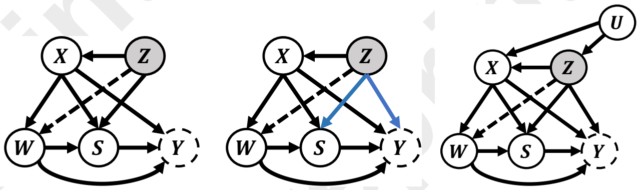
Long-Term Individual Causal Effect Estimation via Identifiable Latent Representation Learning

Ruichu Cai[1,2] , **Junjie Wan**[1] , **Weilin Chen**[1] , **Zeqin Yang**[1,4] , **Zijian Li**[3] , **Peng Zhen**[4] and **Jiecheng Guo**[4]
[1]School of Computer Science, Guangdong University of Technology, Guangzhou, China
[2]Peng Cheng Laboratory, Shenzhen, China
[3]Mohamed bin Zayed University of Artificial Intelligence, Masdar City, Abu Dhabi
[4]DiDi China Ride Hailing Business Group, Beijing, China
{cairuichu, wj1205131700, chenweilin.chn, youngzeqin, leizigin}@gmail.com, {zhenpeng, jasonguo}@didiglobal.com

Abstract

Estimating long-term causal effects by combining long-term observational and short-term experimental data is a crucial but challenging problem in many real-world scenarios. In existing methods, several ideal assumptions, e.g. latent unconfoundedness assumption or additive equi-confounding bias assumption, are proposed to address the latent confounder problem raised by the observational data. However, in real-world applications, these assumptions are typically violated which limits their practical effectiveness. In this paper, we tackle the problem of estimating the long-term individual causal effects without the aforementioned assumptions. Specifically, we propose to utilize the natural heterogeneity of data, such as data from multiple sources, to identify latent confounders, thereby significantly avoiding reliance on idealized assumptions. Practically, we devise a latent representation learning-based estimator of long-term causal effects. Theoretically, we establish the identifiability of latent confounders, with which we further achieve long-term effect identification. Extensive experimental studies, conducted on multiple synthetic and semi-synthetic datasets, demonstrate the effectiveness of our proposed method.



(a) (b) (c)

Figure 1: Three causal graphs in long-term scenarios with X being the pre-treatment variables, Y being the long-term outcome, Z being the latent confounders, S being short-term outcome, U being the auxiliary variable, and W being the treatment. White nodes denote the observed variables and grey nodes denote the unobserved variables. The dashed edges exist in the observational data but are absent in the experimental data. The dashed node Y means Y can be observed in observational data but not in experimental data. Specifically, Fig. 1a shows the causal graph satisfying the latent confoundedness assumption. Fig. 1b shows the causal graph satisfying the equi-confounding bias assumption, where the blue arrows in Fig. 1b indicate the equal confounding bias. Fig. 1c shows the causal graph of our setting.

1 Introduction

Estimating long-term causal effects is of increasing importance in many domains, such as healthcare, public education, marketing, and public policy [Hohnhold *et al.*, 2015; Chetty *et al.*, 2011; Fleming *et al.*, 1994; Zheng *et al.*, 2025]. In such long-term scenarios, it is usually difficult to conduct randomized control experiments to estimate the causal effects. Hence, a lot of researchers resort to the more easily accessible long-term observations. However, methods based on observational data still suffer from the latent confounding bias problem. Therefore, combining observational data and experimental data has emerged as a promising solution for estimating long-term causal effects [Imbens *et al.*, 2024; Ghassami *et al.*, 2022; Hu *et al.*, 2022].

Existing data combination-based methods estimate long-term effects mainly based on the so-called *surrogate*. As shown in Fig. 1, the surrogate S is the short-term outcome, serving as the supplement or replacement for the long-term outcome Y in observational data. However, the unconfoundedness assumption is usually violated in such observational data due to the existence of latent confounders Z. As a replacement for unconfoundedness assumption, [Athey *et al.*, 2020] propose an assumption named latent unconfoundedness, i.e., $Y(w) \perp W|X, S(w)$ on observational data, implicitly indicating the latent confounders Z cannot affect long-term outcome Y as illustrated in Fig. 1a. Alternatively, to relax the unconfoundedness assumption, [Ghassami *et al.*, 2022] introduces the (conditional) additive equi-confounding bias assumption, i.e., the magnitude of the confounding bias for the short-term and the long-term potential outcome variables are the same, as illustrated in Fig. 1b.

Existing methods, however, encounter a **key challenge**: the ideal assumptions are usually violated in real-world applications, including both the latent unconfoundedness and additive equi-confounding bias assumptions, which limit their practical effectiveness. For example, in studying the effect of driver income (treatment W) on long-term retention (outcome Y) in a ride-hailing platform, driver characteristics (pre-treatment variable X) act as observed confounders affecting both income and retention. However, the drivers' household expenses (latent confounders Z) may also affect drivers' long-term retention Y, violating the latent unconfoundedness assumption. Similarly, the additive equi-confounding bias assumption may be violated since household expenses can influence short- and long-term retention differently, i.e., the confounding bias varies over time rather than remaining constant. Therefore, the strong assumptions in existing methods still significantly limit their applicability.

To address the above challenge, we aim to develop a method without the above assumptions to estimate the individual long-term causal effects as shown in Fig. 2. Specifically, instead of assuming latent unconfoundedness or equi-confounding bias, we explore the identifiability of latent confounders Z to estimate long-term causal effects. To identify latent confounder Z, we resort to an additional auxiliary variable U, which is easily accessible from our readily available prior knowledge, such as the natural heterogeneity of data in real-world applications. Recall the aforementioned drivers' income study example, the data are usually collected from various cities, and the indicator variable of the city can be directly taken as the auxiliary variable. Leveraging the identifiability of Z, we establish the causal effect identification result and propose the corresponding latent representation learning-based estimator for long-term individual causal effects. Overall, our contributions can be summarized as follows:

- We focus on a more general setting for estimating long-term causal effects, as shown in Fig. 1c. As shown in Fig. 1c, the assumed causal graph in our paper is a complete graph, and the causal graphs in existing work [Athey *et al.*, 2020; Ghassami *et al.*, 2022] can be seen as our special cases.
- We theoretically achieve the identifiability of latent confounders. Leveraging the identifiability result, we further establish the identification of long-term individual effects.
- We devise a latent representation learning-based estimator for effect estimation. The effectiveness of our estimator is verified on five synthetic and two real-world datasets.[1]

