the entire sample space.
- DRCFR [24]: DRCFR aims to learn disentangled representations and address selection bias in CATE estimation.
- Twostep linear [35]: Twostep linear method uses OBS data to learn a biased estimate for the treatment effect and then aims to remove the bias using RCT data.
- CorNet [26]: CorNet uses the RCT data to learn a non-linear bias function in the second stage.

5.3.2 Evaluation Metrics. Following previous studies [54, 64], we evaluate the performance of CATE estimation using the *square root of Precision in Estimation of Heterogeneous Effects* (PEHE):

$$\sqrt{\epsilon_{\text{PEHE}}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}((\hat{Y}_i^1 - \hat{Y}_i^0) - (Y_i^1 - Y_i^0))^2},$$

where $\hat{Y}_i^t$ and Y_i^t are the predicted and ground truth values for the potential outcomes of individual i under treatment t. In addition, we also use the *absolute error in Average Treatment Effect* (ATE) to evaluate estimation performance, which is defined as:

$$\epsilon_{\text{ATE}} = \frac{1}{n}\left|\sum_{i=1}^{n}((\hat{Y}_i^1 - \hat{Y}_i^0) - (Y_i^1 - Y_i^0))\right|.$$

The smaller the evaluation metrics, the better the estimation.

5.3.3 Experimental Details. We implement TSPF with a multi-layer perceptron [27] with 2 layers for our representation and reconstruction modules as well as the prediction heads in both stages using the PyTorch framework. We tune the hyper-parameters in the loss functions from $1e-5$ to 0.1, the learning rate from $1e-5$ to $1e-2$, and the weight decay parameter from $1e-5$ to $1e-2$. The batch size is set to 256 in all scenarios and the optimizer is selected in Adam and SGD. We report both in-sample and out-of-sample results for metrics $\sqrt{\epsilon_{\text{PEHE}}}$ and ϵ_{ATE} in our experiments.

Table 2: Ablation study on the pretraining stage and fine-tuning stage, the initialization strategy, and the training strategy of freezing foundational representation module and the initialized parameters.

	IHDP				Jobs			
	In-sample		Out-sample		In-sample		Out-sample	
	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}
TSFP w/o $\mathcal{L}_{rec}$ and $\mathcal{L}_{imb}$	0.25 ± 0.05	0.07 ± 0.02	0.29 ± 0.08	0.07 ± 0.02	0.19 ± 0.03	0.04 ± 0.03	0.18 ± 0.02	0.04 ± 0.03
TSFP w/o $\mathcal{L}_{rec}$	0.21 ± 0.04	0.05 ± 0.02	0.24 ± 0.05	0.05 ± 0.02	0.14 ± 0.01	0.03 ± 0.01	0.13 ± 0.02	0.03 ± 0.01
TSFP w/o $\mathcal{L}_{imb}$	0.18 ± 0.03	0.03 ± 0.02	0.19 ± 0.05	0.05 ± 0.02	0.12 ± 0.02	0.03 ± 0.02	0.10 ± 0.04	0.03 ± 0.01
TSFP w/o $\mathcal{L}_{MI}$ and $\mathcal{L}_{shift}$	0.23 ± 0.03	0.06 ± 0.01	0.26 ± 0.05	0.07 ± 0.02	0.16 ± 0.04	0.03 ± 0.01	0.18 ± 0.04	0.03 ± 0.01
TSFP w/o $\mathcal{L}_{MI}$	0.17 ± 0.04	0.04 ± 0.01	0.18 ± 0.04	0.05 ± 0.02	0.12 ± 0.03	0.02 ± 0.01	0.15 ± 0.03	0.02 ± 0.01
TSFP w/o $\mathcal{L}_{shift}$	0.19 ± 0.04	0.05 ± 0.01	0.21 ± 0.05	0.05 ± 0.02	0.15 ± 0.05	0.03 ± 0.01	0.16 ± 0.04	0.02 ± 0.01
TSFP not freezing representation	0.17 ± 0.02	0.03 ± 0.03	0.20 ± 0.05	0.05 ± 0.04	0.10 ± 0.03	0.01 ± 0.01	0.09 ± 0.03	0.01 ± 0.01
TSFP freezing initialized parameters	0.16 ± 0.08	0.03 ± 0.02	0.19 ± 0.05	0.04 ± 0.02	0.09 ± 0.03	0.01 ± 0.01	0.08 ± 0.04	0.01 ± 0.01
TSFP w/o initialization	0.35 ± 0.09	0.08 ± 0.07	0.38 ± 0.11	0.07 ± 0.06	0.42 ± 0.12	0.06 ± 0.04	0.41 ± 0.13	0.06 ± 0.04
TSFP w/o initialization and $\mathcal{L}_{shift}$	0.43 ± 0.09	0.11 ± 0.07	0.45 ± 0.10	0.12 ± 0.06	0.54 ± 0.13	0.09 ± 0.05	0.53 ± 0.14	0.11 ± 0.05
TSFP w/o Z^U	0.18 ± 0.03	0.05 ± 0.02	0.19 ± 0.04	0.05 ± 0.02	0.14 ± 0.03	0.02 ± 0.01	0.16 ± 0.04	0.02 ± 0.01
TSFP	0.13 ± 0.02	0.02 ± 0.02	0.16 ± 0.04	0.04 ± 0.02	0.09 ± 0.03	0.01 ± 0.01	0.06 ± 0.01	0.01 ± 0.01

