DOMAIN-ADAPTED DIFFUSION MODELS FOR CONDITIONAL INDEPENDENCE TESTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Conditional independence (CI) is a fundamental concept in statistics and machine learning. Recent advances in conditional generative modeling provide flexible tools for generative-model-based CI tests, which rely on an estimated conditional distribution to generate randomized samples. However, errors in estimating this distribution accumulate in existing Type I error bounds, and consistency of the generative estimator alone does not guarantee asymptotic Type I error control. To address this limitation, we formulate conditional generative modeling as a domain adaptation problem and leverage auxiliary data from multiple source domains to improve estimation in the target CI testing domain. We propose Domain-Adapted Diffusion (DA-Diff), a multi-source domain adaptation framework for conditional diffusion models based on weighted empirical risk minimization over both target and source domains. We establish the convergence rate of DA-Diff and show how transferable source data can improve target-domain estimation through an increased effective sample size while controlling transfer bias. Building on DA-Diff, we further propose Domain-Adapted Conditional Independence Testing (DA-CIT) and show that its Type I error satisfies $P(p \leq \alpha) \leq \alpha + o(1)$. Experiments demonstrate that DA-Diff improved conditional generation quality compared with transfer-learning diffusion baselines, while DA-CIT provides strong Type I error control and competitive power.

1 INTRODUCTION

Conditional independence (CI) is a fundamental concept in statistics and machine learning, with broad applications to variable selection, causal structure learning, and sufficient dimension reduction (Pearl, 2009; Spirtes et al., 2000; Cook, 2026; Yang et al., 2025a; He et al., 2025; Wang et al., 2025; 2026; Zhang et al., 2026b). Let $X_0 \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$, $Y_0 \in \mathcal{Y} \subseteq \mathbb{R}^{d_y}$, and $Z_0 \in \mathcal{Z} \subseteq \mathbb{R}^{d_z}$ be random variables. CI testing considers the hypotheses

$$H_0 : X_0 \perp\!\!\!\perp Y_0 | Z_0 \quad \text{versus} \quad H_1 : X_0 \not\perp\!\!\!\perp Y_0 | Z_0. \tag{1}$$

A broad range of CI tests has been developed, including kernel-based, regression-based, and metric-based approaches (Fukumizu et al., 2007; Zhang et al., 2011; Shah & Peters, 2020; Azadkia & Chatterjee, 2021; Banerjee et al., 2026). More recently, the conditional randomization test (CRT) (Candès et al., 2018) has become an important framework for CI testing (Bellot & van der Schaar, 2019; Liu et al., 2022; Li & Liu, 2023; Li et al., 2024; Yang et al., 2025a; Ren et al., 2025; Zhao et al., 2026). A key advantage of the CRT is that, given a correctly specified conditional distribution $P_0(\cdot|Z_0)$, it yields valid inference while allowing flexible choices of test statistics (Berrett et al., 2020). When $P_0(\cdot|Z_0)$ is estimated from data, however, the validity of the resulting test depends critically on its estimation accuracy.

Recent advances in generative modeling provide flexible tools for estimating $P_0(\cdot|Z_0)$ (Goodfellow et al., 2014; Ho et al., 2020; Song et al., 2021; Karras et al., 2022; Fu et al., 2024), leading to several generative-model-based CI tests (Bellot & van der Schaar, 2019; Shi et al., 2021; Ren et al., 2025; Zhao et al., 2026). Among generative models, diffusion models are particularly attractive due to their strong generation quality and have also been shown to perform well for CRT (Yang et al., 2025a). Despite these advantages, a central difficulty is that errors in conditional distribution estimation directly affect the validity of the resulting CRT. To make this dependence explicit, suppose that the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

target dataset contains $2n_0$ independent observations, with n_0 observations used for the CRT and the remaining n_0 observations used to estimate $P_0(\cdot|Z_0)$. Let $\widehat{P}_0(\cdot|Z_0)$ denote the estimated conditional distribution and let p be the resulting CRT p-value. By Theorem 4 of Berrett et al. (2020), under H_0,

$$\mathbb{P}(p \le \alpha) \le \alpha + \mathcal{O}\left(n_0 \cdot \mathbb{E}\left[d_{\mathrm{TV}}\left(P_0(\cdot|Z_0), \widehat{P}_0(\cdot|Z_0)\right)\right]\right). \tag{2}$$

Here, $d_{\mathrm{TV}}\Big(P_0(\cdot|Z_0), \widehat{P}_0(\cdot|Z_0)\Big)$ denotes the total variation (TV) distance between the true and estimated conditional distributions of X_0 given Z_0. The second term in (2) quantifies the excess Type I error caused by conditional distribution estimation. In particular, (2) guarantees asymptotic Type I error control when $\mathbb{E}\left[d_{\mathrm{TV}}\left(P_0(\cdot|Z_0), \widehat{P}_0(\cdot|Z_0)\right)\right] = o(n_0^{-1})$. Note that this is only a sufficient condition, rather than a necessary condition for asymptotic validity. Nevertheless, this sufficient condition is particularly stringent. For perspective, even regular parametric estimators typically achieve only $\mathcal{O}(n_0^{-1/2})$ (Vaart, 1998), while flexible nonparametric conditional generative models generally converge more slowly and their rates depend on d_x and d_z. Thus, when $\widehat{P}_0(\cdot|Z_0)$ is learned only from the n_0 target observations reserved for conditional distribution estimation, existing convergence rates may be insufficient to make the excess term in (2) vanish.