2 Related Works

Variational Auto-encoders for Causal Inference Variational Auto-encoder (VAE) [Kingma and Welling, 2014] is a powerful tool to capture latent structure in different kinds of applications, e.g., image processing [Gregor *et al.*, 2015] and time-series [Chung *et al.*, 2015; Cai *et al.*, 2025]. In causal inference, VAE is used to recover unobserved variables to achieve the identification and estimation of the effects. Without unconfoundedness assumption, CEVAE [Louizos *et al.*, 2017] assumes that latent confounders can be recovered by their proxies and applies VAE to learn confounders. As a follow-up work, TEDVAE [Zhang *et al.*, 2021] and DMAVAE [Xu *et al.*, 2023] decouple the learned latent confounders into several factors to achieve a more accurate estimation of treatment effects in different settings. With the recent development of VAE, nonlinear independent component analysis theory [Hyvarinen and Morioka, 2016] enables the identifiability of recovered variables, e.g., iVAE [Khemakhem *et al.*, 2020] and SIG [Li *et al.*, 2023]. CFDiVAE [Xu *et al.*, 2024] apply iVAE to recover the front-door adjustment variable, achieving effect identification under the front-door criterion [Pearl, 2009]. β-Intact-VAE [Wu and Fukumizu, 2022] utilizes iVAE to recover prognostic scores to estimate effects under a limited overlap setting. **Different** from them, we achieve long-term individual effect identification and estimation by applying iVAE to recover the latent confounders.

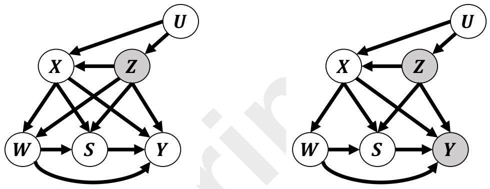
Long-term Causal Inference For decades, many works have explored what a valid surrogate is that can reliably predict long-term causal effects. Different types of criteria are proposed, e.g., prentice criteria [Prentice, 1989] and so on [Frangakis and Rubin, 2002; Lauritzen *et al.*, 2004]. Recently, many works have explored estimating long-term causal effects based on surrogates via data combination. Under the unconfoundedness assumption, LTEE [Cheng *et al.*, 2021] and Laser [Cai *et al.*, 2024] are based on different designed neural networks for long-term causal inference. EETE [Kallus and Mao, 2024] studies the data efficiency from the surrogate and proposes efficient treatment effect estimation. Some works [Wu *et al.*, 2024; Yang *et al.*, 2024a] also focus on balancing short- and long-term rewards under the unconfoundedness assumption. Under surrogacy assumption, SInd [Athey *et al.*, 2019] constructs the Surrogate Index as the substitutions for long-term outcomes in the experimental data to achieve effect identification and [Singh, 2022] propose a kernel ridge regression-based estimator for long-term effect under continuous treatment. As follow-up work, [Athey *et al.*, 2020] assumes latent unconfoundedness assumption, i.e., short-term potential outcomes can mediate the long-term potential outcomes, to identify long-term causal effects. Under this assumption, several methods [Yang *et al.*, 2024b; Chen and Ritzwoller, 2023] are proposed to estimate long-term effects more accurately. Other feasible assumptions are proposed to replace the latent unconfoundedness assumption, e.g., the additive equi-confounding bias assumption [Ghassami *et al.*, 2022; Chen *et al.*, 2025a] and its variant [Chen *et al.*, 2025b]. Based on proximal methods, the sequential structure surrogates are studied [Imbens *et al.*, 2024]. **Different** from them, we focus on estimating long-term individual causal effects in a more general scenario as shown in Fig. 1c.

3 Problem Definition

3.1 Notations

Our notations follow the potential outcome framework [Rubin, 1974]. Let $W \in \{0, 1\}$ be a binary treatment variable. Let $d_\circ$ be the dimension of variable $\circ$. Let $X \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$

[1]The extended version is available at http://arxiv.org/abs/2505.05192.



(a) Observational data (b) Experimental data

Figure 2: Two causal graphs in our setting. The white nodes denote observed variables and the grey nodes denote unobserved variables. Fig. 2a is the causal graph of observational data in our setting. Fig. 2b is the causal graph of experimental data in our setting.

be pre-treatment variable, $Z \in \mathcal{Z} \subseteq \mathbb{R}^{d_z}$ be latent confounders, $S \in \mathcal{S} \subseteq \mathbb{R}^{d_s}$ be the short-term outcome variable, $Y \in \mathcal{Y} \subseteq \mathbb{R}$ be the long-term outcome variable, and $U \in \mathcal{U} \subseteq \mathbb{R}^{d_u}$ be the auxiliary variable. Further, we denote the potential short-term outcomes $S(w) \in \mathbb{R}^{d_s}$ and potential long-term outcomes $Y(w) \in \mathbb{R}$. Denote $G \in \{o, e\}$ be the indicator of the data group, where $G = o$ indicates the observational data, and $G = e$ indicates the experimental data. Let lowercase letters (e.g., x, y) denote the value of random variables. Let lowercase letters with superscript (i) denote the value of the specified i-th unit. Following existing work [Athey *et al.*, 2020; Hu *et al.*, 2022; Ghassami *et al.*, 2022], we consider the data combination setting. We have two types of data: the experimental data $\mathbb{D}_{exp} = \{x^{(i)}, w^{(i)}, s^{(i)}, u^{(i)}, g^{(i)} = e\}_{i=1}^{n_e}$ and the observational data $\mathbb{D}_{obs} = \{x^{(i}, w^{(i)}, s^{(i)}, y^{(i)}, u^{(i)}, g^{(i)} = o\}_{i=1+n_e}^{n_o}$, where n_e, n_o are the sample sizes of experimental and observational data respectively. Our setting is described in Fig. 2.

3.2 Assumptions and Target Estimands

Throughout this paper, we make the following assumptions:

Assumption 1 (Long-term Effect Identification Assumptions)**.** *[Athey* et al.*, 2020; Ghassami* et al.*, 2022]*

- ***A1 [Consistency, Positivity]*** *If $W = w$, then $Y = Y(w)$ and $S = S(w)$. $\forall w, x$, $0 < P(W = w|X = x) < 1$, $0 < P(G = o|W = w, X = x) < 1$.*
- ***A2 [Weak internal validity of observational data]*** *for all $w \in \{0, 1\}$, $W \perp\!\!\!\perp \{Y(w), S(w)\} | X, Z, G = o$.*
- ***A3 [Internal validity of experimental data]*** *for all $w \in \{0, 1\}$, $W \perp\!\!\!\perp \{Y(w), S(w)\} | X, G = e$.*
- ***A4 [External validity of experimental data]*** *for all $w \in \{0, 1\}$, $G \perp\!\!\!\perp \{Y(w), S(w)\} | X$.*

The assumptions above are mild and widely used in existing literature, e.g., [Athey *et al.*, 2020; Ghassami *et al.*, 2022]. A1 is a standard assumption. A2 allows the existence of latent confounders Z. A3 guarantees that the experimental data is unconfounded conditioned on X. A4 allows us to generalize the conditional distribution of potential outcomes between observational and experimental data.

