CDiff: Bridging Conditional Diffusion and Causal Inference with Statistical Guarantees via Principled Distribution Approximation

Anonymous Authors[1]

Abstract

A core challenge in causal inference lies in the fundamental "missing counterfactual" problem: for each unit, only one potential outcome (treated or untreated) is observable. This renders simple comparisons of sample means between treatment groups statistically biased. We present CDiff, a theoretically principled framework for treatment effect estimation under unconfoundedness that leverages recent advances in conditional score-matching diffusion models. Building on statistical learning theory, we formalize how these models' distribution-approximating capabilities enable accurate estimation of conditional outcome distributions given covariates. Our approach facilitates simultaneous imputation of counterfactual outcomes and estimation of individual treatment effects through principled distribution matching and Monte Carlo sampling. We establish key statistical properties of the proposed estimators: consistency, asymptotic normality, and finite-sample error bounds, enabling direct construction of confidence intervals for hypothesis testing. CDiff advances causal inference methodology by integrating state-of-the-art generative models with rigorous statistical guarantees, with applications ranging from personalized medicine to policy evaluation. Experimental validation across multiple domains demonstrates the framework's empirical effectiveness.

1. Introduction

Causal inference faces a fundamental identification challenge: the impossibility of observing both potential outcomes (treated and untreated) for any individual. While existing methods address this missing data problem through assumptions like unconfoundedness, they often rely on restrictive parametric models or fail to scale to complex, high-dimensional data. We propose **CDiff**, a framework that bridges causal inference with modern generative models by leveraging conditional diffusion processes to estimate potential outcome distributions. Our work is motivated by three key observations: (1) diffusion models excel at capturing complex conditional distributions, (2) causal effect estimation inherently requires counterfactual density modeling, and (3) recent advances in score matching provide new theoretical tools for statistical inference. This work makes four advances:

- **Generative Causal Framework:** A conditional diffusion framework that jointly estimates treated/untreated outcome distributions through score matching, enabling simultaneous imputation of counterfactuals and estimation of individual treatment effects (ITE)
- **Theoretical Guarantees:** Finite-sample error bounds for counterfactual prediction, nonparametric convergence rates for potential outcome distributions, and asymptotic normality of the average treatment effect (ATE) estimator with variance characterization
- **Empirical Effectiveness:** State-of-the-art performance across synthetic and real-world benchmarks, significantly outperforming best benchmarks in both ATE and ITE estimation error
- **Generalized Causal Inference:** Extension to unstructured data modalities (images, text, graphs) through differentiable diffusion architectures, overcoming limitations of traditional covariate-based methods

CDiff advances causal methodology by integrating the distribution-approximating power of diffusion models with statistical inference theory. The derived asymptotic variance enables direct construction of confidence intervals for hypothesis testing. Our research demonstrates how deep generative models can overcome longstanding limitations in causal inference.

1.1. Diffusion and Conditional Diffusion Model

The emergence of generative modeling has revolutionized various fields, enabling the development of influential applications in text-to-image and text-to-video generation,

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

such as Stable Diffusion, MidJourney, DALL-E 2, and Firefly (Ramesh et al., 2022; Zhang et al., 2023; Rombach et al., 2022). At the core of these advancements lies the diffusion model, a powerful framework for high-fidelity sample generation. Among these, the Denoising Diffusion Probabilistic Model (DDPM) has become foundational. DDPM leverages variational inference to generate data through a multi-step sampling process. It models a reverse Markov chain to undo a forward diffusion process, wherein incremental noise is added to the original data (e.g., an image) until it becomes pure Gaussian noise (Ho et al., 2020; Sohl-Dickstein et al., 2015). Due to its ability to produce high-quality outputs, DDPM and its variants have quickly set the state of the art across image, audio, and video generation tasks (Dhariwal & Nichol, 2021; Chen et al., 2020; Ho et al., 2022).

One of the most remarkable features of diffusion models is their flexibility to incorporate various forms of input "guidance" to control the generation process. Classifier-free Conditional Diffusion Models (CDMs), for instance, combine conditional and unconditional score estimates to balance sample quality and diversity (Ho & Salimans, 2022a). This adaptability has made diffusion models a versatile tool for generation of various distributions.

The seminal work of Song et al. (2020) provides a unified perspective on various classes of diffusion models, including DDPM and score-based modeling via Langevin Dynamics, by extending these frameworks to a continuous-time formulation. In this formulation, the forward and reverse processes are elegantly described as a stochastic differential equation (SDE) and its corresponding reverse-time SDE, respectively. Subsequent advancements in statistical analysis have rigorously demonstrated that score-matching diffusion models serve as highly effective approximators of probability distributions, offering a strong theoretical foundation for their remarkable success in generative modeling. In particular, progress in conditional score approximation has introduced sample complexity bounds that adapt to the smoothness of the underlying data, achieving near-optimal estimation rates in total variation distance (Oko et al., 2024; Fu et al., 2024).

1.2. Related Work in Causal Inference

A variety of estimators for ATE under the unconfoundedness assumption have been developed in the fields of statistics and econometrics. Many of these estimators rely on nonparametric estimation of the regression function or the propensity score (Hahn, 1998; Hirano et al., 2003; Imbens et al., 2007). These methods derive the asymptotic variance for ATE and construct estimators capable of achieving the semiparametric efficiency bound.

Another significant body of research has focused on estimating ITE. These methods typically take one of two approaches: learning a separate model for each treatment group or incorporating treatment as an input feature with proper adjustments that account for the imbalance between the treated and control group distributions to mitigate the impact of selection bias. Classical approaches include tree-based methods, such as Bayesian Additive Regression Trees (BART) (Chipman et al., 2012), recursive partitioning (Athey & Imbens, 2016), and Causal Forests (Wager & Athey, 2018). Matching methods have also been widely explored, with techniques like one-to-one matching and propensity score matching being proposed to address selection bias (Dehejia & Wahba, 2002; Crump et al., 2008; Lunceford & Davidian, 2004). In recent years, deep learning has emerged as a powerful tool for ITE estimation. Johansson et al. (2016) and Shalit et al. (2017) introduced frameworks that leverage neural networks to model ITE, incorporating techniques to minimize the discrepancy between the treated and control group distributions. Finally, a multi-task learning approach was developed to estimate counterfactuals by modeling the posterior distribution of outcomes (Alaa & Van Der Schaar, 2017). All these methods share the similar objective as they primarily focus on estimating *conditional expectations* $\mathbb{E}[Y|T, X]$ rather than modeling full outcome distributions.

CDiff fundamentally advances this paradigm by learning the complete conditional density $p(y_0, y_1|x)$ through diffusion dynamics. Prior generative attempts like GANITE (Yoon et al., 2018) suffered from three limitations: (1) inefficiency in learning overlapping outcome distributions due to the innate vulnerability of GAN to mode collapse, (2) requirement for separate networks to impute missing outcomes (counterfactual generator), and estimate treatment effects (ITE generator) (3) absence of theoretical guarantees.In contrast, CDiff's conditional score matching provides stable gradient flows allowing for more efficient approximation of treatment/control outcome distributions (Dhariwal & Nichol, 2021; Song et al., 2020), eliminating need for auxiliary networks and yielding significantly better performance across benchmarks.

The recent DiffPO framework (Ma et al., 2024) shares our use of diffusion models but diverges critically in three aspects. First, architectures: DiffPO employs a single treatment-conditioned model $p(y|t, x)$, while CDiff learns separate but partially shared networks $p(y_0|x), p(y_1|x)$ enabling separate covariate-dependent conditioning. This proves essential as treatment group sizes are often imbalanced (e.g., 1:10 ratio), where shared base layers stabilize small-group learning while treatment-specific heads capture distributional shifts. Second, bias mitigation: DiffPO requires decent estimation of propensity scores $\pi(x)$ via an addition neural network to construct its orthogonal diffusion loss. Accurate learning of nuisance parameter is particularly challenging when x is high-dimensional and unstructured. Lastly, inference guarantees: Where DiffPO

provides consistency under correct specifications, CDiff establishes consistency and asymptotic normality of the ATE estimator generously—enabling valid confidence intervals construction without the restrictive Neyman orthogonality assumption. Our theoretical rigor translates to empirical improvements, as CDiff reduces ATE and ITE estimation errors by a significant margin across benchmarks and simulations.

2. Preliminaries

2.1. Problem Formulation

Let D_i denote a dummy variable such that $D_i = 1$ if individual i receives the treatment, and $D_i = 0$ otherwise. The potential outcomes are denoted as $Y_i(1)$ and $Y_i(0)$, corresponding to the treated and untreated states, respectively. The observed outcome is thus expressed as $Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0)$. For the identification of treatment effect, we assume *unconfoundedness*: $D_i \perp (Y_i(0), Y_i(1)) \mid X_i$ and overlap $0 < p(D_i = 1 \mid X_i = x) < 1, \forall x$. Below we offer some different definition of treatment effect.

