Extended Background on Treatment Effect Estimation

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, G. W. Imbens, and Ridder (2003), and G. W. Imbens, Newey, and Ridder (2007). These methods use propensity score as an approach for dimension reduction 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, George, and McCulloch (2012), recursive partitioning Athey and G. Imbens (2016), and Causal Forests Wager and 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 and Wahba (2002), Crump et al. (2008), and Lunceford and Davidian (2004). In recent years, deep learning has emerged as a powerful tool for ITE estimation. F. Johansson, Shalit, and Sontag (2016) and Shalit, F. D. Johansson, and Sontag (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 and 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.

CausalDiff fundamentally advances this paradigm by learning the complete conditional density $p(y_w|x), w \in \{0, 1\}$ through diffusion dynamics. Prior generative attempts like GANITE Yoon, Jordon, and Van Der Schaar (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, CausalDiff's conditional score matching provides stable gradient flows allowing for more efficient approximation of treatment/control outcome distributions Dhariwal and Nichol (2021) and 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 four aspects. First, architectures: DiffPO employs a single treatment-conditioned model $p(y|t, x)$, while CausalDiff 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. G. W. Imbens, Newey, and Ridder (2007) shows that the estimation of the propensity score is not necessary to obtain an asymptotically efficient estimator for treatment effects. Our method directly learns the two potential distributions and avoids estimating the propensity score, which is a nuisance function

that might cause bias for the estimation of ATE. Lastly, inference guarantees: CausalDiff establishes consistency and asymptotic normality of the ATE estimator generously—enabling valid confidence intervals construction. This is important for applications in medical science and policy evaluation because always we want to conclude whether a medicine or a policy has effects instead of merely obtaining an magnitude of the estimator. Fourth, unstructured data performance. Our theoretical rigor translates to empirical improvements, as CausalDiff reduces ATE and ITE estimation errors by a significant margin across benchmarks and simulations, especially in high-dimensional and unstructured data environment.

Proofs

Definition B.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 B.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 $T = \mathcal{O}(\log n)$.

Assumption B.3. Let C and C_2 be two positive constants and $f \in \mathcal{H}^\beta(B)$. We assume that $f(y, x) \geq C, \forall(y, x)$ and the conditional density function $p(y|x) = e^{-C_2\|x\|_2^2} \cdot f(y, x)$. This assumption imposes upper and lower bounds for $f(y, x)$ and enables a faster approximation.

Proof B.4 (Proof for Theorem2.2)**.** To begin with, we give the key steps for proof of conditional score approximation, following Fu et al. (2024). The idea is to approximate $\nabla p_t(y|x)$ and $p_t(y|x)$ separately. However, even though the data generating process has Holder regularity conditions, there are two challenges. First, support of x is unbounded and it is difficult to derive a uniform approximation of $p_t(y|x)$. Second, the density function $p_t(y|x)$ can be arbitrarily small, which implies that its inverse can be quite large. Existing results

address the ceveats by proper truncation on domain x and the value of $p(y|x)$.

Firstly, we truncate domain of y by an ℓ_∞-ball of radius R, denoted by $\mathcal{D}_1 = \{y : \|y\|_\infty \leq R\}$. We set the score approximation to be uniformly bounded by a constant depending on R and t. The domain truncation induces a small approximation error when the radius R is sufficiently large, as shown below.

Lemma B.5. *For any $R > 1$, x and $t > 0$, we have*

$$\int_{\|y\|_\infty \geq R} p_t(y|x) \lesssim R e^{-C_2' R^2}$$

$$\int_{\|y\|_\infty \geq R} \|\nabla \log p_t(y|x)\|_2^2 p_t(y|x) dy \lesssim \frac{1}{\sigma_t^4} R^3 e^{-C_2' R^2}$$

Secondly, we truncate $p_t(y|x)$ to control the density function.

Lemma B.6. *For any $R > 0$, x and $\epsilon_{low} > 0$,*

$$\int_{\|y\|_\infty \leq R} 1\{p_t(y|x) < \epsilon_{low}\} \, p_t(y|x) \, \mathrm{d}x \lesssim R^d \epsilon_{low}$$

$$\int_{\|y\|_\infty \leq R} 1\{p_t(y|x) < \epsilon_{low}\} \, \|\nabla \log p_t(y|x)\|^2 \, p_t(y|x) \, \mathrm{d}y \lesssim \frac{\epsilon_{low}}{\sigma_t^4} R^{d+2}$$

Proposition B.7. *For any $R > 0$, x and $\epsilon_{low} > 0$, We consider time $t \in [N^{-C_\sigma}, C_\sigma \log N]$ for constants C_σ and C_α. Given any integer $N > 0$, we constrain $(y, x) \in [0,1]^d \times [-C_x\sqrt{\log N}, C_x\sqrt{\log N}]^{d_x}$, where C_x is a constant depending on d, β, B, C_1 and C_2. Then there exists a ReLU neural network class $\mathcal{F}(M_t, W, k, L, K)$ which contains a mapping $s(\mathbf{y}, \mathbf{x}, t)$*

satisfying

$$p_t(y|x)\,\|\nabla \log p_t(y|x) - s(y,x,t)\|_\infty \lesssim \frac{B}{\sigma_t^2} N^{-\beta} (\log N)^{\frac{d+\beta+1}{2}} \quad \textit{for any } t \in [N^{-C_\sigma}, C_\alpha \log N].$$