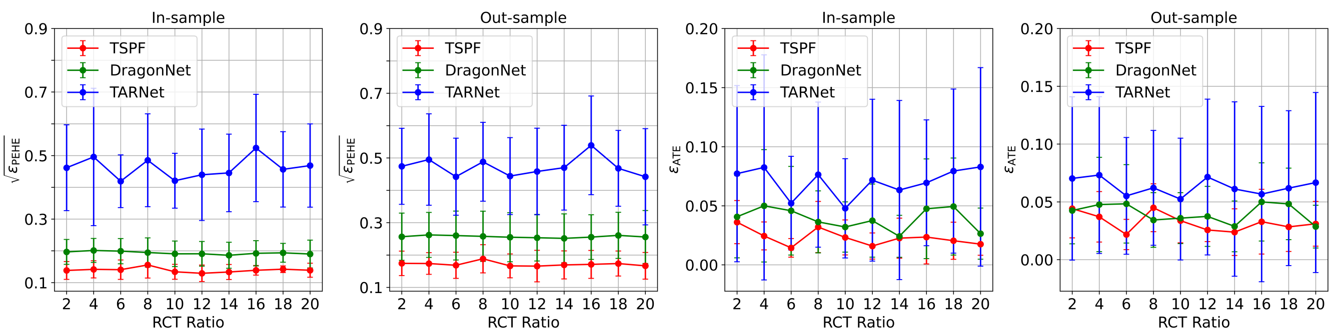

Figure 2: The experiment results for the IHDP dataset under different RCT data ratios in the training data.

5.4 Performance Analysis

Table 1 shows the prediction performance with varying baselines and our methods. First, representation-based methods generally outperform generation-based methods and meta-learners, which shows the effectiveness of causal representation learning. Note that our proposed TSPF exhibits the most competitive performance in most cases, outperforming all the baselines except for the ϵ_{ATE} metric in out-sample scenario. In the comparison of two-step methods, we can see that CorNet [26], which adopts a nonlinear residual module, outperforms the Twostep linear method based on the assumption that the residual $\eta(X_i)$ is linear and additive. In addition, our method stably outperforms the other two-stage methods, showing the effectiveness of our pretraining-finetuning framework.

5.5 Ablation Study

After evaluating the overall performance of our method, we perform ablation studies to validate the effectiveness of each module in our approach. Firstly, we consider employing all the modules in the first stage and taking away the mutual information term and/or the shift regularization term. Then we include all the modules in the second stage and test the effect of removing the reconstruction module and/or the IPM distance. To validate the soundness of our training algorithm, especially the initialization strategy, we also evaluate the performance of TSPF when not freezing the foundational representation module ϕ, freezing the initialized parameters in g_0 and g_1, without the initialization strategy, and without the augmented representation Z^U. The results are presented in Table 2, which shows that each component in our framework plays an important role in accurately estimating CATE. In the first/second stage, $\mathcal{L}_{rec}/\mathcal{L}_{shift}$ is more important than $\mathcal{L}_{imb}/\mathcal{L}_{MI}$. In addition, our initialization strategy significantly improves the performance.