Definition 2.1. The individual treatment effect (ITE) is defined as

$$\tau(x) \triangleq \mathbb{E}\left[Y_i(1) - Y_i(0) | X_i = x\right]$$

Definition 2.2. The average treatment effect (ATE) is defined as

$$\tau_{ATE} \triangleq \mathbb{E}\left[Y_i(1) - Y_i(0)\right]$$

Definition 2.3. The average treatment effect on the treated (ATT) is defined as

$$\tau_{ATT} \triangleq \mathbb{E}\left[Y_i(1) - Y_i(0) | D_i = 1\right]$$

When the true distribution that generates the sample $\mathbf{x}, \mathbf{y}$, denoted as $p_{gt}(\mathbf{x}, \mathbf{y})$, is known, the goodness of estimation of both ATE and ITE can be measured by the following two metrics:

Definition 2.4. The absolute error in average treatment effect, or ϵ_{ATE}, is defined as:

$$\epsilon_{ATE} \triangleq \left| \hat{\mathbb{E}}_{Y_i \sim \mu(\mathbf{x}_i)} \left[Y_i(1) - Y_i(0)\right] - \hat{\mathbb{E}}_{\hat{Y}_i \sim \mathcal{G}(\mathbf{x}_i)} \left[\hat{Y}_i(1) - \hat{Y}_i(0)\right] \right|$$

where $\hat{\mathbf{Y}}_i = [\hat{Y}(0), \hat{Y}(1)]_i^T$ are outcomes predicted or generated from a learned model $\mathcal{G}$ given the observation $\mathbf{x}_i$.

Definition 2.5. The expected precision in estimation of heterogeneous effects, or ϵ_{PEHE}, which according to Hill (2011) is given by:

$$\epsilon_{\text{PEHE}} = \mathbb{E}_{\mathbf{x}_i} \left[\left(\mathbb{E}_{Y_i} \left[Y_i(1) - Y_i(0)\right] - \mathbb{E}_{\hat{Y}_i \sim \mathcal{G}(\mathbf{x}_i)} \left[\hat{Y}_i(1) - \hat{Y}_i(0)\right] \right)^2 \right]$$

The statistical and econometric literature traditionally focuses on direct nonparametric estimation of potential outcome expectations. In contrast, we propose a novel paradigm centered on learning the complete conditional outcome distributions through diffusion-based density estimation. Formally, we partition the dataset into treatment groups: $\mathcal{S}_{\text{treated}} = \{i : D_i = 1\}$ and $\mathcal{S}_{\text{control}} = \{i : D_i = 0\}$. Using the respective subsets $\{(Y_i, X_i) : i \in \mathcal{S}_{\text{treated}}\}$ and $\{(Y_i, X_i) : i \in \mathcal{S}_{\text{control}}\}$, we train two conditional diffusion models: $g_{1,X}(Y) = \hat{p}(Y(1)|X)$ for the treated outcome distribution and $g_{0,X}(Y) = \hat{p}(Y(0)|X)$ for the control outcome distribution. Counterfactual imputation proceeds via Monte Carlo sampling for treated units: $\{\hat{Y}_i(0) : \hat{Y}_i(0) \sim g_{0,X_i}(Y), i \in \mathcal{S}_{\text{treated}}\}$. The average treatment effect for the treated is then estimated as:

$$\hat{\tau}_{\text{ATT}} = \frac{1}{|\mathcal{S}_{\text{treated}}|} \sum_{i \in \mathcal{S}_{\text{treated}}} \left(Y_i(1) - \hat{Y}_i(0) \right).$$

Learning the full distribution, rather than just the expectation, offers several advantages. It enables sampling arbitrary counterfactual realizations for quantile estimation and distributional effect analysis, which leads to variance reduction and robust inference. The method's statistical guarantees rely crucially on controlling the distributional estimation error $\|g_{w,X}(Y) - p(Y(w) \mid X)\|$, as will be discussed later.

2.2. Diffusion with Classifier Free Guidance as a Conditional Distribution Learner

The goal of conditional diffusion models is to generate samples from the conditional data distribution $p(Y|X)$, where P is the probability distribution function and $X \in \mathbb{R}^{d_x}$ is the conditioning information. $Y \in \mathbb{R}^d$ is the dependent variable of our interest. Diffusion model consists of a forward process and a backward process. Song et al. (2020) extends the diffusion framework into continuous time limit, drawing equivalence of forward and backward diffusion process with SDE and reverse-time SDE, respectively. Specifically, the forward process is an Ornstein-Ulhenbeck process:

$$dY_t^x = -\frac{1}{2} Y_t^x dt + dW_t \quad \text{with } Y_0^x \sim P_0(\cdot|x)$$

where W_t is a Wiener process. At any finite time t, we denote $P_t(\cdot \mid x)$ as the marginal conditional distribution. The forward process will terminate at a sufficiently large time T. For sample generation, the backward process reverses the time in forward process:

$$dY_t^{x,\leftarrow} = \left[\frac{1}{2} Y_t^{x,\leftarrow} + \nabla \log p_{T-t}(Y_t^{x,\leftarrow}|x) \right] dt + d\bar{W}_t$$

where $Y_0^{\leftarrow} \sim P_T(\cdot \mid x)$ and $\nabla \log p_{T-t}(Y_t^{x,\leftarrow}|x)$ is the so-called 'conditional score function', which is the gradient of the log probability density function of P_{T-t}. $\bar{W}_t$ is another Wiener process that is independent of W_t. Both the score function $\nabla \log p_t$ and the distribution p_T are unknown,

however, we know $\lim_{T\to\infty} Y_T^x \sim \mathcal{N}(0, I)$. We replace the unknown distribution p_T by the standard Gaussian distribution and denote $\hat{s}(y, x, t)$ as an estimator for $\nabla \log p_t(y|x)$. The estimated score $\hat{s}$ is often parameterized by a deep neural network and takes data, conditional information and time as inputs. The conditional sample generation is to simulate the following backward process

$$d\tilde{Y}_t^{x,\leftarrow} = \left[\frac{1}{2}\tilde{Y}_t^{x,\leftarrow} + \hat{s}(\tilde{Y}_t^{x,\leftarrow}, y, T-t)\right] dt + d\bar{W}_t,$$

where $\tilde{Y}_0^{x,\leftarrow} \sim \mathcal{N}(0, I)$. Denote the distribution of $\tilde{Y}_t^{\leftarrow}$ conditioned on x as $\tilde{P}_{T-t}(\cdot \mid x)$. We use deep neural networks to estimate the conditional score function and a conceptual quadratic loss is defined as: [1]

$$\underset{s\in\mathcal{S}}{arg\min} \int_{t_0}^{T} w(t)\mathbb{E}_{y_t,x}\left[\|s(y_t, x, t) - \nabla \log p_t(y_t|x)\|_2^2\right] dt,$$

where $w(t)$ is a time dependent reweighting function, for instance, $w(t) = \frac{1}{T}$. $\mathcal{S}$ is a class of deep neural networks. However, such an objective function is not computable using samples, as the score function $\nabla \log p_t$ is unknown. Instead, we minimize the following objective function,

$$\int_{t_0}^{T} w(t)\mathbb{E}_{y_0,x}\left[\mathbb{E}_{y_t}\left[\|\nabla_{y_t} \log \phi_t(y_t|y_0) - s(y_t, x, t)\|_2^2\right]\right] dt.$$

Here $\phi_t(y_t|y_0)$ denotes the Gaussian transition kernel of the forward process, so that $\nabla_{y_t} \log \phi_t$ admits an analytical form

$$\nabla_{y_t} \log \phi_t(y_t|y_0) = -\frac{y_t - \alpha(t)y_0}{h(t)}$$

where $\alpha(t) = e^{-\frac{1}{2}t}$ and $h(t) = 1 - e^{-t}$. If the true distribution P_0 is smooth enough, then there exists a ReLU neural network that can consistently estimate the conditional score function. In practice, we collect data $\{y_i, x_i\}_{i=1}^n$ and minimize the empirical risk (Ho & Salimans, 2022b)

$$\hat{s} = \underset{s\in\mathcal{S}}{arg\min}\, \hat{\mathcal{L}}(s), \quad \hat{\mathcal{L}}(s) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, x_i, ; s),$$

where $\ell(y_i, x_i; s)$ is given by

$$\int_{t_0}^{T} \frac{1}{T}\mathbb{E}_{y\sim\mathcal{N}(\alpha_t,\sigma_t^2 I)}\left[\|s(y, x, t) - \nabla_y \log \phi_t(y_t|y_0)\|_2^2\right] dt.$$

We measure the quality of the estimator $\hat{s}$ by its mean-squared deviation to the true conditional score function

$$\mathcal{R}(\hat{s}) = \int_{t_0}^{T} \frac{1}{T}\mathbb{E}_{y_t,x}\|\hat{s}(y_t, x, t) - \nabla \log p_t(y_t|x)\|_2^2 dt.$$

Under some regular conditions, this mean-squared deviation converges to zero in probability. Given the trained conditional score network $\hat{s}(y, x, t)$, we obtain the estimated conditional distribution $\hat{P}_0(\cdot|x)$. If the KL divergence of the original distribution $P(\cdot|x)$ is bounded, then the estimated distribution converges to the true distribution in the sense of total variance.