Furthermore, the neural network hyperparameters satisfy

$$M_t = \mathcal{O}\left(\sqrt{\log N}/\sigma_t^2\right), \quad W = \mathcal{O}\left(N^{d+d_x} \log^7 N\right),$$

$$\kappa = \exp(\mathcal{O}(\log^4 N)), \quad L = \mathcal{O}(\log^4 N), \quad K = \mathcal{O}\left(N^{d+d_x} \log^9 N\right).$$

This proposition states that score function can be approximated in the L_∞ sense, which is necessary for the proof of 2.2. Using these results, we claim that the resulting $s(y,x,t) \in \mathcal{F}$ is an L_2 approximator of the score function. In this regard, we reduce the proof of Theorem 2.2 to the verification of this claim. Indeed, choosing $R = C_x\sqrt{\log N} = \sqrt{\frac{2\beta}{C_2}} \log N$ and $\epsilon_{\text{low}} = C_3 N^{-\beta} (\log N)^{\frac{d+\beta}{2}}$, we decompose the L_2 score approximation error as

$$\int_{\mathbb{R}^{d_y}} \|s(x,y,t) - \nabla \log p_t(y|x)\|_2^2 p_t(y|x) dy$$

$$= \int_{\|y\|_\infty > \sqrt{\frac{2\beta}{C_2}} \log N} \|s(y,x,t) - \nabla \log p_t(y|x)\|_2^2 p_t(y|x) dy \quad (A_1)$$

$$+ \int_{\|y\|_\infty \le \sqrt{\frac{2\beta}{C_2}} \log N} 1\{p_t(y|x) < \epsilon_{\text{low}}\} \|s(x,y,t) - \nabla \log p_t(y|x)\|_2^2 p_t(y|x) dy \quad (A_2)$$

$$+\int_{\|y\|_\infty \le \sqrt{\frac{2\beta}{C_2}}\log N} 1\{p_t(y|x) \ge \epsilon_{\text{low}}\}\|s(x,y,t) - \nabla \log p_t(y|x)\|_2^2 p_t(y|x)dy \quad (A_3).$$

Here (A_1) is the truncation error due to the unbounded range of y; (A_2) is the truncation error due to small $p_t(y|x)$. The remaining (A_3) is the approximation error of $s(y,x,t)$ on $\mathcal{D}$. These three terms can be bounded separately.

For the error bound of the estimated score, we employ the following lemma.

Lemma B.8. *Let $m_t = \frac{M_t}{\sqrt{\log N}}$, then for any $s \in \mathcal{F}(M_t, W, \kappa, L, K)$, we have $|\ell(s,x,y)| \lesssim \int_{t_0}^T m_t^2 dt = M$. In particular, if we take $t_0 = n^{-\mathcal{O}(1)}$ and $T = \mathcal{O}(\log n)$, we have $M = \log t_0$ for $m_t = \frac{1}{\sigma_t}$, and $M = \frac{1}{t_0}$ for $m_t = \frac{1}{\sigma_t^2}$ respectively.*

To convert approximation to statistical complexity bound, we need some techniques in empirical processes. We calculate the covering number of the loss function class $\mathcal{S}(R)$ and the following lemma shows the upper bound of the covering number. Intuitively, covering number measures the complexity of a functional class.

Lemma B.9. *Given $\delta > 0$, when $\|y\|_\infty < R$, the δ-covering number $N(\delta)$ of the loss function class $\mathcal{S}(R)$ satisfies*

$$N(\delta) \lesssim \left(\frac{2L^2(W\max(R,T)+2)\kappa^L W^{L+1}\log N}{\delta}\right)^{2K}$$

Then we can decompose the generalization error and bound them separately. Next, we bound second moment for the score estimation and finite KL divergence with respect to the

standard Gaussian, we can adopt Girsanov's Theorem and bound the KL divergence between the two distributions. We restate the lemma as follows.

Lemma B.10 (See Theorem 2 in S. Chen et al. (2022))**.** *Let p_0 be a probability distribution and let $Y = \{Y_t\}_{t\in[0,T]}$ and $Y' = \{Y'_t\}_{t\in[0,T]}$ be two stochastic processes that satisfies the following SDEs:*

$$dY_t = s(Y_t, t)dt + dW_t, Y_0 \sim p_0$$

$$dY'_t = s(Y'_t, t)dt + dW_t, Y'_0 \sim p_0$$

Suppose that $\int p_t(x)\|(s-s')(x,t)\|^2 dx \le C$ for any $t \in [0,T]$. Then we have

$$KL(p_T|p'_T) \le \int_0^T \frac{1}{2}\int p_t(x)\|(s-s')(x,t)\|^2 dxdt$$

To prove Theorem 2.3, we also need to bound the diffused distribution at the early stopping time t_0, which is presented as follows. Under the assumptions, $TV(P(\cdot|x), P_{t_0}(\cdot|x)) = \mathcal{O}(\sqrt{t_0}\log^{\frac{d+1}{2}}\frac{1}{t_0})$.

Therefore, for any $s \in \mathcal{F}, x \in [0,1]^{d_x}$, we have

$$\int_y p_t(y|x)\|s(x,y,t) - \nabla \log p_t(y|x)\|^2 dx \lesssim \int_y p_t(y|x)\frac{\|y\|^2 + C}{\sigma_t^4}dx \lesssim \frac{1}{\sigma_t^4}.$$

Here we invoke the bound on the score function and the bound on ReLU network $\|s\|_\infty \le M_t \lesssim \frac{\log N}{\sigma_t^2}$ for the first inequality, and we use the subGaussian property of $p_t(y|x)$.