5.6 In-Depth Analysis

5.6.1 RCT Data Ratio. To validate whether our framework performs well with limited RCT data, we further explore the effect of using different ratios of RCT data in the training set. We conduct experiments with a sequence of increasing ratios from 2% to 20%. For each scenario, the optimal network size and hyper-parameters are finetuned to ensure fair comparison. The results in Figure 2 show that our method is robust to the change of RCT data ratio and outperforms the baselines under different RCT ratios.

5.6.2 Unmeasured Confounding Strength. Our approach aims to achieve unbiased estimation of CATE in the presence of unmeasured confounding. To verify the validity under different strengths of

Table 3: Experimental results on the IHDP dataset with unmeasured confounding strength $c = 10$ and $c = 50$. The best result is bolded and the second best is underlined.

	unmeasured confounding strength $c = 10$				unmeasured confounding strength $c = 50$			
	In-sample		Out-sample		In-sample		Out-sample	
Methods	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}
T-learner	0.42 ± 0.06	0.03 ± 0.02	0.47 ± 0.08	0.06 ± 0.05	0.48 ± 0.05	0.05 ± 0.02	0.53 ± 0.06	0.04 ± 0.03
S-learner	0.99 ± 0.26	0.05 ± 0.04	1.29 ± 0.34	0.13 ± 0.11	1.17 ± 0.11	0.05 ± 0.03	1.51 ± 0.19	0.11 ± 0.10
DR-learner	0.69 ± 0.12	0.06 ± 0.03	0.72 ± 0.13	0.10 ± 0.05	0.73 ± 0.11	0.07 ± 0.04	0.80 ± 0.13	0.08 ± 0.06
Causal Forest	1.93 ± 0.31	0.08 ± 0.07	1.85 ± 0.28	0.24 ± 0.22	1.98 ± 0.18	0.13 ± 0.06	2.04 ± 0.34	0.29 ± 0.19
TARNet	0.26 ± 0.06	0.03 ± 0.02	0.27 ± 0.13	0.04 ± 0.04	0.52 ± 0.20	0.09 ± 0.09	0.53 ± 0.24	0.11 ± 0.11
CEVAE	3.09 ± 0.92	0.19 ± 0.16	3.19 ± 1.22	0.41 ± 0.33	3.50 ± 0.51	0.25 ± 0.16	3.39 ± 0.45	0.65 ± 0.32
SCIGAN	2.74 ± 0.68	0.62 ± 0.28	2.77 ± 0.89	0.56 ± 0.30	2.94 ± 0.42	0.46 ± 0.25	2.90 ± 0.35	0.75 ± 0.38
DragonNet	0.19 ± 0.04	0.03 ± 0.02	0.23 ± 0.05	0.06 ± 0.07	0.23 ± 0.05	0.06 ± 0.07	0.29 ± 0.13	0.07 ± 0.04
DESCN	0.36 ± 0.17	0.04 ± 0.03	0.42 ± 0.26	0.08 ± 0.09	0.46 ± 0.14	0.09 ± 0.05	0.55 ± 0.23	0.08 ± 0.06
DRCFR	0.70 ± 0.32	0.09 ± 0.08	0.92 ± 0.74	0.18 ± 0.22	1.21 ± 0.80	0.30 ± 0.20	1.17 ± 0.72	0.26 ± 0.18
Twostep linear	0.70 ± 0.16	0.29 ± 0.25	0.83 ± 0.30	0.30 ± 0.39	0.76 ± 0.21	0.42 ± 0.23	0.83 ± 0.25	0.33 ± 0.31
CorNet	0.23 ± 0.12	0.05 ± 0.04	0.27 ± 0.13	0.04 ± 0.04	0.39 ± 0.19	0.08 ± 0.08	0.38 ± 0.18	0.07 ± 0.07
TSFP (ours)	**0.15 ± 0.04**	**0.02 ± 0.02**	**0.15 ± 0.04**	**0.02 ± 0.02**	**0.15 ± 0.03**	**0.02 ± 0.01**	**0.17 ± 0.07**	**0.02 ± 0.02**