[1] t_0 is an early-stopping time to prevent the blow-up of score functions.

2.3. Theoretical Results for Distribution Approximation

Fu et al. (2024) presents a sharp statistical theory of distribution estimation using conditional diffusion models. The key to the theoretical complexity bound lies in an approximation result for the conditional score function, which is shown in Theorem 2.6.

Suppose the conditional distribution has a density $p(y|x) \in \mathcal{H}^\beta(B)$ for a Holder index β and constant B (see Appendix A.1). Moreover, there exist constants C_1, C_2 such that for all x, the density function $p(y|x) \le C_1 e^{-\frac{1}{2}C_2\|y\|_2^2}$.

Theorem 2.6 (Conditional Score Function Approximation)**.** *For the class of ReLU score network (see AppendixA.2) , if we choose the parameters, stopping time and terminal time appropriately (also see AppendixA.2), then*

$$E_{\{x_i,y_i\}_{i=1}^n}\mathcal{R}(\hat{s}) = \mathcal{O}\left(\frac{1}{t_0}n^{-\frac{\beta}{d+d_x+\beta}}(\log n)^{\max(17, d+\frac{\beta}{2}+1)}\right)$$

Given a conditional score network $\hat{s}(y, x, t)$, we establish quantifiable bounds on distribution estimation accuracy. For a given covariate vector x, let $\hat{P}(\cdot|x)$ denote the generated distribution obtained through early-stopped diffusion sampling using the estimated score $\hat{s}$. The following theorem bounds the total variation (TV) distance between the learned distribution $\hat{P}(\cdot|x)$ and the true conditional distribution $P(\cdot|x)$, providing rigorous quality guarantees:

Theorem 2.7. *Assume in addition that there exists a constant C such that $KL(P(\cdot|y)|\mathcal{N}(0, I)) \le C < \infty$ for all y. Taking the early-stopping time $t_0 = n^{-\frac{\beta}{4(d+d_x+\beta)}}$ and the terminal time $T = \frac{2\beta}{d+d_x+2\beta}\log n$, it holds that*

$$\mathbb{E}_{\{y_i,x_i\}_{i=1}^n}\mathbb{E}_x\left(TV(\hat{P}_{t_0}(\cdot|x), P(\cdot|x))\right) = \mathcal{O}\left(n^{-\frac{\beta}{4(d+d_x+\beta)}}(\log n)^{\max(9, \frac{d}{2}+\frac{\beta}{4}+1)}\right)$$

These theorems lay down critical theoretical foundations for using diffusion models in causal inference. By guaranteeing accurate approximation of conditional outcome distributions, it enables reliable estimation of the ATE and related causal quantities. In Section 3, we further prove that treatment effect estimators derived through this framework achieve both consistency and asymptotic normality.

3. Model and Estimator

3.1. Treatment Effect Estimator and Properties

To estimate the individual treatment effect $\tau(x)$, we first learn the conditional distribution $f(X|W)$ using the full

sample, obtaining $\hat{f}(X|W)$. We generate n synthetic covariates $\{X_i^G\}_{i=1}^n$ via the mapping: $X_i^G = h(W_i)$, where $h(\cdot)$ corresponds to the conditional network $\hat{f}(X|\cdot)$. In practice, this step can be omitted if n is sufficiently large to approximate the population distribution, or the support of interest $\mathcal{X}$ is fully covered by the observed sample $\{X_i\}_{i=1}^n$. This paper will focus on this case.

Core Estimation Procedure Assuming access to n observations $\{(X_i, Y_i)\}_{i=1}^n$, we generate n^G counterfactual pairs for each $X_i = x$ based on density $(Y_i^G(0)|X_i = x) \sim g_{0,X}(Y)$ and $(Y_i^G(1)|X_i = x) \sim g_{1,X}(Y)$, where $g_{0,X}(Y)$ and $g_{1,X}(Y)$ represent conditional outcome density functions that are *separately learned* via conditional diffusion models. For notation simplicity we write $(Y_i^G(0), Y_i^G(1)|X_i = x) = (Y_{i,x}^G(0), Y_{i,x}^G(1))$. The individual treatment effect estimator is then:

$$\hat{\tau}^G(x) = \frac{1}{n^G} \sum_{i=1}^{n^G} \left(Y_{i,x}^G(1) - Y_{i,x}^G(0)\right).$$

Here we propose our main theorems, which substantiate that a framework based on conditional diffusion models can be effective in estimating treatment effects. Detailed proofs can be found in Appendix B. For asymptotic analysis, we replace G by G_n to indicate that the generated pairs $(Y_{i,x}^{G_n}(0), Y_{i,x}^{G_n}(1))$ also depends on the sample size n.

Theorem 3.1. *Suppose assumptions in 2.7 holds and we obtain an estimator $\hat{\tau}^G$ for the treatment effect based on the score network in 2.6. Then the convergence rate of $\hat{\tau}^{G_n}(x)$ is*

$$\left|\hat{\tau}^{G_n}(x) - \tau(x)\right| = O((n^G)^{-\frac{1}{2}}) + O\left(n^{-\frac{\beta}{4(d+d_x+\beta)}} (\log n)^{\max(9, \frac{d}{2}+\frac{\beta}{4}+1)}\right).$$

For asymptotic analysis, let both n and n^G go to infinity and $n^G = n^{\frac{1}{3}}$, then consistency and asymptotical normality are obtained, i.e.

$$\lim_{n\to\infty} \lim_{n^G\to\infty} \hat{\tau}^{G_n}(x) = \tau(x)$$

$$\sqrt{n^G}(\hat{\tau}^{G_n}(x) - \tau(x)) \xrightarrow{d} \mathcal{N}(0, V_x)$$

where $V_x = \mathbb{E}(Y_{i,x}(1) - Y_{i,x}(0))^2 - (\mathbb{E}[Y_{i,x}(1) - Y_{i,x}(0)])^2$

So τ_{ATE} can be consistently estimated as $\hat{\tau}^{G_n} = \frac{1}{n}\sum_{x\in\{X_i\}_{i=1}^n} \hat{\tau}^{G_n}(x)$. We now propose the following theorem for statistical inference for ATE, which is crucial as it quantifies the average effectiveness of a policy or treatment on a population..

Theorem 3.2 (Asymptotic Properties of ATE Estimator)**.** *Let $V_x^{G_n} \triangleq \mathbb{E}\left[\left(Y_{i,x}^{G_n}(1) - Y_{i,x}^{G_n}(0) - \tau(x)\right)^2\right]$ denote the conditional variance of individual treatment effects. Assuming standard regularity, ATE estimator satisfies:*

(i) ***Consistency****: $\hat{\tau}^{G_n} \xrightarrow{p} \tau$ as $n, n^G \to \infty$*

(ii) ***Asymptotic Normality****:*

$$\sqrt{\frac{n}{n^G}}\left(\hat{\tau}^{G_n} - \tau\right) \Big/ \sqrt{\frac{1}{n}\sum_{x\in\{X_k\}_{k=1}^n} V_x^{G_n}} \xrightarrow{d} \mathcal{N}(0,1)$$

where the variance estimator

$$\hat{V}_x^{G_n} \triangleq \frac{1}{n^G}\sum_{i=1}^{n^G}\left(Y_{i,x}^{G_n}(1) - Y_{i,x}^{G_n}(0) - \hat{\tau}^{G_n}(x)\right)^2$$

is consistent: $\hat{V}_x^{G_n} \xrightarrow{p} V_x^{G_n}$.

Theorem 3.2 allows for construction of confidence interval for ATE.