Now we use another backward process as a transition term between $Y_t^{\leftarrow}$ and $\widetilde{Y}_t^{\leftarrow}$, which is defined as

$$dY_t^{\leftarrow} = \left[\frac{1}{2}Y_t^{\leftarrow} + \nabla \log p_{T-t}(Y_t^{\leftarrow}|x)\right] dt + dW_t \quad \text{with} \quad Y_0^{\leftarrow} \sim \mathcal{N}(0, I).$$

We denote the distribution of $Y_t^{\leftarrow}$ conditional on x by $P_{T-t}^{\leftarrow}(\cdot|x)$.

Since $Y^{\leftarrow}$ and $Y^{\leftarrow}$ are obtained through the same backward SDE but with different initial distributions, by Data Processing Inequality and Pinsker's Inequality, we have

$$\begin{aligned}
\mathrm{TV}(P_{t_0}(\cdot|x), P'_{t_0}(\cdot|x)) &\lesssim \sqrt{\mathrm{KL}(P_{t_0}(\cdot|x), P'_{t_0}(\cdot|x))} \\
&\lesssim \sqrt{\mathrm{KL}(P_T(\cdot|y)\|\mathcal{N}(0, I))} \\
&\lesssim \sqrt{\mathrm{KL}(P_0(\cdot|y)\|\mathcal{N}(0, I))}\exp(-T)
\end{aligned} \tag{1}$$

Thus, we could decompose the TV bound into

$$\begin{aligned}
\mathrm{TV}(P(\cdot|), \widetilde{P}_{t_0}(\cdot|x)) &\lesssim \mathrm{TV}(P(\cdot|x), P_{t_0}(\cdot|x)) + \mathrm{TV}(P_{t_0}(\cdot|x), P'_{t_0}(\cdot|y)) + \mathrm{TV}(P'_{t_0}(\cdot|y), \widetilde{P}_{t_0}(\cdot|y)) \\
&\lesssim \sqrt{t_0 \log^{(d+1)/2}} \cdot \frac{1}{t_0} + \exp(-T) \qquad (2) \\
&\quad + \sqrt{\int_{t_0}^{T} \frac{1}{2} \int_y p_t(y|x)\|\widehat{s}(x, y, t) - \nabla \log p_t(y|x)\|^2 dy dt} \qquad (3)
\end{aligned}$$

By taking expectation with respect to x, we have

$$\begin{aligned}
\mathbb{E}_x\left[\mathrm{TV}(P_{t_0}(\cdot|x), \widetilde{P}_{t_0})\right] &\lesssim \sqrt{t_0 \log^{(d+1)/2}} \cdot \frac{1}{t_0} + \exp(-T) + \mathbb{E}_x\left[\sqrt{\int_{t_0}^{T} \frac{1}{2} \int p_t(y|x)\|\widehat{s} - \nabla \log p_t\|^2 dy dt}\right] \\
&\lesssim \sqrt{t_0 \log^{(d+1)/2}} \cdot \frac{1}{t_0} + \exp(-T) + \sqrt{\frac{T}{2}\mathcal{R}(\widehat{s})},
\end{aligned}$$

where we invoke Jensen's inequality for the second inequality. Now we set $T = C_\alpha \log n$ for the constant $C_\alpha = \frac{2\beta}{d+d_y+2\beta}$ and take expectation with respect to $\{(y_i, x_i)\}_{i=1}^n$. Under the assumptions, again by Jensen's Inequality, we have

$$\mathbb{E}_{\{y_i,x_i\}_{i=1}^n}\left[\mathbb{E}_x\left[\mathrm{TV}(P_{t_0}, \widetilde{P}_{t_0})\right]\right] \lesssim \sqrt{t_0 \log^{(d+1)/2}} \cdot \frac{1}{t_0} + n^{-\frac{2\beta}{d+d_y+2\beta}} + \sqrt{\log \frac{1}{t_0} \cdot n^{-\frac{\beta}{2(d+d_y+\beta)}} (\log n)^{c'(\beta)}}$$

where $c'(\beta) = \max(9, \frac{\beta+1}{2})$. We can take $t_0 = n^{-\frac{4\beta}{d+d_x+2\beta}-1}$ so that

$$\sqrt{t_0 \log^{(d+1)/2}} \cdot \frac{1}{t_0} \lesssim n^{-\frac{2\beta}{d+d_y+2\beta}}, \quad \text{for sufficiently large } n$$

Thus, we bound the expected total variation by

$$\mathbb{E}_{\{(y_i,x_i)\}_{i=1}^n}\left[\mathbb{E}_y\left[\mathrm{TV}(P_{t_0}, \widetilde{P}_{t_0})\right]\right] = \mathcal{O}\left(n^{-\frac{2\beta}{d+d_x+2\beta}} (\log n)^{c'(\beta)+1/2}\right)$$

The proof is complete. □

Proof B.11 (Proof for Proposition 3.2)**.** We follow the approach of Hahn (1998). In calculating the variance bounds of average treatment effect τ, we refer to Bickel, Klaassen, Ritov, and Wellner (1993, Section 3.3). First, the tangent space is characterized. The density of (Y_0, Y_1, D, X) (with respect to some σ-finite measure) is given by

$$\bar{q}(y_0, y_1, d, x) = f(y_0, y_1 \mid x) p(x)^d (1 - p(x))^{1-d} f(x)$$

where $f(y_0, y_1 \mid x)$ and $f(x)$ denote the conditional distribution of (Y_0, Y_1) given X, and

the marginal distribution of X, respectively. The density of (Y, D, X) is then equal to

$$q(y, d, x) = [f_1(y \mid x)p(x)]^d \, [f_0(y \mid x)(1 - p(x))]^{1-d} \, f(x)$$

where $f_1(\cdot \mid x) = \int f(y_0, \cdot \mid x)dy_0$, and $f_0(\cdot \mid x) = \int f(\cdot, y_1 \mid x)dy_1$. Consider a regular parametric submodel