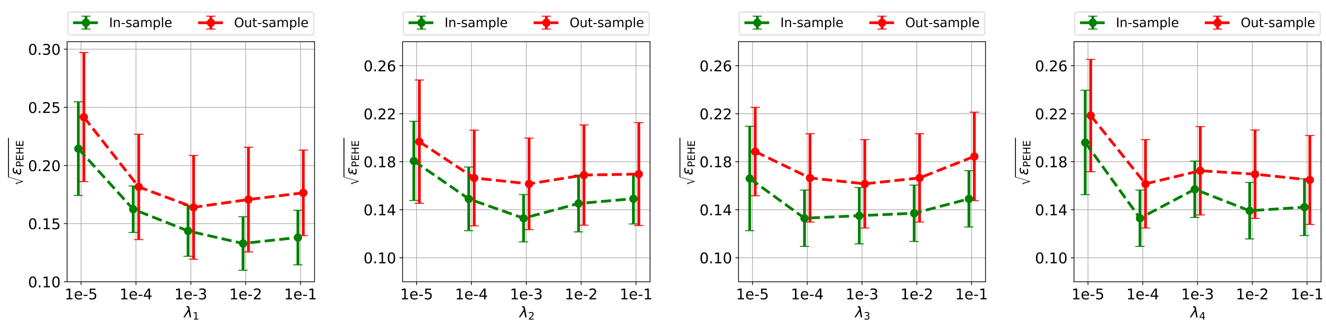

Figure 3: Sensitivity analysis on the reconstruction loss strength λ_1, covariate imbalancing loss strength λ_2, mutual information loss strength λ_3 and RCT shift loss strength λ_4 on the IHDP dataset.

unmeasured confounding, we add a comparison of the performance under low ($c = 10$) and high ($c = 50$) unmeasured confounding effect, which is shown in Table 3. The results show that our method stably outperforms baselines with varying c.

5.7 Sensitivity Analysis

Despite there are many hyper-parameters in our TSPF framework, they present in two separate stages, i.e, λ_1 and λ_2 in the first stage, while λ_3 and λ_4 in the second stage. As a result, the hyper-parameter search space is significantly reduced, compared with simultaneously searching four hyper-parameters. We explore the sensitivity of the hyper-parameters on the IHDP dataset and the results are shown in Figure 3. For all four hyper-parameters, the best result is achieved with moderate values. The performance dramatically drops with low values of hyper-parameters, which further shows the effectiveness of each component in our method.

6 Conclusion

This paper studies the CATE estimation in the presence of unmeasured confounding fusing both large-scale OBS data and small-scale RCT data. We propose a two-stage pretraining-finetuning framework to tackle the overfitting problem caused by the small-scale RCT data. Specifically, the foundational representation learned in the first stage is used to adjust for the *measured* confounding in the OBS data. The augmented representation learned in the second stage is used to adjust for the *unmeasured* confounding guided by the RCT data. To avoid overfitting caused by the small-scale RCT data in the second stage, instead of training a separate network, we further propose to partially initialize the network parameters from the pretrained network from the first stage. Compared to the previous CATE estimation methods that combine OBS and RCT data, our approach has the advantage of not restricting the data-generating process (e.g., linearity or additive noise assumptions) and not suffering from overfitting. The extensive semi-synthetic and real-world experiments conducted on two widely-used public datasets demonstrate the superiority of our method.

Acknowledgments

HL was supported by National Natural Science Foundation of China (623B2002). MG was supported by ARC DE210101624 and ARC DP240102088.

References

[1] Cande V Ananth and Enrique F Schisterman. 2018. Hidden biases in observational epidemiology: the case of unmeasured confounding. *BJOG: An International Journal of Obstetrics and Gynaecology* 125, 6 (2018), 644.

[2] Joshua D Angrist, Guido W Imbens, and Donald B Rubin. 1996. Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* 91, 434 (1996), 444–455.

[3] Serge Assaad, Shuxi Zeng, Chenyang Tao, Shounak Datta, Nikhil Mehta, Ricardo Henao, Fan Li, and Lawrence Carin. 2021. Counterfactual representation learning with balancing weights. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1972–1980.

[4] Susan Athey and Guido W Imbens. 2015. Machine learning methods for estimating heterogeneous causal effects. *Stat* 1050, 5 (2015), 1–26.

[5] Heejung Bang and James M Robins. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics* 61, 4 (2005), 962–973.

[6] Florent Bédécarrats, Isabelle Guérin, and François Roubaud. 2020. *Randomized control trials in the field of development: A critical perspective*. Oxford University Press.