Definition 3.3 (Confidence Intervals)**.** Let $z_{1-\alpha/2}$ denote the $(1-\alpha/2)$-quantile of the standard normal distribution. Given the variance estimator (1), define the standard error:

$$\hat{\sigma}_n \triangleq \sqrt{\frac{1}{n}\sum_{x\in\{X_k\}_{k=1}^n} V_x^{G_n}}$$

The $(1-\alpha)$-confidence interval ($\text{CI}_{1-\alpha}$) for τ is given by

$$\left[\hat{\tau}^{G_n} - z_{1-\alpha/2}\cdot\hat{\sigma}_n\sqrt{\frac{n^G}{n}},\ \hat{\tau}^{G_n} + z_{1-\alpha/2}\cdot\hat{\sigma}_n\sqrt{\frac{n^G}{n}}\right].$$

3.2. Model Architecture

We formally present **CDiff**, a conditional score-based diffusion framework that jointly learns counterfactual distributions $P(Y(0)|X)$ and $P(Y(1)|X)$ through shared representation learning. Building on conditional denoising diffusion framework, CDiff extends the score-matching objective to leverage structural similarities between treatment arms while preserving causal identifiability.

As shown in Fig. 1, CDiff employs: 1) A **shared condition block** that encodes treatment-invariant features from covariates X, and 2) **Parallel counterfactual condition heads** that *separately* learn treatment-specific score functions ϵ_0, ϵ_1 via conditionally modulated denoising. Within each block, CDiff deploys the **Condition Merge Modules** that combines covariate information with diffusion time-step embeddings:

$$h^{(l+1)} = \text{BN}(\text{GELU}(\mathbf{W}_l[h^{(l)} \oplus \mathcal{T}(X)])),$$

where $\mathcal{T}(\cdot)$ denotes cosine embeddings for continuous/binary covariates and $\oplus$ is feature concatenation.

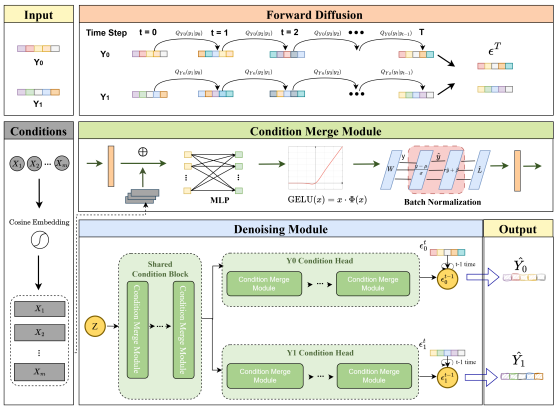

Figure 1. **A Diagram of CDiff**

For each treatment $w \in \{0, 1\}$, CDiff learns the score function $\epsilon_w(t, Y(w), X)$ via:

$$\epsilon_w^t = s_\theta(t, Y(w)^t, X),$$

where $Y(w)^t$ is the noisified outcome at diffusion step t. The shared encoder enables information transfer between treatment arms through:

$$\mathcal{L}_\theta = \sum_{w=0}^{1} \mathbb{E}_{t,Y(w),X} \left[\|\epsilon_w^t - \nabla_{Y(w)^t} \log p_t(Y(w)^t|X)\|_2^2 \right].$$

Starting from Gaussian noise $Y(w)^T \sim \mathcal{N}(0, I)$, CDiff iteratively denoises outcomes using ancestral sampling:

$$Y(w)^{t-1} = \mu_\theta(Y(w)^t, t, X) + \sigma_t \xi, \quad \xi \sim \mathcal{N}(0, I),$$

where μ_θ combines the learned score ϵ_w^t with scheduled noise levels $\{\sigma_t\}_{t=1}^T$.

4. Simulations

In the first simulation, we evaluate the performance of our model on a synthetic dataset inspired by the setup in Yoon et al. (2018). Specifically, we generate 10,000 10-dimensional feature vectors, $\mathbf{x}$, sampled from $\mathcal{N}\left(\mathbf{0}, \Sigma^2\right)$, where $\Sigma = 0.5 \times (\Omega + \Omega^T)$ and $\Omega \sim \mathcal{U}\left((-1, 1)^{10\times 10}\right)$. The outcome $\mathbf{y}$ conditional on observation $\mathbf{x}$ is given by $\mathbf{y} = \mathbf{w}_y^T \mathbf{x} + \mathbf{n}_y$, where $\mathbf{w}_y^T \sim \mathcal{U}\left((-0.1, 0.1)^{10\times 2}\right)$, and $\mathbf{n}_y \sim \mathcal{N}\left(\mathbf{0}^{2\times 1}, 0.1^2 \times I^{2\times 2}\right)$. Similar to Yoon et al. (2018)'s approach, we evaluate the robustness of our model against various levels of selection bias, by comparing with various benchmarks. We generate 10,000 treated or control samples from $\mathbf{x}_1 \sim \mathcal{N}\left(\mu_1, \Sigma^2\right)$ or $\mathbf{x}_0 \sim \mathcal{N}\left(\mu_0, \Sigma^2\right)$, respectively. For each trial, μ_0 is fixed and we vary μ_1 to generate data with different levels of selection bias, as measured by Kullback-Leibler divergences (KL divergences) between distribution of $\mathbf{x}_0$ and $\mathbf{x}_1$. As is shown in Figure 2, Our model outperforms some of the best benchmarks (GANITE and DiffPO) across levels of selection bias.

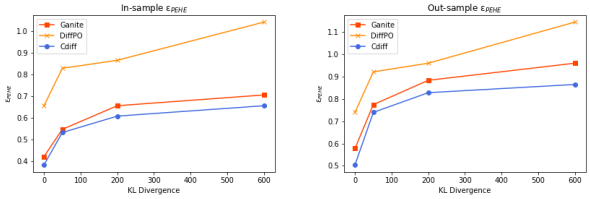

Figure 2. **Estimation of ϵ_{PEHE} under various levels of selection bias**. We selected four levels of KL-Divergence, corresponding to zero, low, medium or high selection bias. The table reports the mean and standard deviation (STD) from 100 independent trials.

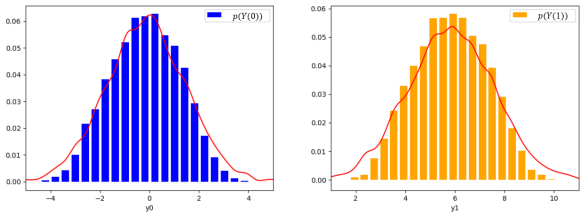

Figure 3. **CDiff-generated samples (colored histograms) versus ground truth $Y(w)$ distributions (red curves).**

Figure 3 showcases CDiff's ability to recover ground-truth potential outcome distributions under high selection bias, demonstrating almost complete alignment between generated distributions and ground truth. This is true for both treatment arms and control arms. The ability for CDiff to approximates true conditional density ($p(Y(0)|X)$ and $p(Y(1)|X)$) is demonstrated in Appendix D.

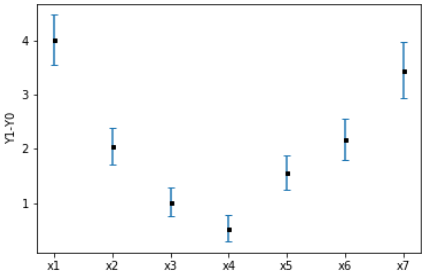

Figure 4. **True ITE and predicted confidence intervals.** Blue bars show 95% CIs for different x values (black dots: true $\tau(x)$).

Finally, we validate CDiff's asymptotic guarantees via sampling across the covariate space $\mathcal{X}$, selecting seven x with ground truth ITE $\tau(x) \in [0.5, 4]$. Following Def. 3.3, we construct 95% confidence intervals ($\text{CI}_{95\%}$) for each $\tau(x)$ through 10^2 Monte Carlo trials. Figure 4 demonstrates statistically tight coverage as 100% of CIs contain $\tau(x)$ (mean interval width 0.36 ± 0.09).

5. Experiments

5.1. Datasets

To address the challenge that counterfactual outcomes are unobservable in real-world settings, we followed the standard approach used by many previous studies (Chipman et al., 2012; Shalit et al., 2017; Alaa & Van Der Schaar, 2017; Yoon et al., 2018), evaluating our method on the semi-synthetic IHDP dataset and the real-world Jobs dataset.

IHDP The Infant Health and Development Program (IHDP) dataset (Hill, 2011) is widely used for the ITE estimation. It consists of 747 instances (139 treated, 608 controls) with 25-dimensional covariates. The outcome values were *synthesized* using the NPCI package under setting 'A', as implemented by Dorie (2016).