$$[f_1(y \mid x, \theta)p(x, \theta)]^d \, [f_0(y \mid x, \theta)(1 - p(x, \theta))]^{1-d} \, f(x, \theta)$$

which equals $q(y, d, x)$ when $\theta = \theta_0$. The corresponding score is given by

$$s(d, y, x \mid \theta) \equiv d \cdot s_1(y \mid x, \theta) + (1 - d) \cdot s_0(y \mid x, \theta) + \frac{d - p(x, \theta)}{p(x, \theta)(1 - p(x, \theta))} \cdot \dot{p}(x, \theta) + t(x, \theta)$$

where

$$s_1(y \mid x, \theta) = \frac{d}{d\theta} \log f_1(y \mid X, \theta),$$

$$s_0(y \mid x, \theta) = \frac{d}{d\theta} \log f_0(y \mid X, \theta),$$

$$\dot{p}(x, \theta) = \frac{d}{d\theta} p(x, \theta),$$

$$t(x, \theta) = \frac{d}{d\theta} \log f(X, \theta).$$

Then we obtain the tangent space of this model

$$\mathscr{S} = \{d \cdot s_1(y \mid x) + (1-d) \cdot s_0(y \mid x) + a(x) \cdot (d - p(x)) + t(x)\}$$

where $\int s_j(y \mid x) f_j(y \mid x) dy = 0 \ \forall x, \ j = 0, 1, \ \int t(x) f(x) dx = 0$, and $a(x)$ is any square-integrable measurable function of x.

Now, the average treatment effect is shown to be pathwise differentiable. For the parametric submodel under consideration, we find that

$$\tau(\theta) = \iint y f_1(y \mid x, \theta) f(x, \theta) dy dx - \iint y f_0(y \mid x, \theta) f(x, \theta) dy dx.$$

Thus,

$$\begin{aligned}\frac{\partial \tau(\theta_0)}{\partial \theta} = &\iint y s_1(y \mid x, \theta_0) f_1(y \mid x) f(x) dy dx + \int \beta_1(x) t(x, \theta_0) f(x) dx \\ &- \iint y s_0(y \mid x, \theta_0) f_0(y \mid x) f(x) dy dx - \int \beta_0(x) t(x, \theta_0) f(x) dx\end{aligned}$$

Let

$$F_\beta(Y, D, X) = \frac{D}{p(X)} \cdot (Y - \tau_1(X)) - \frac{1-D}{1-p(X)} \cdot (Y - \tau_0(X)) + \tau(X) - \tau,$$

For the parametric submodel whose score is given by $s(d, y, x \mid \theta)$, we have

$$\frac{\partial \beta(\theta_0)}{\partial \theta} = \mathbb{E}\left[F_\beta(Y, D, X) \cdot s(D, Y, X \mid \theta_0)\right]$$

from which we conclude that τ is a differentiable parameter. The variance bounds are the

expected squares of the projections of $F_\tau$on $\mathscr{S}$. Because $F_\tau \in \mathscr{S}$, the projections on $\mathscr{S}$ are themselves, and the variance bounds are the expected squares of the projections of F_β and F_γ. □

Proof B.12 (Proof for Theorem 3.3)**.** We first show that, generally, the nonparametric imputation method is efficient ,which is adapted from Hahn (1998). The asymptotic variance of $\hat{\tau}^{G_n}$ can be obtained as follows. This estimator takes the form

$$\frac{1}{n}\sum m(Z_i, \hat{h}_1, \hat{h}_2, \hat{h}_3)$$

where

$$h_1(x) = \mathbb{E}[D_i Y_i \mid X_i = x]$$

$$h_2(x) = \mathbb{E}[(1 - D_i) Y_i \mid X_i = x]$$

$$h_3(x) = \mathbb{E}[D_i \mid X_i = x]$$

and $\hat{h}_1, \hat{h}_2, \hat{h}_3$ are their estimators. Z_i denotes the observation for individual i. Let $h_1(\theta), h_2(\theta), h_3(\theta)$ denote the corresponding functions under some parametric submodel which equals the true model at $\theta = \theta_0$. Because $m(X_i, h_1, h_2, h_3)$ depends on h_1, h_2, h_3 only through their values $h_1(X_i), h_2(X_i), h_3(X_i)$, it follows that

$$\frac{\partial}{\partial\theta}\mathbb{E}[m(X_i, h_1(\theta), h_2(\theta), h_3(\theta))] = \frac{\partial}{\partial\theta}\mathbb{E}\left[\sum_{j=1}^{3} h_j(\theta)\cdot\delta_j(X_i)\right],$$

for $\delta_j(x) = \frac{\partial}{\partial h_j} m(x, h_1(\theta), h_2(\theta), h_3(\theta))\Big|_{\theta=\theta_0}$.

Notice that $\delta_1(x) = \frac{1}{p(x)}$, $\delta_2(x) = -\frac{1}{1-p(x)}$, and $\delta_3(x) = -\frac{\tau_1(x)}{p(x)} + \frac{\tau_0(x)}{1-p(x)}$. Newey's (1994)

Proposition 4 then suggests that the above estimator has the asymptotic influence function equal to

$$\frac{D_i(Y_i - \tau_1(X_i))}{p(X_i)} - \frac{(1 - D_i)(Y_i - \tau_0(X_i))}{1 - p(X_i)} + (\tau_1(X_i) - \tau_0(X_i) - (\tau_1 - \tau_0)),$$

so that its asymptotic variance equals the efficiency bound.