[7] Marc F Bellemare, Jeffrey R Bloem, and Noah Wexler. 2020. The Paper of How: Estimating Treatment Effects Using the Front-Door Criterion. *Oxford Bulletin of Economics and Statistics* (2020).

[8] Ioana Bica, James Jordon, and Mihaela van der Schaar. 2020. Estimating the effects of continuous-valued interventions using generative adversarial networks. *Advances in Neural Information Processing Systems* 33 (2020), 16434–16445.

[9] Diogo GC Britto, Paolo Pinotti, and Breno Sampaio. 2022. The effect of job loss and unemployment insurance on crime in Brazil. *Econometrica* 90, 4 (2022), 1393–1423.

[10] Marie-Laure Charpignon, Bella Vakulenko-Lagun, Bang Zheng, Colin Magdamo, Bowen Su, Kyle Evans, Steve Rodriguez, Artem Sokolov, Sarah Boswell, Yi-Han Sheu, et al. 2022. Causal inference in medical records and complementary systems pharmacology for metformin drug repurposing towards dementia. *Nature Communications* 13, 1 (2022), 7652.

[11] David Cheng and Tianxi Cai. 2021. Adaptive combination of randomized and observational data. *arXiv:2111.15012* (2021).

[12] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. 2020. Club: A contrastive log-ratio upper bound of mutual information. In *International Conference on Machine Learning*. PMLR, 1779–1788.

[13] Hugh A Chipman, Edward I George, and Robert E McCulloch. 2010. BART: Bayesian additive regression trees. (2010).

[14] Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. 2024. Causal inference methods for combining randomized trials and observational studies: a review. *Statist. Sci.* 39, 1 (2024), 165–191.

[15] Giovanni Corrao, Federica Nicotra, Andrea Parodi, Antonella Zambon, Davide Soranna, Franca Heiman, Luca Merlino, and Giuseppe Mancia. 2012. External adjustment for unmeasured confounders improved drug–outcome association estimates based on health care utilization data. *Journal of Clinical Epidemiology* 65, 11 (2012), 1190–1199.

[16] Rajeev H Dehejia and Sadek Wahba. 2002. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics* 84, 1 (2002), 151–161.

[17] Jacob Dorn, Kevin Guo, and Nathan Kallus. 2024. Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding. *J. Amer. Statist. Assoc.* (2024), 1–12.

[18] AlexanderM Franks, Alexander D'Amour, and Avi Feller. 2019. Flexible sensitivity analysis for observational studies without observable implications. *J. Amer. Statist. Assoc.* (2019).

[19] Dennis Frauen and Stefan Feuerriegel. 2022. Estimating individual treatment effects under unobserved confounding using binary instruments. *arXiv:2208.08544* (2022).

[20] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. 2015. Learning with a Wasserstein loss. *Advances in Neural Information Processing Systems* 28 (2015).

[21] Isabel R Fulcher, Ilya Shpitser, Stella Marealle, and Eric J Tchetgen Tchetgen. 2020. Robust inference on population indirect causal effects: the generalized front door criterion. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82, 1 (2020), 199–214.

[22] Ruocheng Guo, Jundong Li, and Huan Liu. 2020. Learning individual causal effects from networked observational data. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 232–240.

[23] Fernando Pires Hartwig, Linbo Wang, George Davey Smith, and Neil Martin Davies. 2023. Average causal effect estimation via instrumental variables: the no simultaneous heterogeneity assumption. *Epidemiology* 34, 3 (2023), 325–332.

[24] Negar Hassanpour and Russell Greiner. 2019. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*.

[25] Tobias Hatt, Jeroen Berrevoets, Alicia Curth, Stefan Feuerriegel, and Mihaela van der Schaar. 2022. Combining observational and randomized data for estimating heterogeneous treatment effects. *arXiv:2202.12891* (2022).

[26] Tobias Hatt, Daniel Tschernutter, and Stefan Feuerriegel. 2022. Generalizing off-policy learning under sample selection bias. In *Uncertainty in Artificial Intelligence*. PMLR, 769–779.

[27] Simon Haykin. 1994. *Neural Networks: A Comprehensive Foundation*. Macmillan.