In this paper, our **task** is to estimate the long-term individual treatment effects (ITE) given $\mathbb{D}_{exp}, \mathbb{D}_{obs}$, defined as:

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

as well as long-term average treatment effects (ATE), defined as:

$$\tau = \mathbb{E}[\tau(x)]. \tag{2}$$

4 Methodology

In this section, we present our end-to-end long-term causal effect estimator. Overall, as shown in Fig. 3, our estimator consists of three modules: short-term potential outcome estimation, latent representation learning, and ITE estimation. In the short-term potential outcome estimation module, we train an estimator for $p(S(w)|W, X)$ using experimental data, as it is identifiable as $p(S|W, X)$. In the latent representation learning module, we leverage variational inference to learn the latent representation of confounders Z. The pre-treatment variable X, treatment W and the short-term potential outcome $S(w)$, obtained from the short-term potential outcome estimation module, are jointly treated as proxies for Z, ensuring sufficient information is available to recover Z. Additionally, the auxiliary variable U is used as a prior, guaranteeing the identifiability of the latent confounder Z, as demonstrated in the theoretical analysis (see Section 5).

In the ITE estimation module, based on learned Z, we conduct an estimator to learn the potential outcomes in treated and control groups, resulting in the final estimator of $\tau(x)$. Note that the first module is trained on experimental data to ensure the identification of short-term potential outcomes, and the others are trained on observational data since the long-term outcome is only observed in observational data.

4.1 Short-term Potential Outcome Estimation

We employ a multilayer perceptron (MLP) to model the distribution of $p(S(w)|X)$ as our short-term potential outcome estimator. Since we can access short-term experimental data, $p(S(w)|X)$ can be rewritten as $p(S|X, W = w)$ on experimental data. To estimate that, inspired by Tarnet [Johansson *et al.*, 2022], we use two heads of MLP for the estimation. Specifically, we can model each dimension of $S(w)$ as a Gaussian distribution as follows:

$$p(S|W, X) = \prod_{i=0}^{d_s} \mathcal{N}(\mu = \hat{\mu}_{S_i}, \sigma^2 = \hat{\sigma}^2_{S_i}), \tag{3}$$

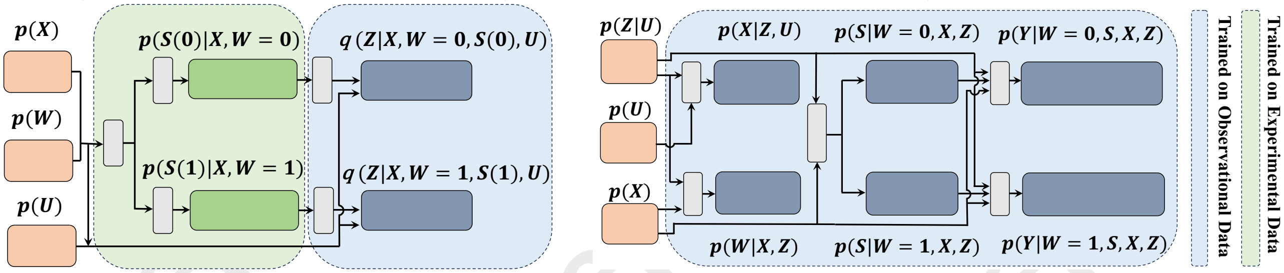
where $\hat{\mu}_{S_i}$ and $\hat{\sigma}_{S_i}$ are the mean and variance of the Gaussian distribution parametrized by the MLPs. We use the negative log-likelihood of Eq. (3) as the objective function $\mathcal{L}_{S(w)}$ for the short-term potential outcome estimator as follows:

$$\begin{aligned}\mathcal{L}_{S(w)} &= -\mathbb{E}_{q_{\mathbb{D}_{exp}}}[\log p(S(w)|X)] \\ &= -\mathbb{E}_{q_{\mathbb{D}_{exp}}}[\log p(S|X, W)],\end{aligned} \tag{4}$$

where $q_{\mathbb{D}_{exp}}$ is the empirical data distribution given by $\mathbb{D}_{exp}$.

4.2 Latent Representation Learning

In the latent representation learning step, we employ iVAE to recover latent confounders Z, as shown in Fig. 3. This module consists of two networks: an inference network and a generative network. Specifically, for the inference network, the auxiliary variable U serves as additional information and



(a) Inference network

(b) Generative network

Figure 3: Overall architecture of the generative and inference networks for our model. Grey nodes represent MLP, green nodes correspond to the distribution trained on experimental data and blue nodes correspond to the distribution trained on observational data.

thus our prior distribution is $p(Z|U)$. We further use the posterior distribution $q(Z|S(1), S(0), X, W, U)$ to approximate the prior, where the short-term potential outcomes are obtained by the short-term potential outcome estimator discussed in the previous section. For the generative network, we reconstruct the treatment W, the short-term outcome S and the pre-treatment covariate X.

Following exiting VAE-based works [Louizos *et al.*, 2017], we choose the prior $p(Z|U)$ as Gaussian distribution:

$$p(Z|U) = \prod_{i=0}^{d_z} \mathcal{N}(Z_i|\hat{\mu}_i, \hat{\sigma}_i^2), \tag{5}$$

where $\hat{\mu}_i$ and $\hat{\sigma}_i$ are the mean and variance of the Gaussian distribution parametrized by the MLPs.

To approximate the prior, we model the posterior distribution $q(Z|S(1), S(0), X, W, U)$ as Gaussian distribution:

$$\begin{aligned} q_0(Z|S(0), X, U) &= \prod_{i=0}^{d_z} \mathcal{N}(\mu = \hat{\mu}_{Z_i|W=0}, \sigma^2 = \hat{\sigma}^2_{Z_i|W=0}), \\ q_1(Z|S(1), X, U) &= \prod_{i=0}^{d_z} \mathcal{N}(\mu = \hat{\mu}_{Z_i|W=1}, \sigma^2 = \hat{\sigma}^2_{Z_i|W=1}), \\ & q(Z|W, S(1), S(0), X, U) \\ =& W \cdot q_1(Z|S(1), X, U) + (1 - W) \cdot q_0(Z|S(0), X, U), \end{aligned} \tag{6}$$

where $\hat{\mu}_{Z_i|W=0}$ and $\hat{\sigma}_{Z_i|W=0}$ are the mean and variance of the Gaussian distribution parametrized by MLPs whose inputs are X, W, U and estimated $S(W)$, and similarly for $\hat{\mu}_{Z_i|W=1}$ and $\hat{\sigma}_{Z_i|W=1}$.