For semiparametric efficiency, we use the argument of Newey (1994). Assumption 5.1 in Newey (1994) requires a local linearization, which is shown below.

Assumption B.13. There is a function $D(z, h)$ that is linear in h such that for all h with $\|h - h_0\|$ small enough,

$$\|m(z, h) - m(z, h_0) - D(z, h - h_0)\| \leq b(z)\|h - h_0\|^2$$

and

$$\mathbb{E}[b(z)]\|\hat{h} - h_0\|^2 \xrightarrow{p} 0$$

This condition requires that the remainder term from a linearization be small. It is analogous to $m(z, h) - m(z, h_0) - \frac{\partial m(z,h_0)}{\partial h}(h - h_0)$ in the parametric case. The first part of this condition is satisfied with

$$D(z, h - h_0) = \frac{h_1 - h_{01}}{h_{03}} - \frac{h_2 - h_{02}}{h_{03}} - [\frac{h_{01}}{h_{03}^2} + \frac{h_{02}}{(1 - h_{03})^2}](h_3 - h_{03})$$

and $b(z) = C(1 + |\tau_1(x)| + |\tau_0(x)|)$.

Now the key step is to show that the first step nonparametric estimation of $\hat{h}$ has

convergence rates faster than $n^{-\frac{1}{4}}$. Suppose we use a diffusion model to estimate the conditional distribution $P(\cdot \mid x)$, obtaining an estimate $\hat{P}_{t_0}(\cdot \mid x)$ that satisfies the following total variation bound:

$$\mathbb{E}_{\{y_i,x_i\}_{i=1}^n}\left[\mathbb{E}_x\left[\mathrm{TV}\left(\hat{P}_{t_0}(\cdot \mid x), P(\cdot \mid x)\right)\right]\right] = \mathcal{O}\left(n^{-\frac{\beta}{d+d_x+2\beta}}(\log n)^{\max(\frac{19}{2},\frac{\beta+2}{2})}\right)$$

Now consider estimating the expectation using n^G i.i.d. samples $Y_1^{G_n}, \ldots, Y_{n^G}^{G_n} \sim \hat{P}_{t_0}(\cdot \mid x)$. Let

$$\hat{\mu}(x) = \frac{1}{n^G}\sum_{i=1}^{n^G} Y_i^{G_n}, \quad \mu(x) = \mathbb{E}_{Y\sim P(\cdot|x)}[Y]$$

Then the estimation error can be decomposed as

$$|\hat{\mu}(x) - \mu(x)| \le \left|\hat{\mu}(x) - \mathbb{E}_{\hat{P}_{t_0}(Y|X=x)}[Y]\right| + \left|\mathbb{E}_{P(Y|X=x)}[Y] - \mathbb{E}_{\hat{P}_{t_0}(Y|X=x)}[Y]\right|$$

By the central limit theorem, the first term corresponds to the Monte Carlo sampling error, which satisfies

$$\mathbb{E}\left|\hat{\mu}(x) - \mathbb{E}_{\hat{P}_{t_0}(Y|X=x)}[Y]\right| = \mathcal{O}\left(\frac{1}{\sqrt{n^G}}\right)$$

and the second term is bounded by the total variation distance:

$$\left|\mathbb{E}_{P(Y|X=x)}[Y] - \mathbb{E}_{\hat{P}_{t_0}(Y|X=x)}[Y]\right| \le 2 \cdot \mathrm{TV}(\hat{P}_{t_0}(Y|X=x), P(Y|X=x))$$

When n is sufficiently large, the log term $(\log n)^{\max(\frac{19}{2},\frac{2+\beta}{2})}$ can be ignored. Combining

the two terms, we conclude that

$$\mathbb{E}|\hat{\mu}(x) - \mu(x)| = \mathcal{O}\left(\frac{1}{\sqrt{n^G}}\right) + \mathcal{O}\left(n^{-\frac{\beta}{d+d_y+2\beta}}\right)$$

This shows that the overall error consists of a generation error governed by n^G and a statistical error governed by the training sample size n. Now we control the convergence rate by choosing appropriate parameters. Let $n^G > n^{\frac{1}{2}}$, then

$$\begin{aligned}\mathbb{E}|\hat{\mu}(x) - \mu(x)| &< \mathcal{O}(n^{-\frac{1}{4}}) + \mathcal{O}(n^{-\frac{1}{d/\beta+d_y/\beta+2}}) \\ &< \mathcal{O}(n^{-\frac{1}{4}}) + \mathcal{O}(n^{-\frac{1}{4}}) \\ &= \mathcal{O}(n^{-\frac{1}{4}})\end{aligned}$$

The second inequality hold when β is large, which means the data is smooth enough.

With further regularity conditions that

Assumption B.14 (Stochastic Equicontinuity)**.**

$$\sum_{i=1}^{n}[D(z_i, \hat{h} - h_0) - \int D(z, \hat{h} - h_0)dF_0]/\sqrt{n} \xrightarrow{p} 0$$

and

Assumption B.15 (Mean-Square Continuity)**.** (a) There is $\alpha(z)$ and a measure $\hat{F}$ such that $E[\alpha(z)] = 0$ and $E[\|\alpha(z)\|^2] < \infty$; and for all $\|\hat{h} - h_0\|$ small enough, $\int D(z, \hat{h} - h_0)dF_0 = \int \alpha(z)d\hat{F}$. (b) For the empirical distribution $\tilde{F} = \frac{1}{n}\sum_{i=1}^{n} 1\{z_i \leq z\}$, $\sqrt{n}[\int \alpha(z)d\hat{F} - \int \alpha(z)d\tilde{F}] \xrightarrow{p} 0$.