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

[29] Jeroen Hoogland, Joanna IntHout, Michail Belias, Maroeska M Rovers, Richard D Riley, Frank E. Harrell Jr, Karel GM Moons, Thomas PA Debray, and Johannes B Reitsma. 2021. A tutorial on individualized treatment effect prediction from randomized trials with a binary endpoint. *Statistics in Medicine* 40, 26 (2021), 5961–5981.

[30] Van-Nam Huynh, Vladik Kreinovich, and Songsak Sriboonchitta. 2016. *Causal inference in econometrics*. Springer.

[31] Guido W Imbens. 2003. Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review* 93, 2 (2003), 126–132.

[32] Guido W Imbens. 2014. *Instrumental variables: An econometrician's perspective*. Technical Report. National Bureau of Economic Research.

[33] Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

[34] Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *International Conference on Machine Learning*. PMLR, 3020–3029.

[35] Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. 2018. Removing hidden confounding by experimental grounding. *Advances in Neural Information Processing Systems* 31 (2018).

[36] Edward H. Kennedy. 2023. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics* 17, 2 (2023), 3008 – 3049. https://doi.org/10.1214/23-EJS2157

[37] Lingjing Kong, Guangyi Chen, Petar Stojanov, Haoxuan Li, Eric P Xing, and Kun Zhang. 2025. Towards Understanding Extrapolation: a Causal Lens. *Advances in Neural Information Processing Systems* 37 (2025).

[38] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences* 116, 10 (2019), 4156–4165.

[39] Baohong Li, Haoxuan Li, Ruoxuan Xiong, Anpeng Wu, Fei Wu, and Kun Kuang. 2024. Learning Shadow Variable Representation for Treatment Effect Estimation under Collider Bias. In *International Conference on Machine Learning*. PMLR.

[40] Haoxuan Li, Yue Liu, Zhi Geng, and Kun Zhang. 2024. A Local Method for Satisfying Interventional Fairness with Partially Known Causal Graphs. *Advances in Neural Information Processing Systems* (2024).

[41] Haoxuan Li, Kunhan Wu, Chunyuan Zheng, Yanghao Xiao, Hao Wang, Zhi Geng, Fuli Feng, Xiangnan He, and Peng Wu. 2024. Removing hidden confounding in recommendation: a unified multi-task learning approach. *Advances in Neural Information Processing Systems* 36 (2024).

[42] Haoxuan Li, Yanghao Xiao, Chunyuan Zheng, and Peng Wu. 2023. Balancing unobserved confounding with a few unbiased ratings in debiased recommendations. In *Proceedings of the ACM Web Conference 2023*. 1305–1313.

[43] Xitong Li, Jörn Grahl, and Oliver Hinz. 2022. How do recommender systems lead to consumer purchases? A causal mediation analysis of a field experiment. *Information Systems Research* 33, 2 (2022), 620–637.

[44] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. Causal effect inference with deep latent-variable models. *Advances in Neural Information Processing Systems* 30 (2017).

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

[46] Judea Pearl. 2009. *Causality*. Cambridge university press.

[47] Mattia Prosperi, Yi Guo, Matt Sperrin, James S Koopman, Jae S Min, Xing He, Shannan Rich, Mo Wang, Iain E Buchan, and Jiang Bian. 2020. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence* 2, 7 (2020), 369–375.

[48] JAMES M Robins and MA Hernán. 2016. Causal inference.

[49] Richard W Robins, Avshalom Caspi, and Terrie E Moffitt. 2000. Two personalities, one relationship: both partners' personality traits shape the quality of their relationship. *Journal of Personality and Social Psychology* 79, 2 (2000), 251.

[50] Paul R Rosenbaum. 1987. Model-based direct adjustment. *J. Amer. Statist. Assoc.* 82, 398 (1987), 387–394.

[51] Paul R Rosenbaum and Donald B Rubin. 1983. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological)* 45, 2 (1983), 212–218.

[52] Kara E Rudolph, Nicholas Williams, and Iván Díaz. 2024. Using instrumental variables to address unmeasured confounding in causal mediation analysis. *Biometrics* 80, 1 (2024), ujad037.

[53] Abhin Shah, Karthikeyan Shanmugam, and Murat Kocaoglu. 2024. Front-door adjustment beyond markov equivalence with limited graph knowledge. *Advances in Neural Information Processing Systems* 36 (2024).

