B Proof of Conditional Sliced Score Matching

In this section, we extend the conclusions of [49] to our conditional score function. We begin by summarizing the most commonly used notations. Let the dataset be $\mathcal{D} = \{(x_i, y_i, z_i)\}_{i=1}^n$ and the conditional probability be $p(x|z)$. The model score function, denoted as $s(x, z; \theta)$, corresponding to $p(x, z; \theta)$ where θ is restricted to a parameter space Θ. The goal of $s(x, z; \theta)$ is to approximate the true score function $\nabla_x \log p(x, z)$. The Hessian of $\log p(x, z)$ w.r.t. x is represented as $\nabla_x s(x, z; \theta)$. Additionally, we introduce v as a random vector of the same dimension as x, referred to as the projection vector, with p_v denoting its distribution.

Assumption B.1 (Regularity of conditional scores). *For any z, $s(x, z; \theta)$ and $\nabla_x \log p(x|z)$ are both differentiable w.r.t. x. Additionally, we assume that they satisfy $\mathbb{E}_{x\sim p(x|z)}[\|s(x, z; \theta)\|_2^2] < \infty$ and $\mathbb{E}_{x\sim p(x|z)}[\|\nabla_x \log p(x|z)\|_2^2] < \infty$.*

Assumption B.2 (Regularity of projection vectors). *The projection vectors satisfy $\mathbb{E}_{v\sim p_v}[\|v\|_2^2] < \infty$, and $\mathbb{E}_{v\sim p_v}[vv^T] \succ 0$.*

Assumption B.3 (Boundary conditions). *For any z, for all $\theta \in \Theta$, the score satisfy $\lim_{\|x\|\to\infty} s(x, z; \theta)p(x|z) = 0$.*

Assumption B.4 (Identifiability). *The model family $\{p(x, z; \theta) \mid \theta \in \Theta\}$ is well-specified, i.e., $p(x, z) = p(x, z; \theta^*)$. Furthermore, $p(x, z; \theta) \neq p(x, z; \theta^*)$ whenever $\theta \neq \theta^*$.*

Assumption B.5 (Positiveness). *The probability density function satisfies $p(x, z; \theta) > 0$, $\forall\theta \in \Theta$ and $\forall(x, z)$.*

Lemma B.1. *Assume $s(x, z; \theta)$, $\nabla_x \log p(x, z)$ and p_v satisfy some regularity conditions (Assumption B.1, Assumption B.2). Under proper boundary conditions (Assumption B.3), we have*

$$\begin{aligned}\mathcal{L}_\theta(z; p_v) &= \frac{1}{2}\mathbb{E}_{\substack{v\sim p_v \\ x\sim p(x|z)}} \left\{\left[v^T s(x, z; \theta) - v^T \nabla_x \log p(x|z)\right]_2^2\right\} \\ &= \mathbb{E}_{v\sim p_v}\mathbb{E}_{x\sim p(x|z)} \left\{v^T \nabla_x s(x, z; \theta)v + \frac{1}{2}\left[v^T s(x, z; \theta)\right]^2\right\} + C(z),\end{aligned} \tag{15}$$

where $C(z)$ is a constant w.r.t. θ.

Proof. For a fix z, , our proof follows the approach of [49]. To enhance readability, we recount the key details. Since expectations are bounded under Assumptions B.1 and B.2, we expand $\mathcal{L}_\theta(z; p_v)$ as

$$\begin{aligned}\mathcal{L}_\theta(z; p_v) &= \frac{1}{2}\mathbb{E}_{\substack{v\sim p_v \\ x\sim p(x|z)}} \left\{\left[v^T s(x, z; \theta) - v^T \nabla_x \log p(x|z)\right]_2^2\right\} \\ &= \frac{1}{2}\mathbb{E}_{v\sim p_v}\mathbb{E}_{x\sim p(x|z)}\Big\{[v^T s(x, z; \theta)]^2 + \left[v^T \nabla_x \log p(x|z)\right]^2 \\ &\qquad - 2\left[v^T s(x, z; \theta)\right]\left[v^T \nabla_x \log p(x|z)\right]\Big\} \\ &= \mathbb{E}_{v\sim p_v}\mathbb{E}_{x\sim p(x|z)}\Big\{-[v^T s(x, z; \theta)]\,[v^T \nabla_x \log p(x|z)] \\ &\qquad + \frac{1}{2}\left[v^T s(x, z; \theta)\right]^2\Big\} + C(z),\end{aligned} \tag{16}$$

where we have absorbed the term related to $\nabla_x \log p(x|z)$ into $C(z)$ since it does not depend on θ. Next, we show that

$$\begin{aligned}&-\mathbb{E}_{v\sim p_v}\mathbb{E}_{x\sim p(x|z)} \left\{\left[v^T s(x, z; \theta)\right]\left[v^T \nabla_x \log p(x|z)\right]\right\} \\ &\qquad = \mathbb{E}_{v\sim p_v}\mathbb{E}_{x\sim p(x|z)} \left[v^T \nabla_x s(x, z; \theta)v\right].\end{aligned} \tag{17}$$

This can be shown by first calculating that

$$\begin{aligned}&\mathbb{E}_{v\sim p_v}\mathbb{E}_{x\sim p(x|z)} \left\{\left[v^T s(x, z; \theta)\right]\left[v^T \nabla_x \log p(x|z)\right]\right\} \\ &= \mathbb{E}_{v\sim p_v} \int p(x|z)\left[v^T s(x, z; \theta)\right]\left[v^T \nabla_x \log p(x|z)\right]dx \\ &= \mathbb{E}_{v\sim p_v} \int \left[v^T s(x, z; \theta)\right]\left[v^T \nabla_x p(x|z)\right]dx \\ &= \mathbb{E}_{v\sim p_v} \sum_{i=1}^{d_x} \int \left[v^T s(x, z; \theta)\right]v_i \frac{\partial p(x|z)}{\partial x_i}dx,\end{aligned} \tag{18}$$

where recall that $x \in \mathbb{R}^{d_x}$. Then, applying multivariate integration by parts, we have

$$\begin{aligned}&\Bigg|\mathbb{E}_{v\sim p_v} \sum_{i=1}^{d_x} \int \left[v^T s(x, z; \theta)\right]v_i \frac{\partial p(x|z)}{\partial x_i}dx \\ &\qquad + \mathbb{E}_{v\sim p_v} \sum_{i=1}^{d_x} \int \left[v_i p(x|z)\right]v^T \frac{\partial s(x, z; \theta)}{\partial x_i}dx\Bigg| \\ &= \Bigg|\mathbb{E}_{v\sim p_v} \sum_{i=1}^{d_x}\Big\{\lim_{x_i\to+\infty} \left[v^T s(x, z; \theta)\right]v_i p(x|z) \\ &\qquad - \lim_{x_i\to-\infty} \left[v^T s(x, z; \theta)\right]v_i p(x|z)\Big\}\Bigg| \\ &\leq \sum_{i=1}^{d_x} \lim_{x_i\to+\infty} \sum_{j=1}^{d_x} \mathbb{E}_{v\sim p_v}|v_i v_j||s_j(x, z; \theta)p(x|z)| \\ &\qquad + \sum_{i=1}^{d_x} \lim_{x_i\to-\infty} \sum_{j=1}^{d_x} \mathbb{E}_{v\sim p_v}|v_i v_j||s_j(x, z; \theta)p(x|z)| \\ &\overset{\text{(i)}}{\leq} \sum_{i=1}^{D} \lim_{x_i\to\infty} \sum_{j=1}^{D} \sqrt{\mathbb{E}_{v\sim p_v} v_i^2 \mathbb{E}_{v\sim p_v} v_j^2} \cdot |s_j(x, z; \theta)p(x|z)| \\ &\qquad + \sum_{i=1}^{d_x} \lim_{x_i\to-\infty} \sum_{j=1}^{d_x} \sqrt{\mathbb{E}_{v\sim p_v} v_i^2 \mathbb{E}_{v\sim p_v} v_j^2} \cdot |s_j(x, z; \theta)p(x|z)| \overset{\text{(ii)}}{=} 0,\end{aligned}$$

where $s_j(x, z; \theta)$ denotes the j-th component of $s(x, z; \theta)$. In the above derivation, (i) is due to Cauchy-Schwarz inequality and (ii) is from the Assumption B.2 and B.3 that $\mathbb{E}_{v\sim p_v}[\|v\|^2] < \infty$ and $\lim_{\|x\|\to\infty} s(x, z; \theta)p(x|z) = 0$. As a result, for Eq. (18), we have

$$\begin{aligned}&\mathbb{E}_{v\sim p_v} \sum_{i=1}^{d_x} \int \left[v^T s(x, z; \theta)\right]v_i \frac{\partial p(x|z)}{\partial x_i}dx \\ &= -\,\mathbb{E}_{v\sim p_v} \sum_{i=1}^{d_x} \int \left[v_i p(x|z)\right]v^T \frac{\partial s(x, z; \theta)}{\partial x_i}dx \\ &= -\,\mathbb{E}_{v\sim p_v} \int p(x|z)\left[v^T \nabla_x s(x, z; \theta)v\right]dx \\ &= -\,\mathbb{E}_{v\sim p_v}\mathbb{E}_{x\sim p(x|z)}\left[v^T \nabla_x s(x, z; \theta)v\right],\end{aligned} \tag{19}$$

which proves Eq. (17) and the proof is completed. □

Lemma B.1 derives $\mathcal{L}_\theta(z;p_v)$ while avoiding terms related to $\nabla_x \log p(x|z)$. The overall objective is then obtained by marginalization, i.e., $\mathcal{L}_\theta(p_v) := \mathbb{E}_{z\sim p(z)}[\mathcal{L}_\theta(z;p_v)]$. Next, in Lemma B.2, we establish key properties of $\mathcal{L}_\theta(p_v)$ showing that the solution satisfying $\mathcal{L}_\theta(p_v) = 0$ corresponds to the optimal parameter θ^*.