Jobs The Jobs dataset, developed by LaLonde (1986), is a widely used *real-life* benchmark for causal effect estimation. It features a randomized study conducted by the National Supported Work program, with 297 treated and 425 control samples. Observational data (the PSID group, 2490 controls) was later added by Smith & Todd (2005). The treatment corresponds to receiving job training, while the observed outcomes are income and employment status.

Since counterfactual outcomes are unobservable, two evaluation metrics are commonly used for this dataset: the error of the average treatment effect on the treated (ϵ_{ATT}) and the policy risk ($\hat{R}_{pol}(\pi_f)$). According to Shalit et al. (2017) and Smith & Todd (2005), the true τ_{ATT} is defined as:

$$\tau_{ATT} = \frac{1}{|T_1 \cap E|} \sum_{\mathbf{x}_i \in T_1 \cap E} Y_1(\mathbf{x}_i) - \frac{1}{|T_0 \cap E|} \sum_{\mathbf{x}_i \in T_0 \cap E} Y_0(\mathbf{x}_i),$$

where T_1 and T_0 are the treated and control groups, respectively, and E is the randomized controlled trial subset. The empirical estimate of ϵ_{ATT} is given by:

$$\hat{\epsilon}_{ATT} = |\tau_{ATT} - \frac{1}{|T_1 \cap E|} \sum_{\mathbf{x}_i \in T_1 \cap E} \hat{Y}_1(\mathbf{x}_i) - \hat{Y}_0(\mathbf{x}_i)|.$$

The policy risk, another metrics that evaluates treatment assignment policies, is given by:

$$\hat{R}_{pol}(\pi_f) = 1 - (\mathbb{E}[Y_1|\pi_f(\mathbf{x}) = 1, t = 1] \cdot p(\pi_f = 1)) + \mathbb{E}[Y_0|\pi_f(\mathbf{x}) = 0, t = 0] \cdot p(\pi_f = 0).$$

5.2. Results

On both datasets, we evaluate the in-sample and out-of-sample estimations of τ_{ATE} and τ_{ATT} from our model against several benchmark models. The baselines include Balancing Linear Regression (BLR) and Balancing Neural Network (BNN) (Johansson et al., 2016), k-Nearest Neighbors (k-NN) (Crump et al., 2008), Bayesian Additive Regression Trees (BART) (Chipman et al., 2012), Causal Forests (C Forest) (Wager & Athey, 2018), Treatment-Agnostic Representation Network (TARNET), Counterfactual Regression with Wasserstein Distance (CFR_{Wass}) (Shalit et al., 2017), Multi-task Gaussian Process (CMGP) (Alaa & Van Der Schaar, 2017), GANITE (Yoon et al., 2018), and DiffPO (Ma et al., 2024).

Dataset	**IHDP** ($\hat{\epsilon}_{ATE}$)		**Jobs** ($\hat{\epsilon}_{ATT}$)	
Methods	In-sample	Out-sample	In-sample	Out-sample
CDiff	**.09 ± .02**	**.12 ± .02**	**.01 ± .01**	**.04 ± .01**
DiffPO	.47 ± .04	.59 ± .05	.05 ± .02	.09 ± .04
GANITE	.43 ± .05	.49 ± .05	**.01 ± .01**	.06 ± .03
BLR	.72 ± .04	.93 ± .05	**.01 ± .01**	.08 ± .03
BNN	.37 ± .03	.42 ± .03	.04 ± .01	.09 ± .04
k-NN	.14 ± .01	.90 ± .05	.21 ± .01	.13 ± .05
BART	.23 ± .01	.34 ± .02	.02 ± .00	.08 ± .03
C Forest	.18 ± .01	.40 ± .03	.03 ± .01	.07 ± .03
TARNET	.26 ± .01	.28 ± .01	.05 ± .02	.11 ± .04
CFR_{Wass}	.25 ± .01	.27 ± .01	.04 ± .01	.09 ± .03
CMGP	.11 ± .10	.13 ± .12	.06 ± .06	.09 ± .07

Table 1. **Estimation errors of τ_{ATE} and τ_{ATT} on IHDP and Jobs datasets**, featuring comparison between our model and current SOTAs. Mean and STD values are computed over multiple independent runs. The best results are highlighted in **bold**.

Table 1 demonstrates CDiff's superiority in estimating population-level treatment effects (τ_{ATE}, τ_{ATT}) on IHDP and Jobs datasets. Our framework achieves SOTA out-of-sample performance over all benchmarks, across both datasets (the difference is statistically significant $p < 0.05$ except for the highly volatile CMGP). While CDiff shows substantial in-sample improvements on IHDP, its Jobs in-sample performance is on par with the best benchmark models (GANITE and BLR).

We further evaluate CDiff using two alternative metrics: the expected precision in heterogeneous effect estimation (ϵ_{PEHE}) and policy risk ($R_{\text{pol}}(\pi_f)$). As shown in Table 2, CDiff demonstrates superior policy risk minimization, achieving a 30% reduction in out-of-sample $R_{\text{pol}}(\pi_f)$ compared to the second-best benchmark, GANITE ($p < 0.05$). While CDiff significantly outperforms most benchmarks on ϵ_{PEHE} (e.g., GANITE, BART, and Causal Forest; $p < 0.05$), it exhibits modest gaps to CMGP, TARNET, and CFR_{Wass}. This reflects CDiff's implicit tradeoff: a slight performance degradation in extreme quantiles for over performance at population-level (ATE). ϵ_{PEHE}, as sum of point-wise squared-errors, inherently favors point-estimate methods like CMGP and TARNET that prioritize local accuracy.

In summary, across diverse evaluation scenarios—spanning simulation, semi-synthetic, and real-world datasets with varying sample sizes (1,000–10,000 units), selection bias,

Dataset	**IHDP** ($\sqrt{\hat{\epsilon}_{PEHE}}$)		**Jobs** ($\hat{R}_{pol}(\pi_f)$)	
Methods	In-sample	Out-sample	In-sample	Out-sample
CDiff	1.7 ± .2	1.8 ± .3	**.08± .01**	**.11± .01**
DiffPO	2.8 ± .2	3.1 ± .4	.19 ± .02	.22 ± .03
GANITE	1.9 ± .4	2.4 ± .4	.13 ± .01	.14 ± .01
BLR	5.8 ± .3	5.8 ± .3	.22 ± .01	.25 ± .02
BNN	2.2 ± .1	2.1 ± .1	.20 ± .01	.24 ± .02
k-NN	2.1 ± .1	4.1 ± .2	.22 ± .00	.26 ± .02
BART	2.1 ± .1	2.3 ± .1	.23 ± .00	.25 ± .02
C Forest	3.8 ± .2	3.8 ± .2	.19 ± .00	.20 ± .02
TARNET	.88 ± .02	.95 ± .02	.17 ± .01	.21 ± .01
CFR_{Wass}	**.71 ± .02**	**.76 ± .02**	.17 ± .02	.21 ± .01
CMGP	.65 ± .44	.77 ± .11	.17 ± .03	.24 ± .05

Table 2. **Estimation of ϵ_{PEHE} or $R_{pol}(\pi)$ on IHDP and Jobs datasets**, featuring comparison between our model and current SOTAs. Mean and STD values are computed over multiple independent runs. The best results are highlighted in **bold**.

and treatment group imbalances (1:5 to 1:10)—CDiff achieves statistically significant superiority in 14 of 16 cases (88%). This includes consistent first-rank performance on both population-level (ATE/ATT) and individual-level (ITE) metrics under in-sample and out-of-sample regimes. The framework balances between asymptotic quality for ATE and high precision in ITE, positioning itself as a versatile tool for both large-scale policy evaluation and individual intervention optimization.

5.3. Discussion

Ablation Study To understand the contributions of individual components in our framework, we perform ablation studies on IHDP by systematically modifying key elements of the model. Additional ablation results are presented in Appendix D.

Noise Formulation Table 3 compares three approaches: 1) CDiff: Joint prediction of ϵ_0^t, ϵ_1^t via parallel heads; 2) $\Delta noise$: Predicts ϵ_0^t and residual $\Delta\epsilon^t = \epsilon_1^t - \epsilon_0^t$; 3) $Y_1 - Y_0$: Separate diffusion of $Y_0/(Y_1 - Y_0)$ with scores ϵ_0^t, ϵ_{1-0}^t. CDiff significantly over-performs the other two types of residual-like formulations, demonstrating that joint modeling of potential outcomes improves distribution matching by leveraging covariate-outcome dependencies.