In the generative network, for a continuous variable, we parametrize the distribution as a Gaussian with its mean and variance both given by MLPs. For a binary variable, we use a Bernoulli distribution parametrized by an MLP similarly. Thus, we employ the following distributions for $p(X|Z, U)$:

$$\begin{aligned} p(X|Z, U) &= \prod_{i=0}^{d_X} \mathcal{N}(\mu = \hat{\mu}_{X_i}, \sigma^2 = \hat{\sigma}^2_{X_i}) \\ \text{or} \quad p(X|Z, U) &= \prod_{i=0}^{d_X} \textbf{Bern}(\pi = \hat{\pi}_{X_i}), \end{aligned} \tag{7}$$

where $\hat{\mu}_{X_i}$ and $\hat{\sigma}_{X_i}$ are the mean and variance of the Gaussian distribution parametrized by MLPs in the generative network when the variable is continuous, and $\hat{\pi}_{X_i}$ is the mean of Bernoulli distribution parametrized by the generative network when the variable is binary. Similarly, we employ the following distributions for $p(W|X, Z)$ and $p(S|W, X, Z)$:

$$\begin{aligned} p(W|X, Z) &= \textbf{Bern}(\pi = \hat{\pi}_{W_i}), \\ p(S|W, X, Z) &= \prod_{i=0}^{d_s} \mathcal{N}(\mu = \hat{\mu}_{S_i'}, \sigma^2 = \hat{\sigma}^2_{S_i'}) \\ \text{or} \quad p(S|W, X, Z) &= \prod_{i=0}^{d_s} \textbf{Bern}(\pi = \hat{\pi}_{S_i'}), \end{aligned} \tag{8}$$

where $\hat{\mu}_{S_i'}$, $\hat{\sigma}_{S_i'}$, $\hat{\pi}_{S_i'}$ and $\hat{\pi}_{W_i}$ are all parametrized by the generative network. We then use the negative variational Evidence Lower Bound (ELBO) as the objective function for the inference and generative networks (see Appendix E for the derivations):

$$\begin{aligned} \textbf{ELBO} = \mathbb{E}_{q_{\mathbb{D}_{obs}}}[\mathbb{E}_{q(Z|S(0),S(1),X,U,W)}[&\log p(Z|U) \\ + \log p(X|Z, U) + \log p(W|X, Z) &+ \log p(S|W, X, Z) \\ - \log q(Z|S(0), S(1), X, U, W)]], \end{aligned} \tag{9}$$

where $q_{\mathbb{D}_{obs}}$ is the empirical data distributions given by $\mathbb{D}_{obs}$.

4.3 ITE Estimation

To obtain the outcome Y, we introduce an auxiliary distribution that helps predict long-term outcome Y. Specifically, we employ the following distribution for $p(Y|W, S, X, Z)$:

$$p(Y|W, S, X, Z) = \mathcal{N}(\mu = \hat{\mu}_{y_i}, \sigma^2 = \hat{\sigma}^2_{y_i}), \tag{10}$$

where $\hat{\mu}_{y_i}$ and $\hat{\sigma}_{y_i}$ are the mean and variance of the Gaussian distribution parametrized by MLPs. We then use the negative log-likelihood as its objective function:

$$\mathcal{L}_Y = -\mathbb{E}_{q_{\mathbb{D}_{obs}}}[\mathbb{E}_{q(Z|S(0),S(1),X,U,W)}[\log p(Y|W, S, X, Z)]]. \tag{11}$$

Overall, our final objective function $\mathcal{L}$ is

$$\mathcal{L} = -\textbf{ELBO} + \mathcal{L}_{S(w)} + \mathcal{L}_Y. \tag{12}$$

As a result, after training our method on experimental and observational data, given specific unit $x^{(i)}, u^{(i)}$, our final estimator yields long-term potential outcomes $\hat{y}(1)^{(i)}, \hat{y}(0)^{(i)}$ on the treated and control group respectively. Thus the estimated long-term individual effect of $x^{(i)}, u^{(i)}$ is

$$\hat{\tau}(x^{(i)}) = \hat{y}(1)^{(i)} - \hat{y}(0)^{(i)}. \tag{13}$$

5 Theoretical Analysis

In this section, we present the identifiability result of our model and the identification of long-term individual causal effects. If we can correctly identify the latent confounders Z, the long-term individual causal effect can be identified based on the learned representation of Z. We first prove that Z is identifiable up to a simple transformation. Leveraging the identifiability result of Z, we further prove that the long-term individual causal effect is identifiable.

5.1 Identifiability of Latent Confounders

To clearly introduce the latent confounders identifiability result, we first denote Z_i as the i-th dimension of Z. The identifiability of latent confounders means that, for each ground-truth latent confounder Z_i, there exist a corresponding estimated latent confounder $\hat{Z}_i$ and an invertible function $h_i : \mathbb{R} \to \mathbb{R}$, such that $Z_i = h_i(\hat{Z}_j)$. Please refer to Appendix A for the formal definition of identifiability.

We show that latent confounders can be identified up to permutation and invertible component-wise transformations.

Theorem 1. *Suppose the data-generation process follows Fig. 2 and the following conditions hold:*

- *Smooth and Positive Density: The probability density function of latent confounders is smooth and positive, i.e., $p_{Z|U}$ is smooth and $p_{Z|U} > 0$ over $\mathcal{Z}$ and $\mathcal{U}$.*
- *Conditional Independence: Conditioned on U, each Z_i is independent, i.e., $\forall i, j \in \{1, ..., d_z\}$, $i \neq j$, $\log p_{Z|U}(Z|U) = \sum_i^{d_z} q_i(Z_i, U)$ where q_i is the log density of the conditional distribution, i.e., $q_i := \log p_{Z_i|U}$.*
- *Linear Independence: For any $Z \in \mathcal{Z} \subseteq \mathbb{R}^{d_z}$, there exist $2d_z + 1$ values of U, i.e., u_j with $j = 0, 1, ..., 2d_z$, such that the $2d_z$ vectors $w(Z, u_j) - w(Z, u_0)$ with $j = 1, ..., 2d_z$, are linearly independent, where vector $w(Z, U)$ is defined as follows:*

$$w(Z,U) = \Bigg(\frac{\partial q_1(Z_1,U)}{\partial Z_1}, \dots, \frac{\partial q_{d_z}(Z_{d_z},U)}{\partial Z_{d_z}}, \frac{\partial^2 q_1(Z_1,U)}{\partial {Z_1}^2}, \dots, \frac{\partial^2 q_{d_z}(Z_{d_z},U)}{\partial {Z_{d_z}}^2} \Bigg). \tag{14}$$

By modeling the aforementioned data generation process in Fig. 2, latent confounders Z are identifiable.

Proof is given in Appendix C. The first two conditions are standard in the identifiability of existing nonlinear ICA works, e.g., [Kong *et al.*, 2022; Khemakhem *et al.*, 2020]. More importantly, the third condition means that the auxiliary variable contains enough information, i.e., at least $2d_z+1$ distinct values of U. This assumption is plausible due to the nature of the heterogeneity of data, e.g., data from 11 cities can ensure the identifiability of Z with up to 5 dimensions. Please refer to Appendix B for more implications of these conditions.