Lemma B.2. *Assume the model family is well-specified and identifiable (Assumption B.4). Assume further that the densities are all positive (Assumption B.5). When p_v satisfies Assumption B.2, we have*

$$\mathcal{L}_\theta(p_v) := \mathbb{E}_{z\sim p(z)}[\mathcal{L}_\theta(z;p_v)] = 0 \Leftrightarrow \theta = \theta^*. \tag{20}$$

Proof. e obtain the proof by extending the results of [49] to the conditional case. Under Assumptions B.4 and B.5, for all (x,z), $p(x,z) = p(x,z;\theta^*) > 0$. Recall the definition of the loss function:

$$\mathcal{L}_\theta(p_v) := \frac{1}{2}\mathbb{E}_{\substack{v\sim p_v\\(x,z)\sim p(x,z)}}\left\{\left[v^T s(x,z;\theta) - v^T\nabla_x \log p(x|z)\right]_2^2\right\}. \tag{21}$$

Hence, $\mathcal{L}_\theta(p_v) = 0$ implies for all (x,z), $\mathbb{E}_{v\sim p_v}\left\{\left[v^T s(x,z;\theta) - v^T\nabla_x \log p(x|z)\right]_2^2\right\} = 0$. Further,

$$\begin{aligned}
&\mathbb{E}_{v\sim p_v}\left\{\left[v^T s(x,z;\theta) - v^T\nabla_x \log p(x|z)\right]_2^2\right\} = 0\\
\Leftrightarrow\ &\mathbb{E}_{v\sim p_v}\Big\{v^T\left[s(x,z;\theta) - \nabla_x \log p(x|z)\right]\\
&\qquad\cdot\left[s(x,z;\theta) - \nabla_x \log p(x|z)\right]^T v\Big\} = 0\\
\Leftrightarrow\ &\left[s(x,z;\theta) - \nabla_x \log p(x|z)\right]^T\cdot\mathbb{E}_{v\sim p_v}\left[vv^T\right]\\
&\qquad\cdot\left[s(x,z;\theta) - \nabla_x \log p(x|z)\right] = 0.
\end{aligned} \tag{22}$$

By the Assumption B.2, $\mathbb{E}_{v\sim p_v}\left[vv^T\right]$ is positive definite. Therefore, Eq. (22) implies that for any (x,z), the equality $s(x,z;\theta) = \nabla_x \log p(x|z)$ holds. Integrating both sides w.r.t. x, yields:

$$s(x,z;\theta) = \nabla_x \log p(x|z) \Leftrightarrow p(x,z;\theta) = p(x,z) + C_0, \tag{23}$$

where note that $p(x,z;\theta)$ is the probability density function corresponding to the score function $s(x,z;\theta)$ and C_0 is a constant. Since both $p(x,z;\theta)$ and $p(x,z)$ are normalized probability density functions, thus $C_0 = 0$. Therefore, by Assumption B.4, we conclude that $\theta = \theta^*$. The remain proof for the right-to-left direction is trivial. □

Thus, the process of finding the optimal parameters is equivalent to minimizing the loss objective $\mathcal{L}_\theta(p_v)$. In the main paper, we omit constant terms that are independent of θ, i.e., the final optimization objective is given by

$$\mathcal{J}_\theta = \mathbb{E}_{\substack{v\sim p_v\\(x,z)\sim p(x,z)}}\left\{v^T\nabla_x s(x,z;\theta)v + \frac{1}{2}\left[v^T s(x,z;\theta)\right]^2\right\}. \tag{24}$$

Note that $\theta^* = \underset{\theta\in\Theta}{\arg\min}\,\mathcal{J}_\theta$. Further, in practice, a finite sample approximation is used, which is expressed as:

$$\widehat{\mathcal{J}}_\theta = \frac{1}{nm}\sum_{i,j}^{n,m}\Big\{v_{ij}^T\nabla_{x_i} s(x_i,z_i;\theta)v_{ij}^T + \frac{1}{2}\left[v_{ij}^T s(x_i,z_i;\theta)\right]^2\Big\}. \tag{25}$$

Next, we prove the consistency of $\hat{\theta}_{n,m} := \underset{\theta\in\Theta}{\arg\min}\,\widehat{\mathcal{J}}_\theta$. The following additional assumptions are needed.

Assumption B.6. *The parameter space Θ is compact.*

Assumption B.7 (Lipschitz continuity). *Both the term $\nabla_x s(x,z;\theta)$ and $s(x,z;\theta)s(x,z;\theta)^T$ are Lipschitz continuous in terms of Frobenious norm, i.e., for all $\theta_1,\theta_2 \in \Theta$, $||\nabla_x s(x,z;\theta_1) - \nabla_x s(x,z;\theta_2)||_F \le L_1(x,z)||\theta_1-\theta_2||_2$, $||s(x,z;\theta_1)s(x,z;\theta_1)^T - s(x,z;\theta_2)s(x,z;\theta_2)^T||_F \le L_2(x,z)||\theta_1-\theta_2||_2$. In addition, we require the Lipschitz constant satisfy $\mathbb{E}_{(x,z)}[L_1^2(x,z)] < \infty$ and $\mathbb{E}_{(x,z)}[L_2^2(x,z)] < \infty$.*

Assumption B.8 (Bounded moments of projection vectors). *The moments of projection vectors satisfy $\mathbb{E}_{v\sim p_v}[||vv^T||_F^2] < \infty$.*

Lemma B.3 (Uniform convergence of the expected error). *Under the Assumption B.6-B.8, we have*

$$\mathbb{E}_{\substack{v\sim p_v\\(x,z)\sim p(x,z)}}\left[\sup_{\theta\in\Theta}\left|\widehat{\mathcal{J}}_\theta - \mathcal{J}_\theta\right|\right] \le O\left(\mathrm{diam}(\Theta)\sqrt{\frac{d_\Theta}{n}}\right), \tag{26}$$

where $\mathrm{diam}(\cdot)$ *denotes the diameter and d_Θ is the dimension of* Θ.

Proof. The proof follows by modifying the proof of Lemma 3 in [49]. We begin by defining $f(\theta;x,z,v) := v^T\nabla_x s(x,z;\theta)v + \frac{1}{2}(v^T s(x,z;\theta))^2$. Under Assumptions B.7 and B.8, we can show that $f(\theta;x,v)$ is Lipschitz continuous with constant $L(x,z,v)$ satisfying $\mathbb{E}_{(x,z)\sim p(x,z),v\sim p_v}[L^2(x,z,v)] < \infty$. Using a symmetrization trick and chaining technique, we can then derive the bound

$$\begin{aligned}
&\mathbb{E}_{\substack{v\sim p_v\\(x,z)\sim p(x,z)}}\left[\sup_{\theta\in\Theta}\left|\widehat{\mathcal{J}}_\theta - \mathcal{J}_\theta\right|\right]\\
&\le O(1)\,\mathrm{diam}(\Theta)\sqrt{\frac{d_\Theta}{n}}\sqrt{\mathbb{E}_{(x,z)\sim p(x,z),v\sim p_v}[L^2(x,z,v)]},
\end{aligned} \tag{27}$$

which completes the proof. □

Theorem B.4 (Consistency). *Under the Assumption B.1-B.8, $\hat{\theta}_{n,m}$ is consistent, meaning that $\hat{\theta}_{n,m} \xrightarrow{p} \theta^*$ as $n \to \infty$.*

Proof. The proof follows directly from the arguments in Theorem 2 of [49]. Specifically, the objective $\mathcal{J}_\theta$ exhibits similar continuous properties in the compact parameter space Θ as outlined in Lemma B.3. □

C Proof of Langevin Dynamics Conditional Sampling (LDCS)

We start by recalling the process of Langevin dynamics conditional sampling (LDCS). Let the step size be h, the total time be T. For a fixed z, the sampling process that iteratively updates x_{kh} as

$$x_{(k+1)h} = x_{kh} + h\cdot s(x_{kh},z;\hat{\theta}_{n,m}) + \sqrt{2h}\cdot\xi_{kh}, \tag{28}$$

where $\xi_{kh} \sim \mathcal{N}(0,I_{d_x})$ and $x_0 \sim p_0(x|z)$ for initializing. We take $p_0(x|z) = \mathcal{N}(0,I_{d_x})$ in practice. In the following, we analyze the error control of the generated distribution. The proof in this section requires the following additional assumptions.