Architecture Configuration Table 4 evaluates layer allocation between shared and treatment-specific modules, where $A+B$ denotes A stacks of Condition Merge Modules in Shared Condition Block and B stacks in Y_0/Y_1 Condition Heads. CDiff's balanced configuration (2+2) achieves optimal performance among all configurations. In particular, the 0+4 variant (fully independent heads) exhibits the worst performance, underscoring the benefits of shared representations in capturing inherent counterfactual dependencies for the same covariates.

Dataset	**IHDP** ($\hat{\epsilon}_{ATE}$)		**IHDP** ($\sqrt{\hat{\epsilon}_{PEHE}}$)	
Methods	In-sample	Out-sample	In-sample	Out-sample
CDiff	**.09 ± .02**	**.12 ± .02**	**1.7 ± .2**	**1.8 ± .3**
$\Delta noise$	.21 ± .05	.28 ± .07	2.5 ± .6	2.9 ± .8
$Y_1 - Y_0$	.12 ± .03	.16 ± .04	2.1 ± .3	2.6 ± .5

Table 3. **Ablation studies on IHDP of different learning target for two heads.** . Mean and STD values are computed over multiple independent runs.

Dataset	**IHDP** ($\hat{\epsilon}_{ATE}$)		**IHDP** ($\sqrt{\hat{\epsilon}_{PEHE}}$)	
Methods	In-sample	Out-sample	In-sample	Out-sample
$3+1$	.12 ± .02	.16 ± .02	1.9 ± .2	2.0 ± .3
$2+2$ (CDiff)	**.09 ± .02**	**.12 ± .02**	**1.7 ± .2**	**1.8 ± .3**
$1+3$	.17 ± .03	.22 ± .04	2.1 ± .3	2.3 ± .4
$0+4$	.25 ± .07	.33 ± .11	2.8 ± .4	3.1 ± .6

Table 4. **Ablation studies on IHDP of different shared module stacks (Shared Stacks + Unshared Stacks)**. Mean and STD values are computed over multiple independent runs.

Future Directions Several future extensions merit investigation: (1) Generalization to $K > 2$ treatments through parallel denoising heads (see Appendix), preserving parameter efficiency via shared base layers while maintaining the original loss structure. (2) Adaptation to multi-modal data (text, graphs, images) through modality-specific encoders, building on works by Veitch et al. (2020) and Jerzak et al. (2022). (3) Systematic evaluation of estimation robustness under non-Euclidean data geometries. (4) Development of model compression techniques (distillation/quantization) to enhance efficiency without losing statistical rigor.

6. Conclusion

We present CDiff, a causal inference framework that synergizes diffusion-based conditional density estimation with statistics theory. By reformulating counterfactual prediction as a score-matching problem, CDiff achieves: (1) consistent and asymptotically normal ATE estimates under standard regularity conditions, (2) finite-sample error bounds for both ATE and ITE predictions, and (3) SOTA empirical performance across benchmarks with statistical significance. The framework's robustness to strong selection bias, as demonstrated through empirical tests, positions it as a versatile tool for high-stakes applications ranging from personalized immunotherapy design to algorithmic fairness auditing. CDiff advances causal machine learning by demonstrating how modern generative models can transcend their traditional descriptive role to enable principled counterfactual reasoning.

Statement of Impacts

CDiff represents a paradigm shift in causal inference, unifying deep generative modeling with statistical theory to address the core challenge of counterfactual estimation. By leveraging conditional score-matching diffusion, CDiff achieves unprecedented accuracy in learning potential outcome distributions, enabling robust estimation of both population-level effects (ATE/ATT) and individualized treatment responses (ITE). Its SOTA performance across synthetic and real-world benchmarks—including high-dimensional healthcare and policy evaluation datasets—demonstrates practical value for decision-making under uncertainty. The framework's theoretical guarantees (consistency, asymptotic normality) directly enable confidence interval construction, bridging the gap between machine learning flexibility and statistical rigor required for high-stakes applications.

References

Alaa, A. M. and Van Der Schaar, M. Bayesian inference of individualized treatment effects using multi-task gaussian processes. *Advances in neural information processing systems*, 30, 2017.

Athey, S. and Imbens, G. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Chen, N., Zhang, Y., Zen, H., Weiss, R. J., Norouzi, M., and Chan, W. Wavegrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713*, 2020.

Chipman, H. A., George, E. I., and McCulloch, R. E. Bart: Bayesian additive regression trees. *Annals of Applied Statistics*, 6(1):266–298, 2012.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics*, 90(3):389–405, 2008.

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

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Dorie, V. Non-parametrics for causal inference, 2016.

Fu, H., Yang, Z., Wang, M., and Chen, M. Unveil conditional diffusion models with classifier-free guidance: A sharp statistical theory. *arXiv preprint arXiv:2403.11968*, 2024.

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

Hahn, J. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331, 1998.

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

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

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022a.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022b.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D. J. Video diffusion models. *arXiv:2204.03458*, 2022.

Imbens, G. W., Newey, W., and Ridder, G. Mean squared error calculations for average treatment effects. *Working Paper*, 2007.

Jerzak, C. T., Johansson, F., and Daoud, A. Image-based treatment effect heterogeneity. *arXiv preprint arXiv:2206.06417*, 2022.

Johansson, F., Shalit, U., and Sontag, D. Learning representations for counterfactual inference. In *International conference on machine learning*, pp. 3020–3029. PMLR, 2016.

LaLonde, R. J. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pp. 604–620, 1986.

Lunceford, J. K. and Davidian, M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, 23(19):2937–2960, 2004.

Ma, Y., Melnychuk, V., Schweisthal, J., and Feuerriegel, S. Diffpo: A causal diffusion model for learning distributions of potential outcomes. *arXiv preprint arXiv:2410.08924*, 2024.

Oko, K., Akiyama, S., and Suzuki, T. Diffusion models are minimax optimal distribution estimators. *Working Paper*, 2024.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pp. 3076–3085. PMLR, 2017.

Smith, J. A. and Todd, P. E. Does matching overcome lalonde's critique of nonexperimental estimators? *Journal of econometrics*, 125(1-2):305–353, 2005.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.

Veitch, V., Sridhar, D., and Blei, D. Adapting text embeddings for causal inference. In *Conference on Uncertainty in Artificial Intelligence*, pp. 919–928. PMLR, 2020.

Wager, S. and Athey, S. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Yoon, J., Jordon, J., and Van Der Schaar, M. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International conference on learning representations*, 2018.

Zhang, C., Zhang, C., Zhang, M., and Kweon, I. S. Text-to-image diffusion models in generative ai: A survey. *arXiv preprint arXiv:2303.07909*, 2023.

A. Definitions and Assumptions

Definition A.1 (Hölder norm)**.** The Hölder norm are widely used as a measure of smoothness (Györfi et al., 2006). Our study focuses a family of density distributions that lie in a Hölder ball. Let $\beta = s + \gamma > 0$ be a degree of smoothness, where $s = \lfloor \beta \rfloor$ is an integer and $\gamma \in [0, 1)$. For a function $f : \mathbb{R}^d \to \mathbb{R}$, its Holder norm is defined as

$$\|f\|_{\mathcal{H}^\beta} := \max_{s:\|s\|_1<s} \sup_x |\partial^s f(x)| + \max_{s:\|s\|_1=s} \sup_{x\neq z} \frac{|\partial^s f(x) - \partial^s f(z)|}{\|x - z\|_\infty^\gamma}$$

where s is a multi-index. We say a function f is β-Holder if $\|f\|_{\mathcal{H}^\beta} < \infty$. We define a Holder ball of radius $B > 0$ as

$$\mathcal{H}^\beta(B) = \left\{ f : \mathbb{R}^d \to \mathbb{R} \mid \|f\|_{\mathcal{H}^\beta} < B \right\}$$

Definition A.2 (Class of ReLU Score Network and Parameters)**.** The class of ReLU score network is defined as

$$\mathcal{F}(M_t, W, \kappa, L, K) := \{s(y, x, t) = (A_L \sigma(\cdot) + b_L) \circ \cdots \circ (A_1[y', x', t]' + b_1)\}$$

where $A_i \in \mathbb{R}^{d_i \times d_{i+1}}$, $b_i \in \mathbb{R}^{d_i+1}$, $\max d_i \leq W$, $\sup_{y,x} \|s(y, x, t)\|_\infty \leq M_t$, $\max \|A_i\|_\infty \vee \|b_i\|_\infty \leq \kappa$ and $\sum_{i=1}^L (\|A_i\|_0 + \|b_i\|_0) \leq K$. For the conditional approximation, we choose the score network with parameters satisfying

$$M_t = \mathcal{O}(\frac{\sqrt{\log N}}{\sigma_t^2}), W = \mathcal{O}(N \log^7 N)$$

$$\kappa = e^{\mathcal{O}(\log^4 N)}, L = \mathcal{O}(\log^4 N), K = \mathcal{O}(N \log^9 N)$$

where the network size parameter $N = n^{\frac{d+d_x}{d+d_x+\beta}}$; for constants $C_\sigma, C_\alpha > 0$, we take the early stopping time $t_0 = N^{-C_\sigma} < 1$ and the terminal time $T = \mathcal{O}(\log n)$.