5.2 Identifiability of Long-term ITE

Building on the identifiability of latent confounders, in this section, we can further achieve the identification of long-term ITE. As stated in Theorem 1, the latent confounder Z is identified up to simple invertible transformation, i.e., $\hat{Z} = h^{-1}(Z)$. Note the identifiability provides a fine-grained theoretical guarantee, ensuring all information of Z is preserved. Thus, with the learned $\hat{Z}$, the long-term causal effects can be identified, as stated in the following theorem.

Theorem 2. *Under Assumption 1, suppose Theorem 1 hold, and then $\tau(x) = \mathbb{E}[Y(1) - Y(0)|X = x]$ is identifiable.*

The proof is given in Appendix D. Theorem 2 theoretically guarantees the correctness of our model, providing a feasible technology of long-term individual causal effects estimation via learning latent confounders.

6 Experiments

In this section, we verify the effectiveness of our model and the correctness of our theory. Specifically, we answer the following questions:

1. **Can our model identify latent confounders Z?**
2. **Does our model perform well on datasets that follow different existing assumptions?**
3. **Does our model outperform baselines on the real-world datasets?**
4. **Is our method robust to different strengths of latent confounding?**

6.1 Experimental Setup

Datasets Since the ground-truth potential outcome can not be observed in the real world, following existing literature [Louizos *et al.*, 2017; Cheng *et al.*, 2021; Cai *et al.*, 2024; Yang *et al.*, 2024b], we use synthetic and semi-synthetic data to evaluate our method and baselines.

For the synthetic data, we simulate five synthetic datasets in our paper. To validate the generalizability of our method, we first simulate three datasets corresponding to the causal graphs in Table 1. The first synthetic dataset allows all the existence of edges following the assumed causal graph in our paper. The second synthetic dataset follows the latent unconfoundedness assumption [Athey *et al.*, 2020] that rules out the edges from unobserved confounders Z to long-term outcome Y. The third dataset follows the additive equi-confounding bias assumption [Ghassami *et al.*, 2022] that assumes the short-term confounding bias is equal to the long-term one. To further analyze the performances in terms of different strengths of confounding bias, we simulate the fourth synthetic dataset with varying β, which controls the coefficients in the data generation function from Z to W and Z to Y. Finally, we simulate the fifth synthetic dataset to verify that our method is able to identify Z. All data generation details can be found in Appendix F.

For the semi-synthetic data, we use IHDP [Hill, 2011] and TWINS [Almond *et al.*, 2005] to validate our model's performance on complex real-world data. In detail, we reuse their original features and divide them into pre-treatment variables

	Synthetic 1		Synthetic 2		Synthetic 3	
	ϵ_{ATE}	ϵ_{ITE}	ϵ_{ATE}	ϵ_{ITE}	ϵ_{ATE}	ϵ_{ITE}
CEVAE [Louizos *et al.*, 2017]	$3.902_{\pm 0.740}$	$4.162_{\pm 0.781}$	$0.146_{\pm 0.037}$	$0.270_{\pm 0.056}$	$0.877_{\pm 0.161}$	$0.975_{\pm 0.181}$
TEDVAE [Zhang *et al.*, 2021]	$4.356_{\pm 1.078}$	$4.851_{\pm 1.183}$	$0.260_{\pm 0.109}$	$0.397_{\pm 0.111}$	$0.941_{\pm 0.186}$	$1.171_{\pm 0.199}$
LTEE [Cheng *et al.*, 2021]	$4.815_{\pm 1.269}$	$5.726_{\pm 1.662}$	$0.373_{\pm 0.232}$	$0.596_{\pm 0.288}$	$0.985_{\pm 0.174}$	$1.215_{\pm 0.176}$
S-Learner [Künzel *et al.*, 2019]	$2.916_{\pm 0.854}$	$4.185_{\pm 1.027}$	$0.106_{\pm 0.171}$	$0.500_{\pm 0.300}$	$0.208_{\pm 0.159}$	$2.235_{\pm 1.493}$
T-Learner [Künzel *et al.*, 2019]	$5.554_{\pm 2.733}$	$7.687_{\pm 3.529}$	$0.310_{\pm 0.308}$	$0.746_{\pm 0.435}$	$0.836_{\pm 0.917}$	$1.832_{\pm 0.763}$
Imputaion [Athey *et al.*, 2020]	$2.480_{\pm 2.290}$	-	$0.628_{\pm 0.542}$	-	$0.956_{\pm 1.094}$	-
Weighting [Athey *et al.*, 2020]	$11.579_{\pm 6.775}$	-	$1.896_{\pm 1.801}$	-	$0.854_{\pm 0.901}$	-
Equi-naive [Ghassami *et al.*, 2022]	$2.837_{\pm 1.377}$	$4.297_{\pm 2.080}$	$0.153_{\pm 0.145}$	$0.974_{\pm 0.268}$	$\mathbf{0.185}_{\pm 0.190}$	$1.927_{\pm 0.443}$
IF-base [Ghassami *et al.*, 2022]	$9.385_{\pm 7.690}$	-	$1.600_{\pm 2.030}$	-	$4.846_{\pm 3.716}$	-
ICEVAE	$\mathbf{2.402}_{\pm 0.436}$	$\mathbf{3.173}_{\pm 0.418}$	$\mathbf{0.105}_{\pm 0.068}$	$\mathbf{0.137}_{\pm 0.064}$	$0.427_{\pm 0.385}$	$\mathbf{0.695}_{\pm 0.364}$

Table 1: Results of estimation error regarding ATE and ITE on three synthetic datasets. We report mean±std results. - means the method is not applicable. The best is bolded.

X, unobserved confounders Z and the auxiliary variables U according to their real-world meanings. Then we divide the samples into experimental and observational data and generate corresponding treatments, short-term outcomes, and long-term outcomes. The feature division and data generation details can be found in Appendix F.