Assumption C.1 (Smoothness). *For any z, $\log p(x|z)$ is continuously differentiable (C^1) w.r.t. x and is L_z-smooth w.r.t. x, meaning that the conditional score function $\nabla_x \log p(x|z)$ is L_z-Lipschitz. Additionally, we assume $L_z \ge 1$ for all z.*

Assumption C.2 (Log-Sobolev inequality constraints). *For any z, we assume that $p(x|z)$ satisfies a log-Sobolev inequality with constant $C_{z;LS}$. Furthermore, we assume $C_{z;LS} \ge 1$ for all z.*

Assumption C.3 (L^2-accurate). *For any z, the error in the conditional score estimate is bounded in L^2, i.e.*

$$\begin{aligned}&\|\nabla_x \log p(x|z) - s(x,z;\hat\theta_{n,m})\|^2_{L^2(p(x|z))}\\ &:= \mathbb{E}_{x\sim p(x|z)}[\|\nabla_x \log p(x|z) - s(x,z;\hat\theta_{n,m})\|^2] \le \varepsilon_z^2.\end{aligned} \tag{29}$$

Remark. Note that Assumption C.3 is closely related to the result of score matching in the previous step. Recall that, according to Theorem B.4, we have shown that $\hat\theta_{n,m} \xrightarrow{p} \theta^*$ as $n \to \infty$. By the continuous mapping theorem, this implies that for all x, z, $s(x,z;\hat\theta_{n,m}) \xrightarrow{p} s(x,z;\theta^*)$ as $n \to \infty$. Therefore, Assumption C.3 holds asymptotically, but for convenience in stating the following theorem, we include it as an assumption at this stage.

The error bound between the sampled distribution and the data distribution is provided by the following theorem. The proof can be derived by modifying the proof of Theorem 1.2 in [23], assuming an L^2-accurate conditional score function estimate. The primary difference is that the constants in the analysis now depend on z.

Theorem C.1 (LDCS with L^2-accurate score estimate). *Under Assumptions C.1–C.3, consider an accuracy requirement in total variation (TV) distance:* $0 < \varepsilon_{z;TV} < 1$. *Suppose further that the initial distribution satisfies* $d_{\chi^2}\{p_0(x|z)\|p(x|z)\} \le K_z^2$. *Then if*

$$\varepsilon_z \le \frac{\varepsilon^4_{z;TV}}{174080\sqrt{5}d_x L_z^2 C^{5/2}_{z;LS}\max\{\ln(2K_z/\varepsilon^2_{z;TV}), 2K_z\}}, \tag{30}$$

then running LDCS with score estimate $s(x,z;\hat\theta_{n,m})$, *step size* $h = \frac{\varepsilon^2_{z;TV}}{2720 d_x L_z^2 C_{z;LS}}$, *and total time* $T = 4C_{z;LS}\ln\left(\frac{2K_z}{\varepsilon^2_{z;TV}}\right)$, *yields a distribution* $p_T(x|z)$ *satisfying the error bound*

$$d_{TV}\{p_T(x|z), p(x|z)\} \le 2\varepsilon_{z;TV}. \tag{31}$$

Proof. For fixed z, the proof can be obtained by modifying the proof of Theorem 1.2 in [23]. □

To simplify the notation, we introduce universal constants that hold for all z. Specifically, we define $K^2 := \sup_z\{K_z^2\}$, the Lipschitz constant $L = \sup_z\{L_z\}$ and the constant for log-Sobolev inequality as $C_{\mathrm{LS}} := \sup_z\{C_{z;\mathrm{LS}}\}$. Then if we aim to control the accuracy for all z within $2\varepsilon_{TV}$, we require that for all z,

$$\begin{aligned}\varepsilon_z &\le \frac{\varepsilon^4_{\mathrm{TV}}}{174080\sqrt{5}d_x L^2 C^{5/2}_{\mathrm{LS}}\max\{\ln(2K/\varepsilon^2_{\mathrm{TV}}), 2K\}} =: \varepsilon_c,\\ h &= \frac{\varepsilon^2_{\mathrm{TV}}}{2720 d_x L^2 C_{\mathrm{LS}}}, T = 4C_{\mathrm{LS}}\ln\left(\frac{2K}{\varepsilon^2_{\mathrm{TV}}}\right).\end{aligned} \tag{32}$$

Theorem C.2 (Error bound of conditional distribution). *Under Assumptions C.1 and C.2, running LDCS with the score estimate* $s(x,z;\hat\theta_{n,m})$, *with an appropriate step size h, and time T, then for any z, results in a conditional distribution* $p_T(x|z)$ *such that the error guarantee that* $d_{TV}\{p_T(x|z), p(x|z)\} = o_p(1)$.

Proof. By Theorem B.4, we have shown that $\hat\theta_{n,m} \xrightarrow{p} \theta^*$ as $n \to \infty$. Applying the continuous mapping theorem, we obtain that for all x, z, $s(x,z;\hat\theta_{n,m}) \xrightarrow{p} s(x,z;\theta^*)$ as $n \to \infty$. In other equivalent form, for all x, z, for any $\epsilon > 0$, we have

$$\lim_{n\to\infty}\mathbb{P}\left(\|s(x,z;\hat\theta_{n,m}) - s(x,z;\theta^*)\| \le \epsilon\right) = 1. \tag{33}$$

Additionally, since under the condition given by Eq. (32), the event $d_{\mathrm{TV}}\{p_T(x|z), p(x|z)\} \le 2\varepsilon_{\mathrm{TV}}$ will happen, yield:

$$\begin{aligned}&\mathbb{P}\left(d_{\mathrm{TV}}\{p_T(x|z), p(x|z)\} \le 2\varepsilon_{\mathrm{TV}}\right)\\ &\ge \mathbb{P}\left(\mathbb{E}_{x\sim p(x|z)}[\|\nabla_x \log p(x|z) - s(x,z;\hat\theta_{n,m})\|^2] \le \varepsilon_c^2\right)\\ &= \mathbb{P}\left(\mathbb{E}_{x\sim p(x|z)}[\|s(x,z;\theta^*) - s(x,z;\hat\theta_{n,m})\|^2] \le \varepsilon_c^2\right)\\ &\ge \mathbb{P}\left(\|s(x,z;\theta^*) - s(x,z;\hat\theta_{n,m})\|^2 \le \varepsilon_c^2\right).\end{aligned} \tag{34}$$

By setting $\epsilon = \varepsilon_c$, and taking the limit on both sides, we obtain

$$\begin{aligned}&\lim_{n\to\infty}\mathbb{P}\left(d_{\mathrm{TV}}\{p_T(x|z), p(x|z)\} \le 2\varepsilon_{\mathrm{TV}}\right)\\ &\ge \lim_{n\to\infty}\mathbb{P}\left(\|s(x,z;\theta^*) - s(x,z;\hat\theta_{n,m})\|^2 \le \varepsilon_c^2\right) = 1\end{aligned} \tag{35}$$

for any given $\varepsilon_{\mathrm{TV}} \in (0,1)$, thus by definition of "converge in distribution" notion, we have $d_{\mathrm{TV}}\{p_T(x|z), p(x|z)\} = o_p(1)$. □

D Proof of Exchangeablility

In this section, we prove the exchangeability property, which ensures the validity of p-values under certain assumptions.

Proposition D.1 (Exchangeablility of triples). *Let* $\overset{d}{=}$ *denotes equality in distribution. Then under* $\mathcal{H}_0$, *and further assuming that for all* $b \in [B]$, $(X^{(b)}, Y, Z) \overset{d}{=} (X, Y, Z)$, *the resulting random sequence of triples* $(X^{(b)}, Y, Z)_{b=0}^B$ *is exchangeable. Recall that in the main paper, we use* $(X^{(0)}, Y, Z)$ *to represent the observed triple* (X, Y, Z).