Stochastic equicontinuity is essential in empirical process theory and it is a sufficient condition for Donsker class. This assumption is similar to the assumption on the limited complexity of the functional class, and is satisfied by the results on finite covering numbers derived before. Linearization and stochastic equicontinuity involve second order terms and are therefore regularity conditions. They should be satisfied if $m(z, \tau, h)$ is smooth enough and $\hat{h}$ sufficiently well behaved. Mean-square continuity contains first order terms. These conditions are the ones that allow

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} m(z_i, \tau, \hat{h}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \alpha(z_i) \xrightarrow{d} \mathcal{N}(0, \alpha(z_i)\alpha(z_i)')$$

which corresponds to the first step influence function in semiparametric literature. It measures the effect of estimation of nuisance parameter h on the parameter of interest τ. Combining these results, we have

$$\sqrt{n} m(z_i, \hat{\tau}, \hat{h}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [m(z, \tau, h_0) + \alpha(z_i)] + o_p(1)$$

Therefore, using the result in Proposition 3.2, we have

$$\sqrt{n}(\hat{\tau}^{G_n} - \tau) \xrightarrow{d} \mathcal{N}(0, \mathbb{E}[\frac{\sigma_1^2(X_i)}{p(X_i)} + \frac{\sigma_0^2(X_i)}{1 - p(X_i)} + (\tau(X_i) - \tau)^2])$$

□

Training Procedures

Pseudocodes

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

Algorithm 1 Training

repeat

$\mathbf{Y}_0^0 \sim q(\mathbf{Y}_0), \mathbf{Y}_1^0 \sim q(\mathbf{Y}_1)$

$t \sim \text{Uniform}(\{1, \ldots, T\})$

$\boldsymbol{\epsilon}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\boldsymbol{\epsilon}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

Do gradient descent $\nabla_{\boldsymbol{\theta}} \left(\left\| \boldsymbol{\epsilon}_0 - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\sqrt{\bar{\alpha}_t}\mathbf{Y}_0^0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0, t, X) \right\|^2 + \left\| \boldsymbol{\epsilon}_1 - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\sqrt{\bar{\alpha}_t}\mathbf{Y}_1^0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_1, t, X) \right\|^2 \right)$

until converged

Algorithm 2 Sampling

$\tau^T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

$\mathbf{Y}_0^t \leftarrow \tau^T, \mathbf{Y}_1^t \leftarrow \tau^T$

for $t = T$ **down to** 1 **do**

$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

$\mathbf{Y}_0^{t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{Y}_0^t - \frac{1-\alpha_t}{\sqrt{1-\hat{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{Y}_0^t, t, X) \right) + \sigma_t \mathbf{z}$

$\mathbf{Y}_1^{t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{Y}_1^t - \frac{1-\alpha_t}{\sqrt{1-\hat{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{Y}_1^t, t, X) \right) + \sigma_t \mathbf{z}$

end for

return $\mathbf{Y}_0^0, \mathbf{Y}_1^0$

Table 5 Hyperparameters of CausalDiff

Hyper-parameter	Value / Setting
Initialization	Xavier Initialization for the weight matrix and Zero initialization for the 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}

Hyperparameter Settings for Training CausalDiff

Implemention 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.

Datasets

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 and Todd (2005). The treatment corresponds to receiving job training, while the observed outcomes are income and employment status.

Since counterfactual outcomes are unobservable on Jobs, 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, F. D. Johansson, and Sontag (2017) and Smith and 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:

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

Twins The Twins Almond, Chay, and Lee 2005 dataset is derived from all births in the USA between 1989-1991. We follow the settings of GANITE Yoon, Jordon, and Van Der Schaar

2018: It defines the treatment t = 1 as being the heavier twin (and t = 0 as being the lighter twin). The outcome is defined as the 1-year mortality. For each twin-pair, it obtained 30 features relating to the parents, the pregnancy and the birth: marital status; race; residence; number of previous births; pregnancy risk factors; quality of care during pregnancy; and number of gestation weeks prior to birth. Only twins weighing less than 2kg and without missing features (list-wise deletion) are chosen. This creates a complete dataset (without missing data). The final cohort is 11,400 pairs of twins whose mortality rate for the lighter twin is 17.7%, and for the heavier 16.1%. In this setting, for each twin pair it observed both the case t = 0 (lighter twin) and t = 1 (heavier twin); thus, the ground truth of individualized treatment effect is known in this dataset. In order to simulate an observational study, it selectively observes one of the two twins using the feature information (creating selection bias) as follows: $t|x \sim \text{Bern}(Sigmoid(w^T x + n))$, where $w^T \sim U((-0.1, 0.1)^{30\times 1})$ and $n \sim N(0, 0.1)$

ACIC18 The 2018 Atlantic Causal Inference Challenge (ACIC2018) (MacDorman and Atkinson 1998) contains 24 different settings of benchmark datasets. They are designed to benchmark causal inference algorithms with various data-generating mechanisms. Derived from the linked birth and infant death data, ACIC2018 provides 63 distinct data-generating mechanisms with around 40 non-equal-sized samples for each mechanism (n ranges from 1000 to 50000, $d_X = 177$) with a constant CATE but heterogeneous propensity scores for most datasets.