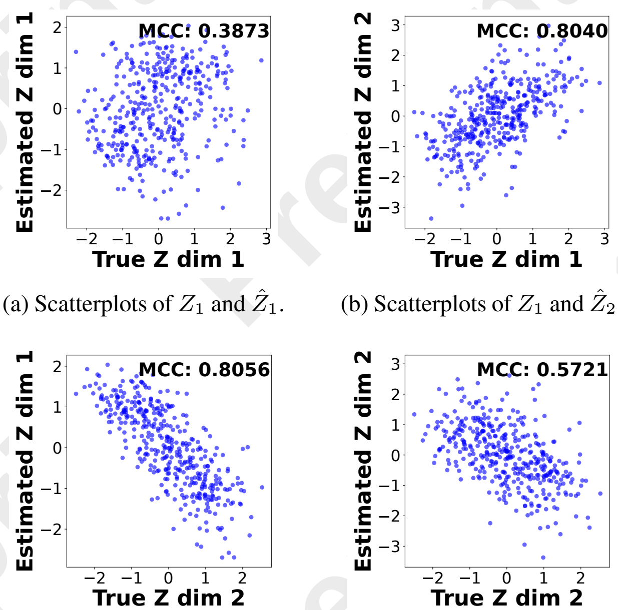
Baselines and Metrics We compare our model **ICEVAE**[2] with the following baselines designed for long-term causal effect, including the **Imputation** and the **Weighting** approaches [Athey *et al.*, 2020], the naive estimator and the efficient influence function-based estimator under Conditional Additive Equi-Confounding Bias assumption [Ghassami *et al.*, 2022], named **Equi-naive** and **IF-based** respectively, and **LTEE** [Cheng *et al.*, 2021]. Besides, since there is a lack of work on estimating heterogeneous long-term causal effects, we use **CEVAE** [Louizos *et al.*, 2017] as one of the baselines, as it is designed for recovered latent confounders in effects estimation. We also compare our model with the follow-up work **TEDVAE** [Zhang *et al.*, 2021]. Finally, we introduce two simple estimators, the **S-Learner** and the **T-Learner** [Künzel *et al.*, 2019] to be baselines, which are implemented using MLPs. Note that the **Imputation** method, the **Weighting** method, and the **IF-based** method are designed for ATE and cannot estimate ITE. The implementation details regarding baselines and our method can be found in Appendix F.

For metrics, to measure the error of average causal effect estimation, we report the mean and the standard deviation(std) of mean square error ϵ_{ATE} on the test set by performing 5 replications, i.e., $\epsilon_{ATE} = (\tau - \hat{\tau})^2$, where τ and $\hat{\tau}$ are the real and estimated average treatment effects on the test set respectively. To measure the error of estimating individual causal effects, we report the mean and std of Precision in the Estimation of Heterogeneous Effect (PEHE) ϵ_{ITE} on the test set by performing 5 replications where $\epsilon_{ITE} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \left(\tau(x^{(i)}) - \hat{\tau}(x^{(i)})\right)^2$, where n_{test} is the test sample size.

[2]Code is available at https://github.com/DMIRLAB/ICEVAE and https://github.com/learnwjj/ICEVAE.

6.2 Results and Analysis

Can our model identify latent confounders Z?
To validate the correctness of Theorem 1, we apply our method to the **Synthetic 5** dataset. As shown in Fig. 4, the latent variables are successfully recovered, with a high MCC metric calculated by the ground-truth Z and estimated Z. Fig. 4 suggests that the latent causal variables are estimated up to permutation and component-wise invertible transformation, i.e., the estimated Z_1 in the figure corresponds to the true Z_2, with an MCC value of 0.8056. The estimated Z_2 corresponds to the true Z_1, with an MCC value of 0.8040. This indicates that our proposed method is able to identify Z, which verifies the correctness of our Theorem 1.



(a) Scatterplots of Z_1 and $\hat{Z}_1$. (b) Scatterplots of Z_1 and $\hat{Z}_2$.

(c) Scatterplots of Z_2 and $\hat{Z}_1$. (d) Scatterplots of Z_2 and $\hat{Z}_2$.

Figure 4: Result on the fifth synthetic dataset. Fig. 4a-4d show the scatterplots between each ground-truth and estimated latent confounder.

	$\beta=1$		$\beta=1.5$		$\beta=3$		$\beta=4.5$		$\beta=5$	
	ϵ_{ATE}	ϵ_{ITE}	ϵ_{ATE}	ϵ_{ITE}	ϵ_{ATE}	ϵ_{ITE}	ϵ_{ATE}	ϵ_{ITE}	ϵ_{ATE}	ϵ_{ITE}
CEVAE	$0.116_{\pm 0.083}$	$0.188_{\pm 0.081}$	$0.324_{\pm 0.115}$	$0.401_{\pm 0.119}$	$3.902_{\pm 0.740}$	$4.162_{\pm 0.781}$	$14.348_{\pm 3.036}$	$15.406_{\pm 3.493}$	$19.403_{\pm 3.152}$	$20.444_{\pm 3.585}$
TEDVAE	$0.097_{\pm 0.039}$	$0.205_{\pm 0.041}$	$0.351_{\pm 0.065}$	$0.498_{\pm 0.068}$	$4.356_{\pm 1.078}$	$4.851_{\pm 1.183}$	$16.543_{\pm 3.757}$	$18.589_{\pm 4.633}$	$22.046_{\pm 3.856}$	$24.036_{\pm 4.571}$
LTEE	$0.048_{\pm 0.034}$	$0.212_{\pm 0.071}$	$0.296_{\pm 0.141}$	$0.510_{\pm 0.220}$	$4.815_{\pm 1.269}$	$5.726_{\pm 1.662}$	$17.678_{\pm 7.534}$	$20.310_{\pm 9.404}$	$23.980_{\pm 7.371}$	$26.740_{\pm 9.062}$
S-Learner	$\mathbf{0.021}_{\pm 0.013}$	$0.414_{\pm 0.091}$	$\mathbf{0.096}_{\pm 0.090}$	$0.617_{\pm 0.098}$	$2.916_{\pm 0.854}$	$4.186_{\pm 1.027}$	$15.382_{\pm 7.409}$	$19.080_{\pm 8.705}$	$18.842_{\pm 9.137}$	$22.609_{\pm 10.251}$
T-Learner	$0.190_{\pm 0.130}$	$0.582_{\pm 0.193}$	$0.209_{\pm 0.211}$	$0.867_{\pm 0.304}$	$5.554_{\pm 2.733}$	$7.687_{\pm 3.529}$	$17.026_{\pm 6.587}$	$21.598_{\pm 7.321}$	$20.068_{\pm 10.22}$	$26.478_{\pm 9.314}$
Imputation	$0.792_{\pm 0.934}$	-	$0.928_{\pm 1.258}$	-	$2.480_{\pm 2.290}$	-	$13.156_{\pm 7.144}$	-	$19.518_{\pm 11.240}$	-
Weighting	$0.861_{\pm 0.648}$	-	$0.639_{\pm 0.339}$	-	$11.579_{\pm 6.775}$	-	$51.634_{\pm 13.346}$	-	$70.104_{\pm 11.687}$	-
Equi-naive	$0.285_{\pm 0.422}$	$0.823_{\pm 0.353}$	$0.247_{\pm 0.272}$	$1.001_{\pm 0.245}$	$2.837_{\pm 1.377}$	$4.297_{\pm 2.080}$	$10.619_{\pm 14.245}$	$14.245_{\pm 7.583}$	$19.314_{\pm 3.186}$	$22.978_{\pm 3.406}$
IF-base	$0.619_{\pm 0.831}$	-	$1.707_{\pm 1.893}$	-	$9.385_{\pm 7.690}$	-	$30.562_{\pm 13.998}$	-	$32.723_{\pm 12.970}$	-
ICEVAE	$0.038_{\pm 0.031}$	$\mathbf{0.069}_{\pm 0.032}$	$0.182_{\pm 0.086}$	$\mathbf{0.217}_{\pm 0.085}$	$\mathbf{2.402}_{\pm 0.436}$	$\mathbf{3.173}_{\pm 0.418}$	$\mathbf{9.897}_{\pm 2.685}$	$\mathbf{11.395}_{\pm 2.841}$	$\mathbf{14.960}_{\pm 3.363}$	$\mathbf{16.467}_{\pm 3.935}$

Table 2: Results of estimation error regarding ATE and ITE on the fourth synthetic dataset with different strengths of confounding bias controlled by β. We report mean$\pm$std results. - means the method is not applicable. The best is bolded.