Proof. A sequence of random variables is said to be exchangeable if its distribution is invariant under variable permutations. By the "representation theorem" [10] for exchangeable sequences of random variables, that show that every sequence of conditionally $i.i.d.$ random variables can be considered as a sequence of exchangeable random variables. Recall the process of our generative model, we start from $i.i.d.$ sequence of init random variables $x_0^{(b)}$, which are iteratively updated as:

$$x^{(b)}_{(k+1)h} = x^{(b)}_{kh} + h\cdot s(x^{(b)}_{kh}, Z;\theta) + \sqrt{2h}\cdot\xi_{kh}, \tag{36}$$

where $\xi_{kh} \sim \mathcal{N}(0, I_{d_x})$. Note that for each step $t = kh$, we can represent the generated process of $x_T^{(b)}$ as

$$x_T^{(b)} = \phi_T(\cdots\phi_t(\cdots\phi_1(x_0^{(b)}; Z, \xi_h); Z, \xi_{kh}); Z, \xi_T), \tag{37}$$

where $\phi_t(x^{(b)}_{(k-1)h}; Z, \xi_{kh}) = x^{(b)}_{(k-1)h} + h\cdot s(x^{(b)}_{(k-1)h}, Z;\theta) + \sqrt{2h}\cdot\xi_{kh}$. By the construction of Eq. (37), since the score function $s(\cdot, z;\theta)$ is measurable and the additional noise ξ_{kh} and Z are independent to the $x_0^{(b)}$, the resulting random sequence of random variables $(X^{(b)}, Y, Z)_{b=1}^B$ is exchangeable according to the "representation theorem", thus completes the proof. □

Next, we show the exchangeability of the statistic derived from the random sequence, as shown in the following corollary.

Corollary D.2 (Exchangeablility of statistics). *Let $\overset{d}{=}$ denotes equality in distribution. Then under $\mathcal{H}_0$, and further assume that for all $b \in [B]$, $(X^{(b)}, Y, Z) \overset{d}{=} (X, Y, Z)$, the resulting random sequence of statistics $[\rho(X^{(b)}, Y, Z)]_{b=0}^{B}$ is exchangeable.*

Proof. By Proposition D.1, the resulting random sequence of triples $(X^{(b)}, Y, Z)_{b=0}^{B}$ is exchangeable. Since ρ is a measurable function, the sequence of statistics $[\rho(X^{(b)}, Y, Z)]_{b=0}^{B}$ is also exchangeable by the "representation theorem". □

Given that the sequence of statistics is exchangeable, we now demonstrate that the p-value obtained by

$$p\text{-value} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\{\rho(X^{(b)}, Y, Z) \geq \rho(X, Y, Z)\}}{1 + B} \tag{38}$$

is valid, as shown in the following proposition.

Proposition D.3 (Valid p-value). *Under $\mathcal{H}_0$, and assuming that for all $b \in [B]$, $(X^{(b)}, Y, Z) \overset{d}{=} (X, Y, Z)$, let the random sequence of statistics be $[\rho(X^{(b)}, Y, Z)]_{b=0}^{B}$. Then the p-value given by Eq. (38) is valid, i.e.,*

$$\mathbb{P}(p\text{-}value \leq \alpha | \mathcal{H}_0) \leq \alpha, \ \textit{for any given } \alpha \in (0, 1). \tag{39}$$

Proof. For simplify, we also write $\mathbb{P}(\cdot|\mathcal{H}_0)$ as $\mathbb{P}_{\mathcal{H}_0}$. For any given $\alpha \in (0, 1)$, we have

$$\begin{aligned}
&\mathbb{P}_{\mathcal{H}_0}(p\text{-value} \leq \alpha) \\
&= \mathbb{P}_{\mathcal{H}_0}\left(\frac{1 + \sum_{b=1}^{B} \mathbf{1}\{\rho(X^{(b)}, Y, Z) \geq \rho(X, Y, Z)\}}{1 + B} \leq \alpha \right) \\
&\leq \mathbb{P}_{\mathcal{H}_0}\left(\sum_{b=1}^{B} \mathbf{1}\{\rho(X^{(b)}, Y, Z) \geq \rho(X, Y, Z)\} \leq \lfloor \alpha(1 + B) \rfloor \right).
\end{aligned} \tag{40}$$

Since the sequence $[\rho(X^{(b)}, Y, Z)]_{b=0}^{B}$ is exchangeable, by the property of order statistics, we have

$$\begin{aligned}
&\mathbb{P}_{\mathcal{H}_0}\left(\sum_{b=1}^{B} \mathbf{1}\{\rho(X^{(b)}, Y, Z) \geq \rho(X, Y, Z)\} \leq \lfloor \alpha(1 + B) \rfloor \right) \\
&= \frac{\lfloor \alpha(1 + B) \rfloor}{1 + B} \leq \alpha,
\end{aligned} \tag{41}$$

which completes the proof. □

Remark. Note that all the above results assume that for all $b \in [B]$, $(X^{(b)}, Y, Z) \overset{d}{=} (X, Y, Z)$, that the distribution of the generated samples perfectly models the conditional distribution. However, in practical applications, the generative model may introduce some error, even if we have provided an upper bound on this error, as analyzed in detail in Sec. B and Sec. C. Therefore the actual p-value estimate will have some deviation compared to the theoretical value caused by the estimation error. As a result, in the next Sec. E, we will further examine the validity of the p-value within our CI testing framework and obtain the Type I error Bound.

E Proof of Type I error Bound

To simplify, we separate the samples used in the previous stage of generative modeling from the samples used in the CI test, and the number of samples is denoted as N and n, respectively. We denote the estimated conditional distribution as $p_{T;N}(x|z)$. We further define $\boldsymbol{X} := (x_1, x_2, ..., x_n)^T, \boldsymbol{Y} := (y_1, y_2, ..., y_n)^T$ and $\boldsymbol{Z} := (z_1, z_2, ..., z_n)^T$ as the vectors formed by n samples. Additionally, for $b \in [B]$, we define $\boldsymbol{X}^{(b)} := (x_1^{(b)}, x_2^{(b)}, ..., x_n^{(b)})^T$ that are the generated vector corresponding to $\boldsymbol{Z}$. Then the estimation of statistic is given by $\hat{\rho} = \rho(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$. Recall that we have defined the estimation of threshold in the main paper as $c_\alpha := \inf\{c \in \mathbb{R} : \mathbb{P}(\hat{\rho} > c) \leq \alpha\}$. The following results give a bound for Type I error given $\boldsymbol{Y}, \boldsymbol{Z}$.

Lemma E.1 (Type I error bound given $\boldsymbol{Y}, \boldsymbol{Z}$). *Assume $\mathcal{H}_0 : X \perp\!\!\!\perp Y|Z$ is true. Under all the Assumptions in Sec. B and C, for any signifiance level $\alpha \in (0, 1)$, given $\boldsymbol{Y}, \boldsymbol{Z}$, the bound for the Type I error is obtained as*

$$\begin{aligned}
\mathbb{P}_{\mathcal{H}_0}(\hat{\rho} > c_\alpha | \boldsymbol{Y}, \boldsymbol{Z}) &\leq \alpha + d_{TV}\{p_{T;N}(\cdot|\boldsymbol{Z}), p(\cdot|\boldsymbol{Z})\} \\
&= \alpha + o_p(1), \ \textit{as } N \longrightarrow \infty.
\end{aligned} \tag{42}$$

Proof. By definition, the statistic $\hat{\rho}$ results in a p-value $< \alpha$ if and only if the observed variables are contained in the set A_α^B, where each element $(\boldsymbol{x}, \boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(B)})$ satisfies

$$\frac{1}{B+1}\left[1 + \sum_{b=1}^{B} \mathbf{1}\{\rho(\boldsymbol{x}^{(b)}, \boldsymbol{Y}, \boldsymbol{Z}) \geq \rho(\boldsymbol{x}, \boldsymbol{Y}, \boldsymbol{Z})\} \right] < \alpha. \tag{43}$$