Earth Observation Dataset The Earth Observation Dataset (Jerzak, F. Johansson, and Daoud 2023) uses satellite imagery to evaluate anti-proverty aid program effect in Africa,

and formulate it as a satellite image-based observational causal inference task. The raw satellite imagery captured by NASA/USGS. Earth Observation Dataset contains 1382 satellite images and the image resolution is 984×984. We follow the simulation process of (Jerzak, F. Johansson, and Daoud 2023) (Sec. 4) to synthesize the data, and ResNet-101He et al. 2016 pre-trained on ImageNet-1K Deng et al. 2009 is adopted to extract feature vectors from satellite images to synthesize outcome variable and treatment dummy. To simulate real-world scenarios, we adopt ResNet-50He et al. 2016 pre-trained on ImageNet-1K Deng et al. 2009 (a different feature extractor) to extract feature vectors from satellite images and conduct experiments.

Additional Simulation and Experiment Results

Additional Simulation Details

For simulation, we generate 10,000 10-dimensional feature vectors, $\mathbf{x}$, sampled from $\mathcal{N}(\mathbf{0}, \Sigma^2)$, where $\Sigma = 0.5 \times (\Omega + \Omega^T)$ and $\Omega \sim \mathcal{U}((-1, 1)^{10\times 10})$. The outcome $\mathbf{y}$ conditional on observation $\mathbf{x}$ is given by $\mathbf{y} = \mathbf{w}_y^T \mathbf{x} + \mathbf{n}_y$, where $\mathbf{w}_y^T \sim \mathcal{U}((-0.1, 0.1)^{10\times 2})$, and $\mathbf{n}_y \sim \mathcal{N}(\mathbf{0}^{2\times 1}, 0.1^2 \times I^{2\times 2})$. Similar to Yoon, Jordon, and Van Der Schaar (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 $\mathbf{x}_1 \sim \mathcal{N}(\mu_1, \Sigma^2)$ or $\mathbf{x}_0 \sim \mathcal{N}(\mu_0, \Sigma^2)$, 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 KL divergences between distribution of $\mathbf{x}_0$ and $\mathbf{x}_1$. Figure 3 showcases CausalDiff's ability to recover ground-truth potential outcome distributions under high selection bias, demonstrating almost complete alignment between generated distributions

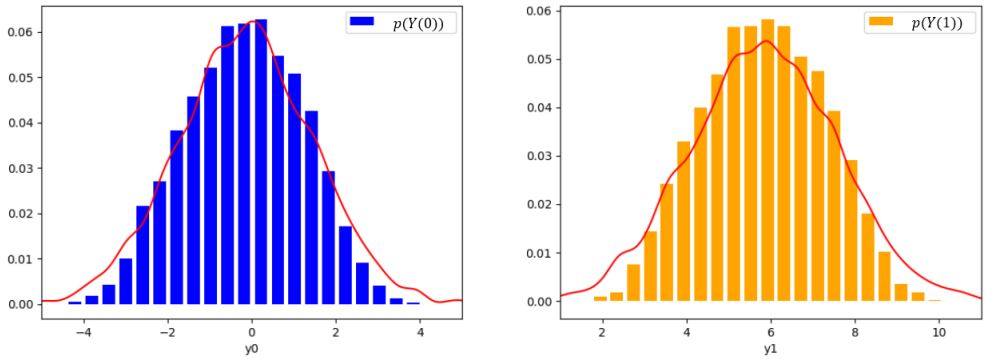

Figure 3 CausalDiff-generated samples (colored histograms) versus ground truth $Y(w)$ distributions (red curves).

and ground truth. This is true for both treatment arms and control arms.

We evaluate CausalDiff's ability to recover $p(Y(d)|X = x)$ under extreme selection bias ($D_{\text{KL}} = 600$), focusing on a fixed covariate value x_0. Using 10^2 generated samples per treatment arm, CausalDiff generates samples that are remarkably close to ground truth distributions (Figure 4)), confirming accurate counterfactual density matching despite limited overlap.

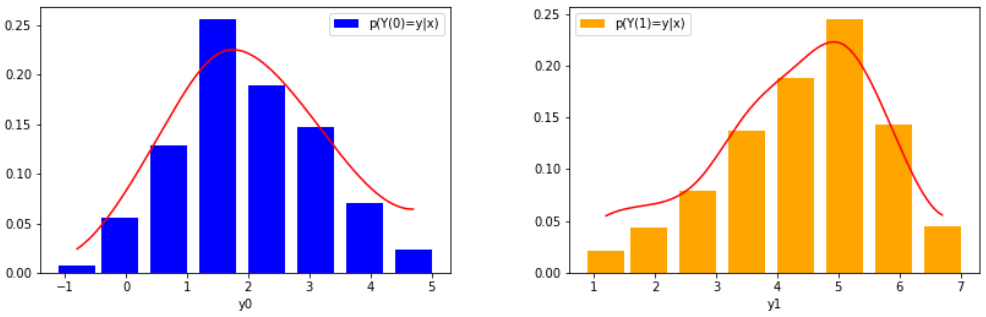

Figure 4 Conditional density estimation under selection bias. CausalDiff-generated samples (colored histograms) versus ground truth $p(Y(d)|X = x_0)$ (red curves).