B. Mathematical Proofs

Proof B.1 (Proof for Theorem 3.1)**.**

$$\sqrt{n^G}(\hat{\tau}^{G_n}(x) - \tau(x)) = \sqrt{n^G}(\hat{\tau}^{G_n}(x) - \hat{\tau}(x)) + \sqrt{n^G}(\hat{\tau}(x) - \tau(x))$$

$$\begin{aligned}
\lim_{n^G\to+\infty} \sqrt{n^G}(\hat{\tau}^{G_n}(x) - \hat{\tau}(x)) &= \lim_{n^G\to+\infty} \sqrt{n^G} \left(\frac{1}{n^G} \sum_{i=1}^{n^G} (Y_{i,x}^{G_n}(1) - Y_{i,x}^{G_n}(0)) - \frac{1}{n^G} \sum_{i=1}^{n^G} (Y_{i,x}(1) - Y_{i,x}(0)) \right) \\
&= \lim_{n^G\to+\infty} \sqrt{n^G} \left(\frac{1}{n^G} \sum_{i=1}^{n^G} (Y_{i,x}^{G_n}(1) - Y_{i,x}(1)) + \frac{1}{n^G} \sum_{i=1}^{n^G} (Y_{i,x}(0) - Y_{i,x}^{G_n}(0)) \right) \\
&= \lim_{n^G\to+\infty} \sqrt{n^G} \left(\frac{1}{n^G} \sum_{i=1}^{n^G} \mathcal{O}\left(n^{-\frac{\beta}{4(d+d_x+\beta)}} (\log n)^{\max(9, \frac{d}{2}+\frac{\beta}{4}+1)} \right) \right) \\
&= \lim_{n^G\to+\infty} (n^G)^{\frac{1}{2}} \mathcal{O}\left(n^{G^{-\frac{3}{4+\frac{d+d_x}{\beta}}}} (\log n^G)^{3\max(9, \frac{d}{2}+\frac{\beta}{4}+1)} \right) \\
&= o_p(1)
\end{aligned}$$

It shows that the first term $\sqrt{n^G}(\hat{\tau}^{G_n}(x) - \hat{\tau}(x))$ converges to 0 in probability, therefore it has no effect on the asymptotic distribution of $\hat{\tau}^{G_n}(x)$. For the second term $\sqrt{n^G}(\hat{\tau}(x) - \tau(x))$, by Central Limit Theorem, we have $\sqrt{n^G}(\hat{\tau}(x) - \tau(x)) \xrightarrow{d} \mathcal{N}(0, Var(Y_{i,x}(1) - Y_{i,x}(0))$. Finally, by Slutsky's theorem,

$$\sqrt{n^G}(\hat{\tau}^{G_n}(x) - \tau(x)) \xrightarrow{d} \mathcal{N}(0, Var(Y_{i,x}(1) - Y_{i,x}(0)))$$

■

Proof B.2 (Proof for Theorem 3.2)**.** Define a triangular array whose nth row has nn^G elements. The first n^G elements consist of a random sample $\{Y_{i,X_1}^{G_n}(1) - Y_{i,X_1}^{G_n}(0) - \tau^{G_n}(X_1)\}_{i=1}^{n^G}$, and the kth n^G elements consist of a random sample $\{Y_{i,X_k}^{G_n}(1) - Y_{i,X_k}^{G_n}(0) - \tau^{G_n}(X_k)\}_{i=1}^{n^G}$. Each element M_{nj} in nth row has mean $\mathbb{E}[M_{nj}] = 0$. Therefore, $Var(M_{nj}) = \mathbb{E}[M_{nj}^2]$. Let the sum of variance in nth row be $s_n^2 = \sum_{j=1}^{nn^G} \mathbb{E}[M_{nj}^2] = n^G \sum_{x\in\{X_k\}_{k=1}^n} \mathbb{E}[(Y_{i,x}^{G_n}(1) - Y_{i,x}^{G_n}(0) - \tau^{G_n}(x))^2]$. By Lindeberg Central Limit Theorem,

$$\frac{1}{s_n} \sum_{x\in\{X_k\}_{k=1}^n} \sum_{i=1}^{n^G} \left(Y_{i,x}^{G_n}(1) - Y_{i,x}^{G_n}(0) - \tau^{G_n}(x) \right) \xrightarrow{d} \mathcal{N}(0,1)$$

Furthermore,

$$\begin{aligned}
\frac{1}{s_n} \sum_{x\in\{X_k\}_{k=1}^n} \sum_{i=1}^{n^G} \left(Y_{i,x}^{G_n}(1) - Y_{i,x}^{G_n}(0) - \tau^{G_n}(x) \right) &= \frac{1}{\frac{1}{nn^G} s_n} \frac{1}{nn^G} \sum_{x\in\{X_k\}_{k=1}^n} \sum_{i=1}^{n^G} \left(Y_{i,x}^{G_n}(1) - Y_{i,x}^{G_n}(0) - \tau^{G_n}(x) \right) \\
&= \frac{1}{\frac{1}{n}\sqrt{n^G \sum_{x\in\{X_k\}_{k=1}^n} V_x^{G_n}}} \left(\left(\hat{\tau}^{G_n} - \tau \right) + \left(\tau - \frac{1}{n} \sum_{x\in\{X_k\}_{k=1}^n} \tau^{G_n}(x) \right) \right) \\
&= \frac{1}{\sqrt{\frac{n^G}{n}}\sqrt{\frac{1}{n} \sum_{x\in\{X_k\}_{k=1}^n} V_x^{G_n}}} \left(\left(\hat{\tau}^{G_n} - \tau \right) + \left(\tau - \frac{1}{n} \sum_{x\in\{X_k\}_{k=1}^n} \tau^{G_n}(x) \right) \right)
\end{aligned}$$

By Law of Large Number, the second term satisfies $\lim_{n\to+\infty} \frac{1}{n} \sum_{x\in\{X_i\}_{i=1}^n} \tau^{G_n}(x) = \tau$. Therefore, $\sqrt{nn^G}(\tau - \frac{1}{n} \sum_{x\in\{X_k\}_{k=1}^n} \tau^{G_n}(x)) = O_p(1)$. Therefore, $\sqrt{\frac{n}{n^G}}(\tau - \frac{1}{n} \sum_{x\in\{X_k\}_{k=1}^n} \tau^{G_n}(x)) = o_p(1)$. Therefore, the second term has no effect on the asymptotic distribution, which implies

$$\sqrt{\frac{n}{n^G}} \frac{1}{\sqrt{\frac{1}{n} \sum_{x\in\{X_k\}_{k=1}^n} V_x^{G_n}}} (\hat{\tau}^{G_n} - \tau) \xrightarrow{d} \mathcal{N}(0,1)$$

■

C. Training Procedures here

C.1. Pseudocodes

Here we present the pseudocode for the conditional score-matching diffusion framework.

Algorithm 1 Training

repeat
  $\mathbf{Y_0^0} \sim q(\mathbf{Y_0}), \mathbf{Y_1^0} \sim q(\mathbf{Y_1})$
  $t \sim \text{Uniform}(\{1, \dots, T\})$
  $\boldsymbol{\epsilon}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \boldsymbol{\epsilon}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  Do gradient descent $\nabla_\theta \left(\left\| \boldsymbol{\epsilon}_0 - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{Y_0^0} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0, t, X) \right\|^2 + \left\| \boldsymbol{\epsilon}_1 - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{Y_1^0} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_1, t, X) \right\|^2 \right)$
until converged

Algorithm 2 Sampling

$\tau^T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
$\mathbf{Y_0^t} = \tau^T, \mathbf{Y_1^t} = \tau^T$
for $t = T$ to 1 **do**
  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\mathbf{Y_0^{t-1}} = \frac{1}{\sqrt{\alpha_t}}(\mathbf{Y_0^t} - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{Y_0^t}, t, X)) + \sigma_t \mathbf{z}$
  $\mathbf{Y_1^{t-1}} = \frac{1}{\sqrt{\alpha_t}}(\mathbf{Y_1^t} - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{Y_1^t}, t, X)) + \sigma_t \mathbf{z}$
end for
return $\mathbf{Y_0^0}, \mathbf{Y_1^0}$

C.2. Hyperparameter Settings for Training CDiff

Blocks	Sets of Hyper-parameters
Initialization	Xavier Initialization for Weight matrix, Zero initialization for bias vector.
Optimization	AdamW
Batch size	1024
Depth of layers	2 + 2 (See Sec 5.3 for details)
Hidden state dimension	512
α, β	$\{0.1, 0.01\}$

Table 5. **Hyperparameters of CDiff**

C.3. Implemental Details

We train all models with AdamW optimizer with initialization learning rate of 1e-4 and weight decay of 1e-2. The cosine annealing learning rate warmup is adopted to stabilize the training process. Following common practice in the Diffusion model, we maintain an exponential moving average (EMA) of eights over training with a decay of .999. We use 1000 time steps during training while 500 time steps during inference. We train the model for about 100,000 steps. The hidden dim of Denoising Module is set to 512.