Let $\hat{\boldsymbol{X}} \sim p_{T;N}(\cdot|\boldsymbol{Z})$ be sampled from the estimated conditional distribution. Then it holds that,

$$\begin{aligned}
\mathbb{P}_{\mathcal{H}_0}(\hat{\rho} > c_\alpha | \boldsymbol{Y}, \boldsymbol{Z}) &= \mathbb{P}_{\mathcal{H}_0}\left((\boldsymbol{X}, \boldsymbol{X}^{(1)}, ..., \boldsymbol{X}^{(B)}) \in A_\alpha^B | \boldsymbol{Y}, \boldsymbol{Z} \right) \\
&= \mathbb{P}_{\mathcal{H}_0}\left((\hat{\boldsymbol{X}}, \boldsymbol{X}^{(1)}, ..., \boldsymbol{X}^{(B)}) \in A_\alpha^B | \boldsymbol{Y}, \boldsymbol{Z} \right) \\
&\quad + \mathbb{P}_{\mathcal{H}_0}\left((\boldsymbol{X}, \boldsymbol{X}^{(1)}, ..., \boldsymbol{X}^{(B)}) \in A_\alpha^B | \boldsymbol{Y}, \boldsymbol{Z} \right) \\
&\quad - \mathbb{P}_{\mathcal{H}_0}\left((\hat{\boldsymbol{X}}, \boldsymbol{X}^{(1)}, ..., \boldsymbol{X}^{(B)}) \in A_\alpha^B | \boldsymbol{Y}, \boldsymbol{Z} \right) \\
&\leq \mathbb{P}_{\mathcal{H}_0}\left((\hat{\boldsymbol{X}}, \boldsymbol{X}^{(1)}, ..., \boldsymbol{X}^{(B)}) \in A_\alpha^B | \boldsymbol{Y}, \boldsymbol{Z} \right) \\
&\quad + d_{\text{TV}}\left\{ (\boldsymbol{X}, \boldsymbol{X}^{(1)}, ..., \boldsymbol{X}^{(B)} | \boldsymbol{Y}, \boldsymbol{Z}), (\hat{\boldsymbol{X}}, \boldsymbol{X}^{(1)}, ..., \boldsymbol{X}^{(B)} | \boldsymbol{Y}, \boldsymbol{Z}) \right\}.
\end{aligned} \tag{44}$$

By the definition of $\hat{\boldsymbol{X}}$, perform the same analysis as in Proposition D.1 and Corollary D.2, we can show that

$$\rho(\hat{\boldsymbol{X}}, \boldsymbol{Y}, \boldsymbol{Z}), \rho(\boldsymbol{X}^{(1)}, \boldsymbol{Y}, \boldsymbol{Z}), ..., \rho(\boldsymbol{X}^{(B)}, \boldsymbol{Y}, \boldsymbol{Z}) \tag{45}$$

are exchangeable conditionally on $\boldsymbol{Y}, \boldsymbol{Z}$. Hence by combining the property of rank test similar to Proposition D.3, we obtain that

$$\mathbb{P}_{\mathcal{H}_0}\left((\hat{\boldsymbol{X}}, \boldsymbol{X}^{(1)}, ..., \boldsymbol{X}^{(B)}) \in A_\alpha^B | \boldsymbol{Y}, \boldsymbol{Z} \right) \leq \alpha. \tag{46}$$

And by the definition of TV distance of probability measures, we can further calculating that

$$\begin{aligned}
&d_{\mathrm{TV}}\Big\{(\boldsymbol{X},\boldsymbol{X}^{(1)},...,\boldsymbol{X}^{(B)}|\boldsymbol{Y},\boldsymbol{Z}),(\hat{\boldsymbol{X}},\boldsymbol{X}^{(1)},...,\boldsymbol{X}^{(B)}|\boldsymbol{Y},\boldsymbol{Z})\Big\}\\
&=\frac{1}{2}\int\big|p_{\hat{\boldsymbol{X}},\boldsymbol{X}^{(1)},...,\boldsymbol{X}^{(B)}|\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{x},\boldsymbol{x}^{(1)},...,\boldsymbol{x}^{(B)})\\
&\qquad-p_{\boldsymbol{X},\boldsymbol{X}^{(1)},...,\boldsymbol{X}^{(B)}|\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{x},\boldsymbol{x}^{(1)},...,\boldsymbol{x}^{(B)})\big|d\boldsymbol{x}d\boldsymbol{x}^{(1)}\cdots d\boldsymbol{x}^{(B)}\\
&=\frac{1}{2}\int\Big|p_{\hat{\boldsymbol{X}}|\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{x})-p_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{Z}}(\boldsymbol{x})\Big|d\boldsymbol{x}\\
&=d_{\mathrm{TV}}\Big\{(\hat{\boldsymbol{X}}|\boldsymbol{Y},\boldsymbol{Z}),(\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{Z})\Big\},
\end{aligned}\tag{47}$$

where the calculation is based on the property that $(\boldsymbol{x},\boldsymbol{x}^{(1)},...,\boldsymbol{x}^{(B)})$ is independent of each other. As a result, we obtain the bound of Type I error rate as

$$\begin{aligned}
\mathbb{P}_{\mathcal{H}_0}(\hat{\rho}>c_\alpha|\boldsymbol{Y},\boldsymbol{Z})-\alpha &\le d_{\mathrm{TV}}\Big\{(\hat{\boldsymbol{X}}|\boldsymbol{Y},\boldsymbol{Z}),(\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{Z})\Big\}\\
&=d_{\mathrm{TV}}\{p_{T;N}(\cdot|\boldsymbol{Z}),p(\cdot|\boldsymbol{Z})\}\\
&\le\sum_{i=1}^{n}d_{\mathrm{TV}}\{p_{T;N}(\cdot|z_i),p(\cdot|z_i)\},
\end{aligned}\tag{48}$$

then by combining Theorem C.2, we complete the proof. □

Next, we show that the Type I error rate can be unconditionally controlled, as shown in Theorem E.2.

Theorem E.2 (Type I error bound). *Assume $\mathcal{H}_0: X \perp\!\!\!\perp Y|Z$ is true. Under all the Assumptions in Sec. B and C, for any signifiance level $\alpha\in(0,1)$, the bound for the Type I error is obtained as*

$$\mathbb{P}(p\text{-}value\le\alpha|\mathcal{H}_0)\le\alpha+o_p(1),\ \ as\ N\to\infty.\tag{49}$$

Proof. Applying Theorem E.1 and Lebesgue dominated convergence theorem, by marginalizing over $\boldsymbol{Y},\boldsymbol{Z}$ and note that the TV distance is upper bounded by 1, thus we have

$$\begin{aligned}
\mathbb{P}(p\text{-value}\le\alpha|\mathcal{H}_0)&\le\alpha+\mathbb{E}_{\boldsymbol{Z}}\big[d_{\mathrm{TV}}\{p_{T;N}(\cdot|\boldsymbol{Z}),p(\cdot|\boldsymbol{Z})\}\big]\\
&\le\alpha+o_p(1),\ \text{as}\ N\to\infty,
\end{aligned}\tag{50}$$

which completes the proof. □

Thus, these theories prove that our test is valid in the sense that asymptotically the Type I error can be well controlled, and more intuitively, when the training samples for the generative model are sufficiently large, the resulting sample distribution is sufficiently close to the true distribution, so that the upper bound of the Type I error rate is precisely controlled to any given α.

F Implementation Details.

F.1 Details of Compared Methods

The compared methods in our experiments are described below.

- **CCIT** [39]: Transforms the CI test into a classification problem, leveraging powerful classifiers such as gradient-boosted trees.
- **RCIT** [52]: Approximates the kernel-based CI test (KCIT) using random Fourier features for scalability.
- **FCIT** [7]: Performs a fast conditional independence test by comparing the mean squared errors (MSE) from regressing Y on X, Z, versus regressing Y on Z along.
- **GCM** [40]: Computes a normalized statistic for conditional independence testing based on the sample covariance between the regression residuals of X and Y given Z.
- **KCIT** [59]: A kernel-based CI test that constructs test statistics using kernel embeddings of the distributions.
- **LPCIT** [38]: Measures conditional dependence by evaluating differences between analytic kernel embeddings of distributions at a finite set of locations.
- **GCIT** [3]: Employs generative adversarial networks (GANs) to model conditional distributions for CI testing.
- **DGCIT** [42]: Uses two GANs to model the conditional distributions $\mathbb{P}(X|Z)$ and $\mathbb{P}(Y|Z)$, and designs a random mapping-based statistic using neural networks.
- **NNLSCIT** [25]: Integrates a classifier-based conditional mutual information estimator. A k-nearest-neighbor local sampling strategy is used to approximate the null hypothesis samples.

Below are the GitHub URLs of the compared methods:

- **CCIT** https://github.com/rajatsen91/CCIT.
- **RCIT**: https://github.com/ericstrobl/RCIT.
- **FCIT**: https://github.com/kjchalup/fcit.
- **GCM**: The R package is available.
- **KCIT**: http://people.tuebingen.mpg.de/kzhang/KCI-test.zip.
- **LPCIT**: https://github.com/meyerscetbon/lp-ci-test.
- **GCIT**: https://github.com/alexisbellot/GCIT.
- **DGCIT**: https://github.com/tianlinxu312/dgcit.
- **NNLSCIT**: https://github.com/LeeShuai-kenwitch/NNLSCIT.