Does our model perform well on datasets that follow different existing assumptions?

We conduct experiments by comparing our method with baselines on three different synthetic datasets that follow different data generation processes. The results are shown in Table 1. Overall, on all three datasets, our method achieves almost the best performance, revealing the generalizability of our method under different assumptions. In detail, on the Synthetic 1 dataset, our method achieves the lowest ITE and ATE estimation error and std, indicating the effectiveness of our method. As for the results of the Synthetic 2 dataset, compared with baselines, our method achieves comparable performance. Note that the Imputation and Weighting methods perform much better on the Synthetic 2 dataset than the Synthetic 1 dataset since the Synthetic 2 dataset is designed following the latent unconfoundedness assumption. Similarly, as for the results of the Synthetic 3 dataset that is generated following the additive equi-confounding bias assumption, Equi-naive can achieve the lowest error in terms of ATE estimation. On this dataset, our method also achieves comparable performance, especially in terms of ITE estimation. Hence, we conclude that our model can perform well on datasets that follow different existing assumptions.

Does our model outperform baselines on real-world datasets?

In Table 3, we evaluate the performance of our model on complex real-world data by comparing each method using two semi-synthetic datasets. The main observations are as follows. Overall, our method achieves the best performance regarding ITE estimation and comparable performance regarding ATE estimation, indicating the effectiveness of our method. Specifically, compared with the VAE-based method, our method performs better, which indicates that the experimental data does help recover latent confounders. Compared with the Imputation and Weighting methods, our method strongly outperforms them, since the unsuitable latent unconfoundedness assumption is made by their methods. In conclusion, we find that our method ICEVAE can outperform baselines on real-world datasets.

Is our method robust to different strengths of latent confounding?

In table 2, we compare our model with baselines on the fourth synthetic dataset with different strengths of confounding bias controlled by β. The main observations are as follows. With the strengths of latent confounding increasing, i.e., β from 1 to 5, all methods perform worse, which is reasonable since a large confounding bias will lead to a significant imbalance of distribution between treated and control groups. When the latent confounding is small, traditional methods yield a comparable performance, since the unconfoundedness assumption almost holds. When the latent confounding is large enough, only our method yields accurate estimations in terms of ATE and ITE, which indicates that our method is robust to the latent confounding. It is because our method can correctly recover the latent confounders Z, and it also reveals the necessity of recovering latent confounders.

	IHDP		TWINS	
	ϵ_{ATE}	ϵ_{ITE}	ϵ_{ATE}	ϵ_{ITE}
CEVAE	$\mathbf{0.004}_{\pm 0.003}$	$0.183_{\pm 0.054}$	$0.641_{\pm 0.521}$	$16.818_{\pm 11.867}$
TEDVAE	$0.011_{\pm 0.019}$	$0.188_{\pm 0.032}$	$0.824_{\pm 1.108}$	$16.657_{\pm 11.978}$
LTEE	$0.015_{\pm 0.014}$	$0.668_{\pm 0.132}$	$1.994_{\pm 2.470}$	$15.667_{\pm 14.297}$
T-Leaner	$0.061_{\pm 0.051}$	$1.060_{\pm 0.214}$	$4.665_{\pm 5.505}$	$5.191_{\pm 5.581}$
S-Leaner	$0.020_{\pm 0.018}$	$0.969_{\pm 0.354}$	$2.536_{\pm 4.101}$	$12.288_{\pm 4.779}$
Imputaion	$0.713_{\pm 0.478}$	-	$46.092_{\pm 43.729}$	-
Weighting	$0.664_{\pm 0.959}$	-	$6.597_{\pm 10.859}$	-
ICEVAE	$0.016_{\pm 0.027}$	$\mathbf{0.178}_{\pm 0.060}$	$\mathbf{0.204}_{\pm 0.229}$	$\mathbf{3.665}_{\pm 2.246}$

Table 3: Results of estimation error regarding ATE and ITE on two semi-synthetic datasets. We report mean$\pm$std results. - means the method is not applicable. The best is bolded.

7 Conclusion

In this paper, we provide a practical solution to estimate the long-term individual causal effects in the presence of latent confounders via identifiable representation learning. Our proposed method takes advantage of the natural heterogeneity of data, e.g., data from multiple cities, to identify latent confounders and further estimate the long-term individual effect, which not only helps us avoid the idealized assumptions of the existing methods, but also renders our approach with theoretical guarantees of identifiability. Extensive experimental results verify the correctness of our theory and the effectiveness of our estimator.

Acknowledgments

This research was supported in part by National Science and Technology Major Project (2021ZD0111501), National Science Fund for Excellent Young Scholars (62122022), Natural Science Foundation of China (U24A20233, 62206064, 62206061, 62476163, 62406078), Guangdong Basic and Applied Basic Research Foundation (2023B1515120020), and CCF-DiDi GAIA Collaborative Research Funds (CCF-DiDi GAIA 202311).

References

[Almond *et al.*, 2005] Douglas Almond, Kenneth Y Chay, and David S Lee. The costs of low birth weight. *The Quarterly Journal of Economics*, 120(3):1031–1083, 2005.

[Athey *et al.*, 2019] Susan Athey, Raj Chetty, Guido W Imbens, and Hyunseung Kang. The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Technical report, National Bureau of Economic Research, 2019.

[Athey *et al.*, 2020] Susan Athey, Raj Chetty, and Guido Imbens. Combining experimental and observational data to estimate treatment effects on long term outcomes. *arXiv preprint arXiv:2006.09676*, 2020.

[Cai *et al.*, 2024] Ruichu Cai, Weilin Chen, Zeqin Yang, Shu Wan, Chen Zheng, Xiaoqing Yang, and Jiecheng Guo. Long-term causal effects estimation via latent surrogates representation learning. *Neural Networks*, 176:106336, 2024.