Finally, we validate CausalDiff's asymptotic guarantees via sampling across the covariate space $\mathcal{X}$, selecting seven x with ground truth ITE $\tau(x) \in [0.5, 4]$. We construct 95% confidence

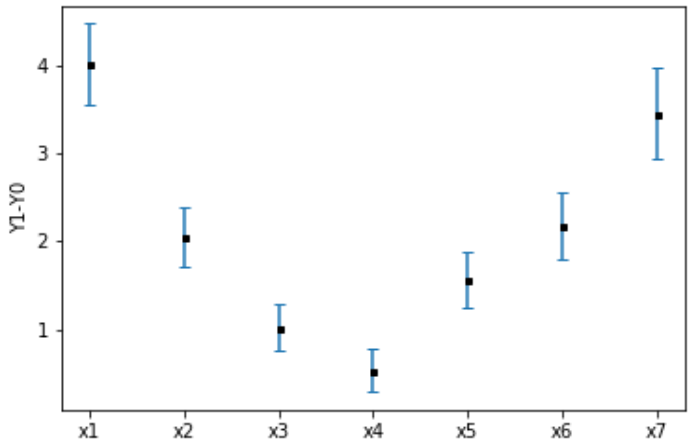

Figure 5 True ITE and predicted confidence intervals. Blue bars show 95% CIs for different x values (black dots: true $\tau(x)$).

intervals ($\text{CI}_{95\%}$) for each $\tau(x)$ through 10^2 Monte Carlo trials. Figure 5 demonstrates statistically tight coverage since all of $\text{CI}_{95\%}$ cover $\tau(x)$ in our experiment (mean interval width 0.36 ± 0.09).

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.

Dataset	**IHDP** ($\hat{\epsilon}_{ATE}$)		**IHDP** ($\sqrt{\hat{\epsilon}_{PEHE}}$)	
Methods	In-sample	Out-sample	In-sample	Out-sample
CausalDiff	**.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.

Alternative Version of Models

Multi-Treatment Extension Figure 6 extends CausalDiff 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 CausalDiff. This hints the potential cost of ignoring counterfactual information flow - a critical design insight for causal architectures.

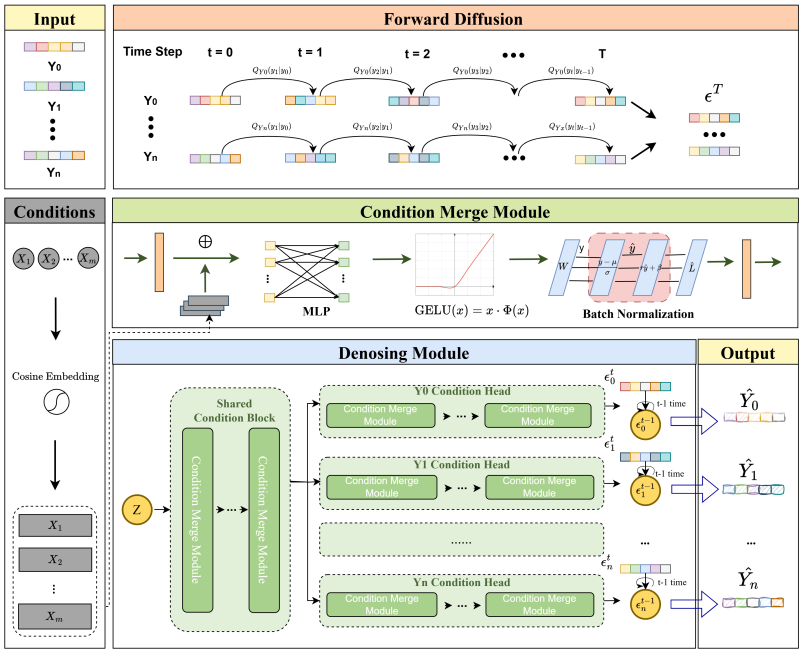

Figure 6 An Multi-Head Version of CausalDiff

Discussion of Limitations

While CausalDiff excels in complex settings with nonlinear responses and high dimensional inputs, its computational intensity makes classical low-dimensional linear regimes (e.g.,

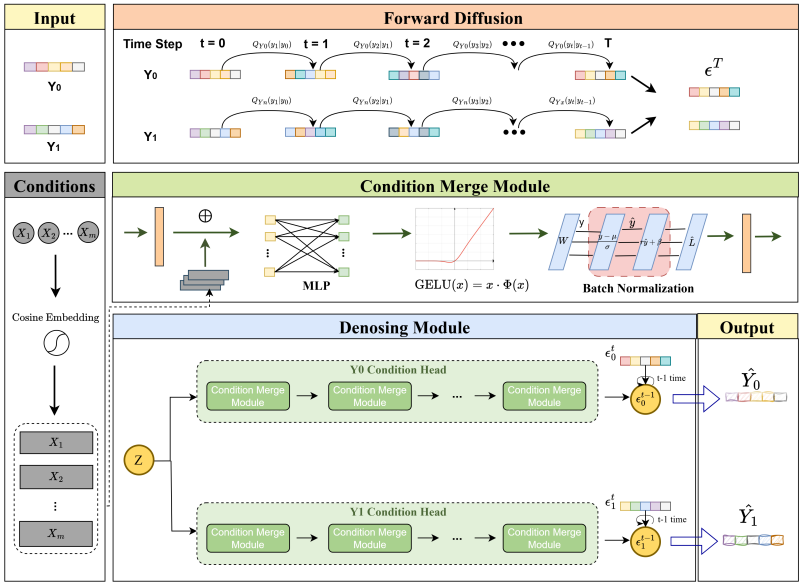

Figure 7 An Alternative Verion of Two-Head CausalDiff

randomized trials with small sample sizes) better served by simpler estimators. This trade-off reflects CausalDiff's design prioritization: sacrificing lightweight computation for unmatched fidelity in modern, data-rich environments. Future work will develop distillation/quantization techniques to enhance efficiency without sacrificing theoretical guarantees.

Statement of Impacts

CausalDiff 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, CausalDiff 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.

Our model can also perform well under scenarios such as multiple treatment and high dimensional unstructured inputs, further broadening its practical impact. The framework's theoretical guarantees (consistency, asymptotic normality, semi-parametric efficiency) directly enable sample-efficient confidence interval construction, bridging the gap between machine learning flexibility and statistical rigor required for high-stakes applications.