F.2 Details of SGMCIT

We give the detailed implementation of SGMCIT as follows.
Model Architecture. The conditional score model is based on a multi-layer perceptron (MLP) with three fully connected layers, each followed by Swish activations. The input size is $d_x + d_z$, while the output size is d_x, with a hidden layer dimension of 64.
Projection Vectors. We set the distribution of the projection vectors to be Gaussian and set the projection number $m = 1$.
Hyperparameters. The model is trained using the Adam optimizer with a learning rate of 1×10^{-4} over 100 epochs. The batchsize is set to 50 by default.
Data Preprocessing. To normalize the input data, we apply min-max scaling, transforming all features into the $[0, 1]$ range. Following this, a logit transformation is applied: $\log(x) - \log(1 - x)$.
Sampling Procedure. The Langevin dynamics conditional sampling process is governed by the step size h and the total number of steps. In this work, we set the step size h to 0.1 and the number of steps to 200. For the samples used to model the null hypothesis distribution, we take a parallel implementation, i.e., we generate B samples in parallel, which greatly reduces the time required to generate the samples utilizing parallel hardware.
Code. For more details, our code is available at https://github.com/jinchenghou123/SGMCIT.

G More Experiments Results

In this section, we provide more experimental results, including some visualization results, as well as results under two additional metrics, with the runtime results.

G.1 Detailed Visualization Results

In this section, we give the detailed visualization results of generative model based methods GCIT, DGCIT and SGMCIT (Ours). The results corresponding to benchmark datasets and high-dimensional confounder setting in the main paper.

Visualization results on Benchmarks. Figs. 7, 8, and 9 demonstrate the performance of SGMCIT, GCIT, and DGCIT in estimating the marginal distribution under different transformation functions: linear, square, cos, tanh, and exp functions. SGMCIT consistently performs well across all settings, with the marginal distributions of the generated samples closely matching those of the observed data. In contrast, both GCIT and DGCIT struggle with highly non-linear transformations, such as cosine and exponential functions.

Visualization results of high-dimensional confounder setting. Fig. 10 compares the performance of GCIT, DGCIT, and SGMCIT in the high-dimensional confounder setting. In this case, both SGMCIT and DGCIT handle the high-dimensional setting effectively, outperforming GCIT. A further analysis of the approximation performance across different regions of the probability density shows that SGMCIT provides accurate approximations in various density regions. In contrast, while DGCIT yields similar overall distribution, it struggles with local accuracy, reflecting its inability to model the conditional distribution in certain areas.

Overall, SGMCIT outperforms both GCIT and DGCIT in most scenarios, highlighting its ability to model complex distribution. Also, the visualization results provide interpretability for the performance of the CI testing in corresponding experimental results.

G.2 Results under Additional Metrics

In this section, we present experimental results using two additional metrics. For context, the main paper focuses on Type I and Type II errors, here we introduce a total of four evaluation metrics.

Performance metrics. We assess performance using four metrics: (1) Type I error rate, which measures validity by ensuring the error rate remains controlled at any significance level α; (2) Testing power, defined as 1− Type II error rate, reflecting the ability to detect conditional dependencies; (3) Kolmogorov-Smirnov (KS) statistic, which under $\mathcal{H}_0$ compares the p-value distribution to a uniform [0,1], with smaller values indicating better uniformity; and (4) Area under the power curve (AUPC), which measures the empirical cumulative distribution of p-values under $\mathcal{H}_1$, with values closer to one indicating higher power.

Results and analysis. The results for Cases 1 and 2 are shown in Fig. 11, while those for Cases 3 and 4 are shown in Fig. 12.

Across all metrics, SGMCIT excels at controlling Type I errors while maintaining high testing power across a variety of function combinations, establishing it as the most reliable method in these experiments. The KS statistic for SGMCIT demonstrates good uniformity of the p-value distribution across a large number of function combination settings, reflecting its effective modeling of the conditional distribution. The AUPC results align closely with the power results, further showcasing SGMCIT's high power. In comparison, while most other methods perform well with the linear and tanh functions, they struggle with some other settings. For instance, DGCIT often fails to control Type I errors effectively, CCIT shows weak performance in terms of testing power, and GCIT exhibits poor p-value uniformity.

G.3 Experimental Results of Running Time

This section presents the results of running time for each method. All experiments are performed on the same device. The runtime for a single test is reported in a high-dimensional confounder setting with standard Gaussian noise.

Results. Fig. 13 shows the performance of all methods. When varying the sample size, we fixed $d_z = 100$, while for varying dimensionality, we set the sample size to 1000.

Analysis. SGMCIT, RCIT, GCM, and GCIT exhibit consistently low runtime, demonstrating strong scalability with respect to sample size. KCIT and LPCIT stand out with significantly longer runtime as the sample size increases. For example, KCIT exceeds 300 seconds for 10,000 samples, while LPCIT approaches 500 seconds.

SGMCIT and GCIT maintain low and stable runtime across all dimensions, demonstrating their efficiency in high-dimensional settings. This can be attributed to the full utilization of parallel hardware by the generative model. DGCIT, while also utilizing generative models, has a longer overall runtime due to the multiple models that need to be trained as well as ineffective statistic design. Notably, LPCIT shows exponential growth in runtime, becoming the slowest method as the dimensionality exceeds 60. For instance, at the 100-dimensional setting, LPCIT's runtime exceeds 200 seconds.

These results demonstrate that SGMCIT is computationally efficient, handling both large sample sizes and high-dimensional conditioning sets effectively. Among generative model-based CI methods, SGMCIT performs the best in terms of runtime efficiency.

G.4 Additional Baseline Results on Real Data

We evaluated all baseline methods on a real-world dataset, using the same experimental setup as in previous sections with a test sample size of 1000. Results are summarized in Table 1.

Table 1: The CI results of 10 methods on real datasets.

Method	p-value
CCIT	0.68
RCIT	0.00
FCIT	0.03
GCM	0.00
KCIT	0.00
LPCIT	0.00
GCIT	0.00
DGCIT	0.00
NNLSCIT	0.27
SGMCIT	0.00

Analysis. Although the ground truth is unknown, most methods—including our proposed SGMCIT—reject the null hypothesis, indicating that X and Y are not conditionally independent given Z. This aligns with our model's conclusion. In contrast, CCIT and NNLSCIT produce higher p-values. However, their poor power in synthetic experiments, where the ground truth is known, suggests that these results may be less reliable.

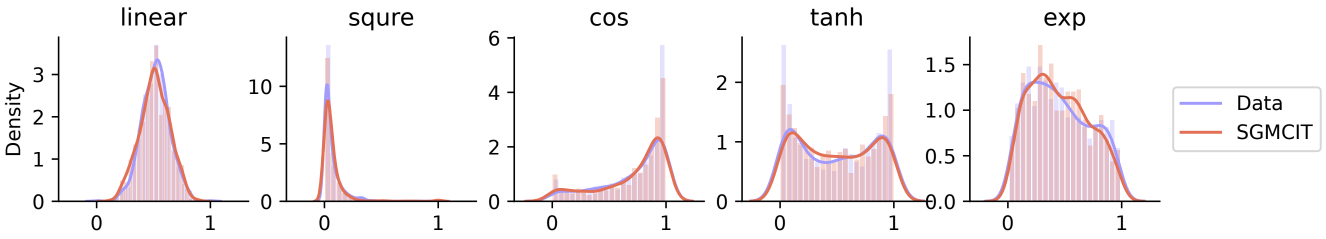


Figure 7: The visualization results of SGMCIT for the marginal distribution of X under Case 4 of benchmark datasets.

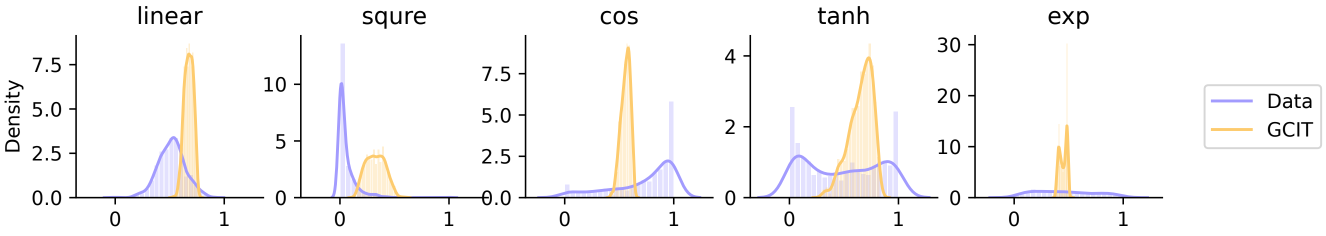


Figure 8: The visualization results of GCIT for the marginal distribution of X under Case 4 of benchmark datasets.