[Cai *et al.*, 2025] Ruichu Cai, Zhifan Jiang, Kaitao Zheng, Zijian Li, Weilin Chen, Xuexin Chen, Yifan Shen, Guangyi Chen, Zhifeng Hao, and Kun Zhang. Learning disentangled representation for multi-modal time-series sensing signals. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 3247–3266, New York, NY, USA, 2025. Association for Computing Machinery.

[Chen and Ritzwoller, 2023] Jiafeng Chen and David M Ritzwoller. Semiparametric estimation of long-term treatment effects. *Journal of Econometrics*, 237(2):105545, 2023.

[Chen *et al.*, 2025a] Weilin Chen, Ruichu Cai, Junjie Wan, Zeqin Yang, and José Miguel Hernández-Lobato. Nonparametric heterogeneous long-term causal effect estimation via data combination. *arXiv preprint arXiv:2502.18960*, 2025.

[Chen *et al.*, 2025b] Weilin Chen, Ruichu Cai, Yuguang Yan, Zhifeng Hao, and José Miguel Hernández-Lobato. Long-term causal inference via modeling sequential latent confounding. *arXiv preprint arXiv:2502.18994*, 2025.

[Cheng *et al.*, 2021] Lu Cheng, Ruocheng Guo, and Huan Liu. Long-term effect estimation with surrogate representation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 274–282, 2021.

[Chetty *et al.*, 2011] Raj Chetty, John N Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. How does your kindergarten classroom affect your earnings? evidence from project star. *The Quarterly journal of economics*, 126(4):1593–1660, 2011.

[Chung *et al.*, 2015] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. *Advances in neural information processing systems*, 28, 2015.

[Fleming *et al.*, 1994] Thomas R Fleming, Ross L Prentice, Margaret S Pepe, and David Glidden. Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and aids research. *Statistics in medicine*, 13(9):955–968, 1994.

[Frangakis and Rubin, 2002] Constantine E Frangakis and Donald B Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.

[Ghassami *et al.*, 2022] AmirEmad Ghassami, Alan Yang, David Richardson, Ilya Shpitser, and Eric Tchetgen Tchetgen. Combining experimental and observational data for identification and estimation of long-term causal effects. *arXiv preprint arXiv:2201.10743*, 2022.

[Gregor *et al.*, 2015] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. In *International conference on machine learning*, pages 1462–1471. PMLR, 2015.

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

[Hohnhold *et al.*, 2015] Henning Hohnhold, Deirdre O'Brien, and Diane Tang. Focusing on the long-term: It's good for users and business. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1849–1858, 2015.

[Hu *et al.*, 2022] Wenjie Hu, Xiaohua Zhou, and Peng Wu. Identification and estimation of treatment effects on long-term outcomes in clinical trials with external observational data. *arXiv preprint arXiv:2208.10163*, 2022.

[Hyvarinen and Morioka, 2016] Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in neural information processing systems*, 29, 2016.

[Imbens *et al.*, 2024] Guido Imbens, Nathan Kallus, Xiaojie Mao, and Yuhao Wang. Long-term causal inference under persistent confounding via data combination. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkae095, 2024.

[Johansson *et al.*, 2022] Fredrik D Johansson, Uri Shalit, Nathan Kallus, and David Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *Journal of Machine Learning Research*, 23(166):1–50, 2022.

[Kallus and Mao, 2024] Nathan Kallus and Xiaojie Mao. On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *Journal of the*

Royal Statistical Society Series B: Statistical Methodology, page qkae099, 2024.

[Khemakhem *et al.*, 2020] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

[Kingma and Welling, 2014] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[Kong *et al.*, 2022] Lingjing Kong, Shaoan Xie, Weiran Yao, Yujia Zheng, Guangyi Chen, Petar Stojanov, Victor Akinwande, and Kun Zhang. Partial disentanglement for domain adaptation. In *International conference on machine learning*, pages 11455–11472. PMLR, 2022.

[Künzel *et al.*, 2019] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

[Lauritzen *et al.*, 2004] Steffen L Lauritzen, Odd O Aalen, Donald B Rubin, and Elja Arjas. Discussion on causality [with reply]. *Scandinavian Journal of Statistics*, 31(2):189–201, 2004.

[Li *et al.*, 2023] Zijian Li, Ruichu Cai, Guangyi Chen, Boyang Sun, Zhifeng Hao, and Kun Zhang. Subspace identification for multi-source domain adaptation. *Advances in Neural Information Processing Systems*, 36:34504–34518, 2023.

[Louizos *et al.*, 2017] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

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

[Prentice, 1989] Ross L Prentice. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in medicine*, 8(4):431–440, 1989.

[Rubin, 1974] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[Singh, 2022] Rahul Singh. Generalized kernel ridge regression for long term causal inference: Treatment effects, dose responses, and counterfactual distributions. *arXiv preprint arXiv:2201.05139*, 2022.

[Wu and Fukumizu, 2022] Pengzhou Abel Wu and Kenji Fukumizu. β-intact-vae: Identifying and estimating causal effects under limited overlap. In *International Conference on Learning Representations*, 2022.

[Wu *et al.*, 2024] Peng Wu, Ziyu Shen, Feng Xie, Wang Zhongyao, Chunchen Liu, and Yan Zeng. Policy learning for balancing short-term and long-term rewards. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 53817–53846. PMLR, 21–27 Jul 2024.

[Xu *et al.*, 2023] Ziqi Xu, Debo Cheng, Jiuyong Li, Jixue Liu, Lin Liu, and Ke Wang. Disentangled representation for causal mediation analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10666–10674, 2023.

[Xu *et al.*, 2024] Ziqi Xu, Debo Cheng, Jiuyong Li, Jixue Liu, Lin Liu, and Kui Yu. Causal inference with conditional front-door adjustment and identifiable variational autoencoder. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[Yang *et al.*, 2024a] Qinwei Yang, Xueqing Liu, Yan Zeng, Ruocheng Guo, Yang Liu, and Peng Wu. Learning the optimal policy for balancing short-term and long-term rewards. *Advances in Neural Information Processing Systems*, 37:36514–36540, 2024.

[Yang *et al.*, 2024b] Zeqin Yang, Weilin Chen, Ruichu Cai, Yuguang Yan, Zhifeng Hao, Zhipeng Yu, Zhichao Zou, Zhen Peng, and Jiecheng Guo. Estimating long-term heterogeneous dose-response curve: Generalization bound leveraging optimal transport weights. *arXiv preprint arXiv:2406.19195*, 2024.

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

[Zheng *et al.*, 2025] Jiabi Zheng, Weilin Chen, Zhiyong Lin, Aqing Yang, and Zhifeng Hao. Long-term causal effects estimation across domains: an invariant surrogate representation learning approach. *International Journal of Machine Learning and Cybernetics*, pages 1–13, 2025.