D. Additional Simulation and Experiment Results can go here

D.1. Conditional Density Simulation

We evaluate CDiff's ability to recover $p(Y(w)|X = x)$ under extreme selection bias ($D_{\text{KL}} = 600$), focusing on a fixed covariate value x_0. Using 10^3 generated samples per treatment arm, CDiff generates samples that are remarkably close to ground truth distributions (Figure 5)), confirming accurate counterfactual density matching despite limited overlap.

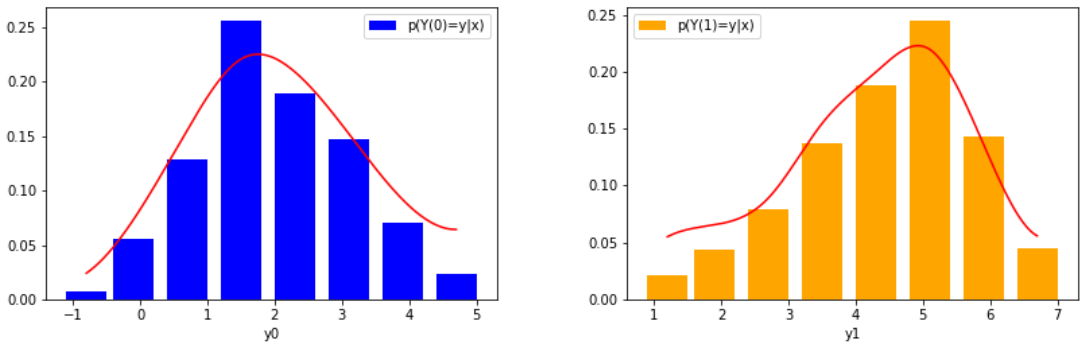

Figure 5. **Conditional density estimation under selection bias.** CDiff-generated samples (colored histograms) versus ground truth $p(Y(w)|X = x_0)$ (red curves).

Dataset	**IHDP** ($\hat{\epsilon}_{ATE}$)		**IHDP** ($\sqrt{\hat{\epsilon}_{PEHE}}$)	
Methods	In-sample	Out-sample	In-sample	Out-sample
CDiff	**.09 ± .02**	**.12 ± .02**	**1.7 ± .2**	**1.8 ± .3**
Scalar Only	.11 ± .03	.14 ± .04	1.8 ± .2	1.9 ± .3
Binary Only	.14 ± .03	.19 ± .05	2.2 ± .3	2.5 ± .4
without CE	.17 ± .03	.25 ± .05	2.6 ± .3	2.8 ± .4

Table 6. **Ablation studies on IHDP of w/o using cosine embedding (CE) for conditions**. Mean and STD values are computed over multiple independent runs.

D.2. Additional Ablation Studies

Table 6 quantifies cosine embedding (CE)'s impact on IHDP's mixed data types (scalar/binary covariates). Ablating CE for either types increases prediction errors, where joint removal yielding further performance degradation. The greater sensitivity to scalar covariates stems from their wider dynamic range—CE stabilizes learning by normalizing heterogeneous input scales, whereas binary features inherently exhibit limited variance.

E. Alternative Version of Models

Multi-Treatment Extension Figure 6 extends CDiff to K treatments via parallel score heads $\{\epsilon_w^t\}_{w=1}^K$ with shared covariate encoding. This preserves sample efficiency while scaling complexity linearly with K.

Ablated Architecture Figure 7 evaluates the $0+4$ configuration (no shared layers between treatment arms), exhibiting 72% higher $\sqrt{\epsilon_{\text{PEHE}}}$ than CDiff. This hints the potential cost of ignoring counterfactual information flow - a critical design insight for causal architectures.

F. Discussion of Limitations

While CDiff excels in complex settings with nonlinear responses and high dimensional inputs, its computational intensity makes classical low-dimensional linear regimes (e.g., randomized trials with small sample sizes) better served by simpler estimators. This trade-off reflects CDiff's design prioritization: sacrificing lightweight computation for unmatched fidelity in modern, data-rich environments. Future work will extend CDiff to multi-modal data (e.g., medical imaging, sensor streams) while developing distillation/quantization techniques to enhance efficiency without sacrificing theoretical guarantees.

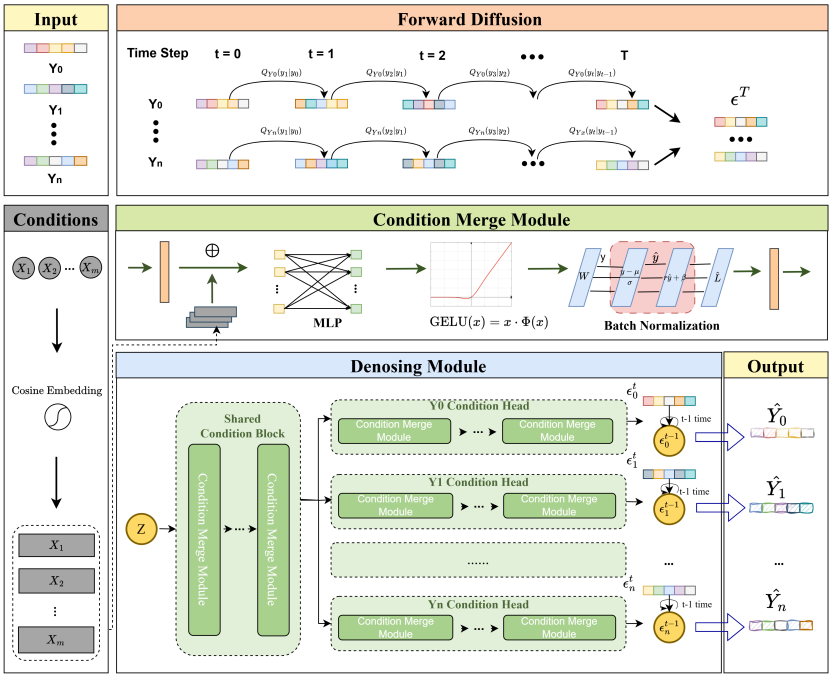

Figure 6. **An Multi-Head Version of CDiff**

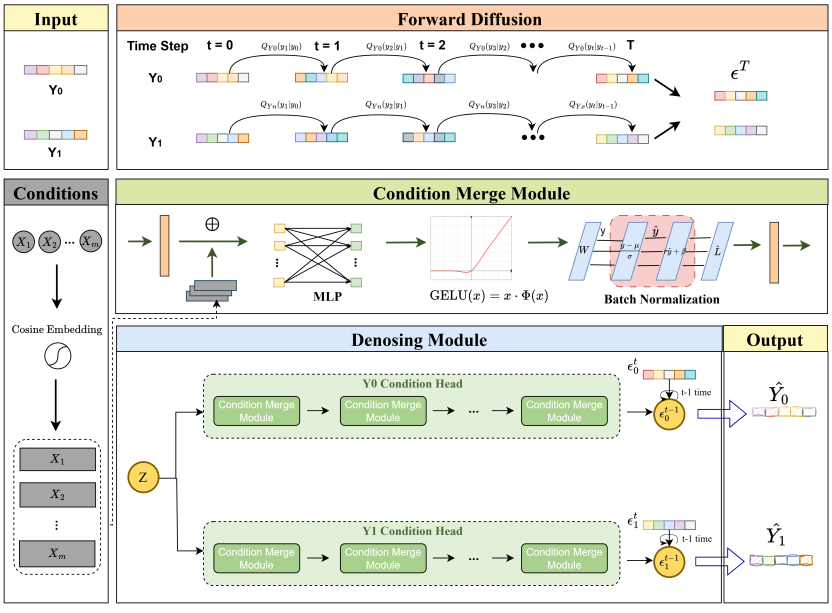

Figure 7. **An Alternative Verion of Two-Head CDiff**