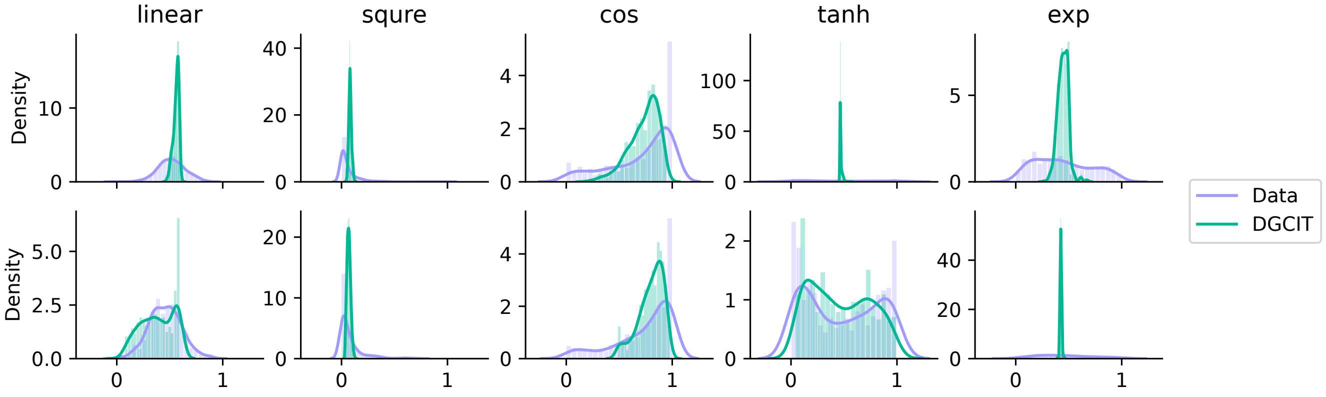


Figure 9: The visualization results of DGCIT for the marginal distribution of X and Y under Case 4. Top row: the results of X. Below row: the results of Y.

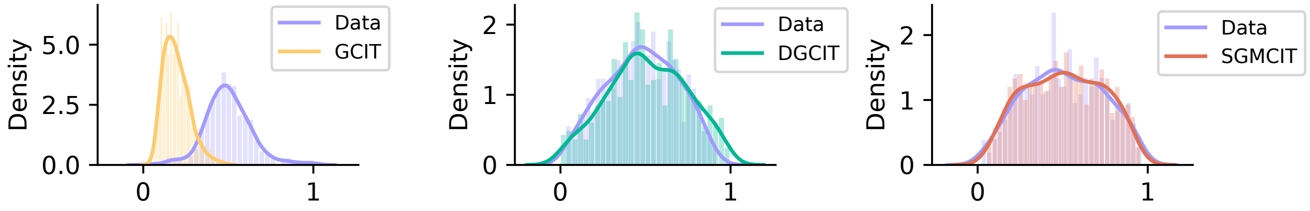


Figure 10: The visualization results of GCIT, DGCIT and SGMCIT for the marginal distribution of X under high-dimensional confounder setting.

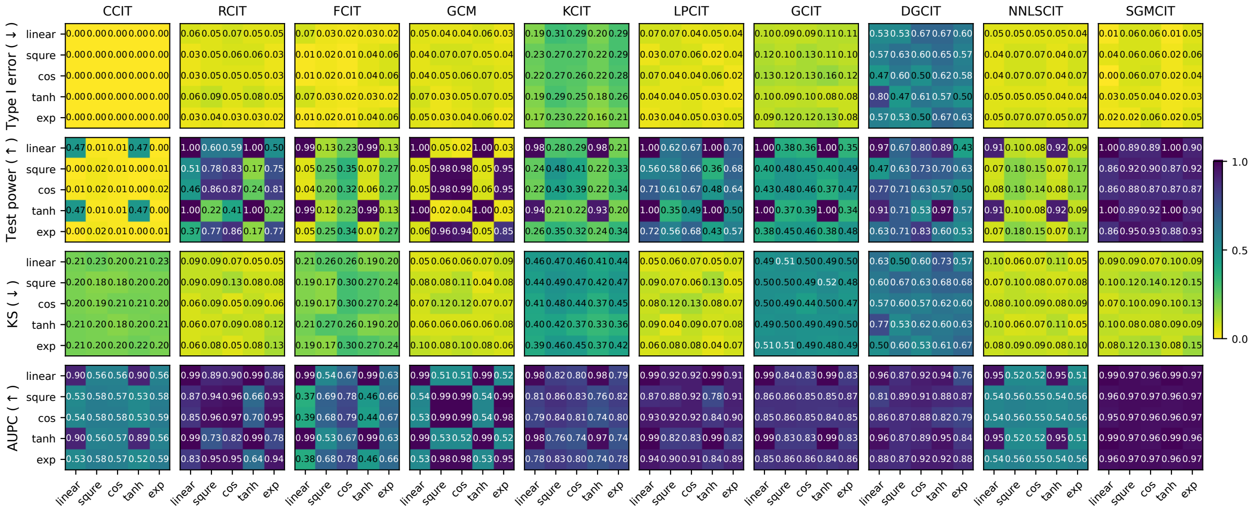


Figure 11: Additional results of conditional independence tests for Cases 1 and 2 on benchmark datasets.

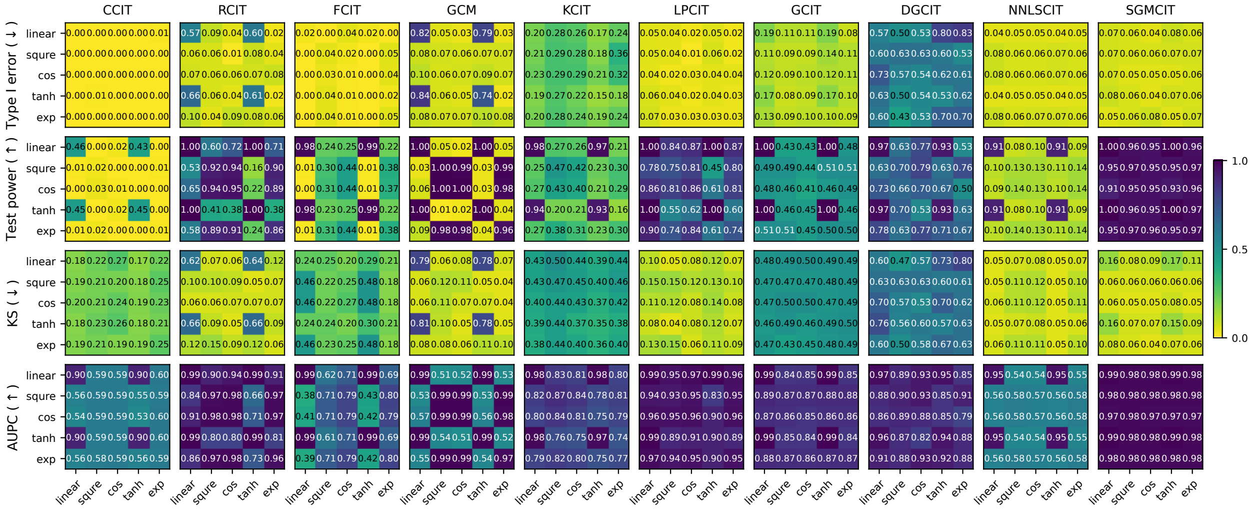


Figure 12: Additional results of conditional independence tests for Cases 3 and 4 on benchmark datasets.

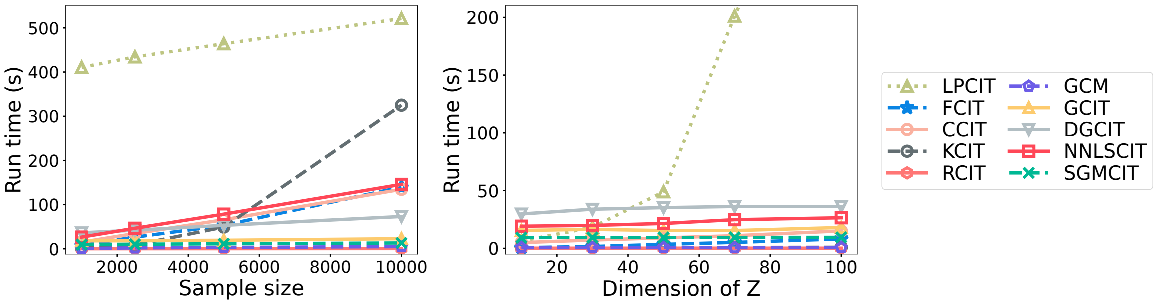


Figure 13: Results of running time. Left: The results w.r.t. the sample size. Right: The results w.r.t. the dimension of Z.