[54] Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*. PMLR, 3076–3085.

[55] Claudia Shi, David Blei, and Victor Veitch. 2019. Adapting neural networks for the estimation of treatment effects. *Advances in Neural Information Processing Systems* 32 (2019).

[56] William Torous, Florian Gunsilius, and Philippe Rigollet. 2021. An optimal transport approach to causal inference. *arXiv:2108.05858* (2021).

[57] Victor Veitch and Anisha Zaveri. 2020. Sense and sensitivity analysis: Simple post-hoc analysis of bias due to unobserved confounding. *Advances in Neural Information Processing Systems* 33 (2020), 10999–11009.

[58] Stefan Wager and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* 113, 523 (2018), 1228–1242.

[59] Hao Wang, Jiajun Fan, Zhichao Chen, Haoxuan Li, Weiming Liu, Tianqiao Liu, Quanyu Dai, Yichao Wang, Zhenhua Dong, and Ruiming Tang. 2024. Optimal transport for treatment effect estimation. *Advances in Neural Information Processing Systems* 36 (2024).

[60] Anpeng Wu, Kun Kuang, Bo Li, and Fei Wu. 2022. Instrumental variable regression with confounder balancing. In *International Conference on Machine Learning*. PMLR, 24056–24075.

[61] Lili Wu and Shu Yang. 2022. Integrative R-learner of heterogeneous treatment effects combining experimental and observational studies. In *Conference on Causal Learning and Reasoning*. PMLR, 904–926.

[62] Peng Wu, Haoxuan Li, Yuhao Deng, Wenjie Hu, Quanyu Dai, Zhenhua Dong, Jie Sun, Rui Zhang, and Xiao-Hua Zhou. 2022. On the Opportunity of Causal Learning in Recommendation Systems: Foundation, Estimation, Prediction and Challenges. In *International Joint Conference on Artificial Intelligence*. Survey Track.

[63] Yunfu Xu and Aiya Li. 2020. The relationship between innovative human capital and interprovincial economic growth based on panel data model and spatial econometrics. *J. Comput. Appl. Math.* 365 (2020), 112381.

[64] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. 2018. Representation learning for treatment effect estimation from observational data. *Advances in Neural Information Processing Systems* 31 (2018).

[65] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. 2019. ACE: Adaptively similarity-preserved representation learning for individual treatment effect estimation. In *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1432–1437.

[66] Liuyi Yao, Yaliang Li, Sheng Li, Mengdi Huai, Jing Gao, and Aidong Zhang. 2021. SCI: subspace learning based counterfactual inference for individual treatment effect estimation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 3583–3587.

[67] Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. 2018. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*.

[68] Emily C Zabor, Alexander M Kaizer, and Brian P Hobbs. 2020. Randomized controlled trials. *Chest* 158, 1 (2020), S79–S87.

[69] Weijia Zhang, Lin Liu, and Jiuyong Li. 2021. Treatment effect estimation with disentangled latent factors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 10923–10930.

[70] Yao Zhang, Alexis Bellot, and Mihaela Schaar. 2020. Learning overlapping representations for the estimation of individualized treatment effects. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1005–1014.

[71] Kailiang Zhong, Fengtong Xiao, Yan Ren, Yaorong Liang, Wenqing Yao, Xiaofeng Yang, and Ling Cen. 2022. Descn: Deep entire space cross networks for individual treatment effect estimation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4612–4620.

[72] Minqin Zhu, Anpeng Wu, Haoxuan Li, Ruoxuan Xiong, Bo Li, Xiaoqing Yang, Xuan Qin, Peng Zhen, Jiecheng Guo, Fei Wu, and Kun Kuang. 2024. Contrastive balancing representation learning for heterogeneous dose-response curves estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 17175–17183.

[73] Xinyuan Zhu, Yang Zhang, Xun Yang, Dingxian Wang, and Fuli Feng. 2024. Mitigating hidden confounding effects for causal recommendation. *IEEE Transactions on Knowledge and Data Engineering* (2024).

[74] Hao Zou, Peng Cui, Bo Li, Zheyan Shen, Jianxin Ma, Hongxia Yang, and Yue He. 2020. Counterfactual prediction for bundle treatment. *Advances in Neural Information Processing Systems* 33 (2020), 19705–19715.