G.5 Comparison with CDCIT [57]

As noted in Sec. 8, we provide a thorough comparison between our proposed method, SGMCIT, and the contemporaneous work CDCIT [57]. Unless otherwise specified, all experiments follow the same setup described in the main paper. This comparison includes evaluations on benchmark datasets, high-dimensional settings, and real-world data. Furthermore, we analyze each method's ability to estimate marginal distributions and compare their computational efficiency. CDCIT is implemented using default settings.

G.5.1 Results on benchmark datasets. We begin by comparing SGMCIT and CDCIT on four standard benchmark cases. In Fig. 14 (a), we report results on Cases 1 and 2, while Fig. 14 (b) presents results for Cases 3 and 4. These cases encompass a range of functional forms and dependency strengths, designed to systematically evaluate both Type I error control and statistical power. Also, we present CDCIT's visualization results under different transformation functions: linear, square, cos, tanh, and exp functions in Fig. 15.
Analysis. Our results show that CDCIT struggles in multiple aspects. It fails to control the Type I error in several settings, which undermines its reliability as a statistical test. More critically, its test power remains consistently low—even when the conditional dependency between variables is strong and should be easily detectable. Additionally, CDCIT often fails to accurately estimate the marginal distribution, particularly under nonlinear transformations such as cos, tanh, or exp. Even in relatively simple cases—such as a linear transformation—its performance is at best moderate. These findings suggest that the conditional diffusion model employed by CDCIT has difficulty modeling complex distributions.

In contrast, SGMCIT consistently performs well across all benchmark cases. It not only achieves strong Type I error control, but also maintains high test power across a variety of functional forms. The generative component of SGMCIT produces accurate marginal estimates even in challenging scenarios, highlighting its effectiveness in modeling complex dependencies and distributions.

G.5.2 Results on High-Dimensional setting. We further examine performance under high-dimensional confounding setting ($d_z = 100$). The results are provided in Fig. 14 (c).
Analysis. It can be observed that CDCIT performs poorly in high-dimensional settings, CDCIT exhibits consistently low test power, failing to detect dependencies on conditioned on Z even when they are pronounced. For example, in the case where Z is high-dimensional and the strength of dependence is strong (i.e., $b = 0.6$), CDCIT still yields unsatisfactory results. This suggests that CDCIT may suffer from an inherent inability to capture intricate conditional relationships in high-dimensional scenarios.

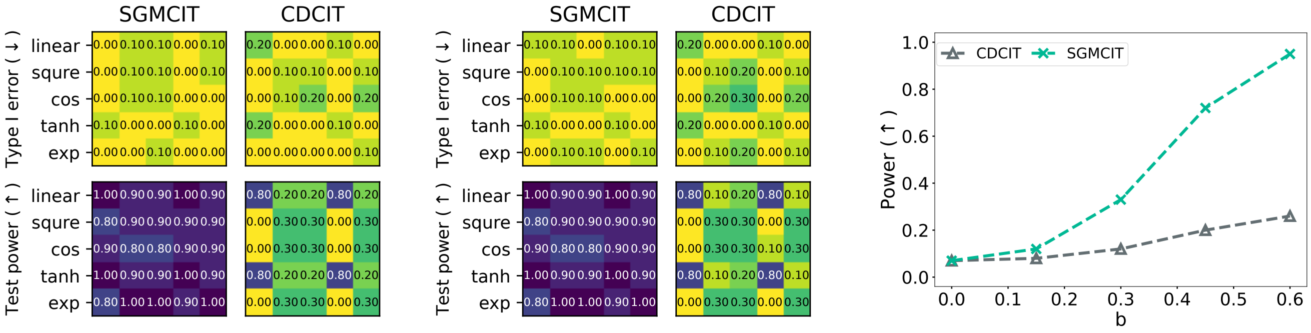
In contrast, our proposed method SGMCIT maintains excellent performance even under these challenging conditions. It achieves both strong Type I error control and high test power, demonstrating robust behavior regardless of the dimensionality of the input variables. These results highlight the advantage of our method in practical applications where high-dimensional data is common and effective CI testing is critical.

G.5.3 Visualization Results on Real Data. To further assess CDCIT's generative capacity in practical scenarios, we visualize its estimated marginal distributions on a real-world dataset. The results are presented in Fig. 16.
Analysis. To compensate for this, we provided CDCIT with a significantly large sample size ($n = 50,000$) during training on real-world datasets. However, even with this increased availability of data, the estimated marginal distributions remained inaccurate, demonstrating that simply increasing the sample size is insufficient to overcome the inherent limitations of method.

In contrast, SGMCIT achieves a highly accurate marginal distribution estimation. This highlights not only its modeling capacity but also its efficiency in data usage.

G.5.4 Running Time Evaluation. We evaluate the computational efficiency of CDCIT compared to our methods. The timing results are shown in Fig. 17.
Analysis. CDCIT is computationally expensive. Even in favorable conditions with low sample size ($n = 1000$) and moderate dimensionality ($d_z = 10$), a single run of CDCIT takes nearly 40 seconds. This is significantly slower than most other methods evaluated. Such runtime requirements may render CDCIT impractical for large-scale or time-sensitive applications.

Our method, SGMCIT, on the other hand, is far more efficient. It achieves faster execution while maintaining high statistical performance, making it well-suited for real-world tasks where both accuracy and speed are essential.



(a)

(b)

(c)

Figure 14: Additional results of CI tests. (a) Additional results of CI tests for Cases 1 and 2 on benchmark datasets. (b) Additional results of CI tests for Cases 3 and 4 on benchmark datasets. (c) Results in the high-dimensional confounder setting.

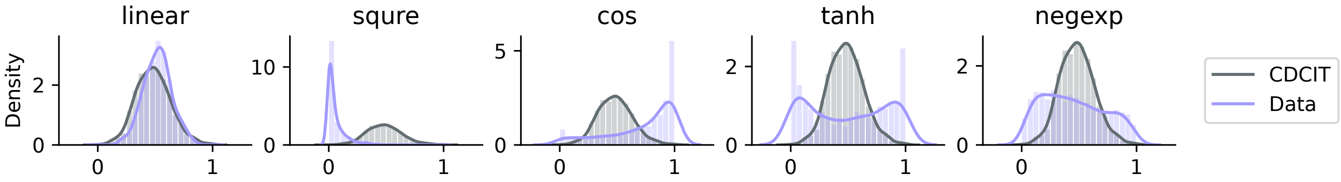


Figure 15: The visualization results of SGMCIT and CDCIT for the distribution of X under Case 4 of benchmark datasets.

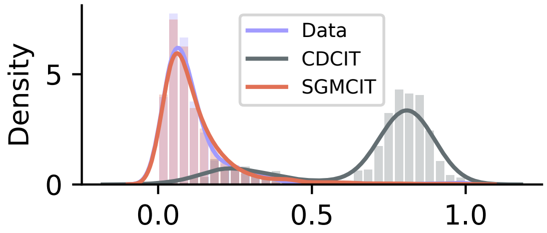


Figure 16: Visualization results of SGMCIT and CDCIT on real data with 50000 sample size.

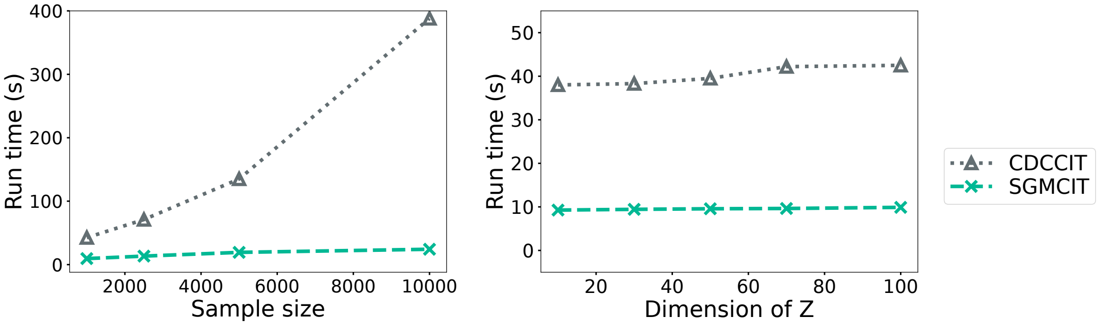


Figure 17: Results of running time of CDCIT and SGMCIT (Ours). Left: The results w.r.t. the sample size. Right: The results w.r.t. the dimension of Z.