Therefore, a natural strategy is to increase the amount of data available for conditional distribution estimation. Existing theory for conditional diffusion models shows that estimation accuracy improves with the growth of sample size (Fu et al., 2024; Yang et al., 2025a), motivating the use of auxiliary data to improve estimation of $P_0(\cdot|Z_0)$. This argument is true when the auxiliary observations and the target observations follow the same distribution. However, in practice, auxiliary data often come from heterogeneous source domains rather than from the target distribution itself. Naively pooling such observations can introduce transfer bias that offsets the benefit of the additional samples. This raises the central question of how to selectively leverage multiple heterogeneous source domains to improve estimation of $P_0(\cdot|Z_0)$ while controlling the bias induced by source–target distribution shifts.

To solve this question, we propose the **Domain-Adapted Diffusion Model (DA-Diff)**, a multi-source domain adaptation framework for conditional diffusion models. DA-Diff learns a target-oriented conditional score model by minimizing a weighted empirical score-matching loss over the target domain and other source domains, where the source-domain weights determine how much information is transferred from each source. Our theoretical analysis reveals an explicit estimation–transfer trade-off: incorporating informative source observations increases the effective sample size and reduces estimation error, whereas discrepancies between source and target conditional scores induce transfer bias. This decomposition provides the theoretical basis for our source-weighting criterion, which favors the statistical benefit of informative sources while penalizing source–target discrepancy. We further translate the target-domain score-matching risk bound into a TV error bound for the learned conditional distribution. Building on this result, we develop the **Domain-Adapted Conditional Independence Test (DA-CIT)**, which uses DA-Diff to generate the randomized samples required by the CRT. Under suitable source–target transferability conditions, we establish asymptotic Type I error control for DA-CIT.

Our contributions are summarized as follows:

1. We propose the **Domain-Adapted Diffusion Model (DA-Diff)**, the first transfer-learning framework for diffusion models based on weighted ERM under a MSDA setting. We establish its theoretical convergence rate in TV distance, providing convergence guarantees for diffusion models trained by jointly leveraging multiple heterogeneous source domains.
2. Building on DA-Diff, we develop the **Domain-Adapted Conditional Independence Test (DA-CIT)**, which first incorporates heterogeneous external data into generative model-based CIT. We further establish theoretical conditions under which the excess Type I error of DA-CIT vanishes asymptotically, i.e., is controlled at the rate $o(1)$.
3. We conduct synthetic experiments to evaluate both components of the proposed framework. The DA-Diff experiments assess conditional-generation accuracy under varying numbers of heterogeneous source domains and compare against various baselines. The DA-CIT experiments evaluate both Type I error control and test power under nonlinear data-generating mechanisms. We further evaluate DA-CIT on the multi-source Sachs dataset to demonstrate its performance on real-world data.

2 PRELIMINARY

2.1 CONDITIONAL RANDOMIZATION TESTS (CRT)

Conditional randomization tests (CRT) is a general framework for testing CI (Candès et al., 2018). Let $D_0 = \{(x_{0,i}, y_{0,i}, z_{0,i})\}_{i=1}^{2n_0}$ denote the target-domain dataset. We reserve the first n_0 observations for testing and use the remaining n_0 observations exclusively for conditional distribution estimation. Suppose that an estimator $\widehat{P}_0(\cdot|Z_0)$ of the conditional distribution $P_0(\cdot|Z_0)$ is available. Conditioning on $\mathbf{Z}_0 := (z_{0,1}, \ldots, z_{0,n_0})^\top \in \mathbb{R}^{n_0 \times d_z}$, we independently generate B pseudo-samples from $\prod_{i=1}^{n_0} \widehat{P}_0(\cdot|z_{0,i})$, yielding $\mathbf{X}_0^{(b)} \in \mathbb{R}^{n_0 \times d_x}$, $b = 1, \ldots, B$. Let $\mathbb{T}(\cdot,\cdot,\cdot) : \mathbb{R}^{n_0 \times d_x} \times \mathbb{R}^{n_0 \times d_y} \times \mathbb{R}^{n_0 \times d_z} \to \mathbb{R}$ be a pre-specified test statistic, with larger values indicating stronger evidence against H_0. Unless otherwise specified, we use distance correlation (dCor) (Székely et al., 2007) as $\mathbb{T}$ in the proposed DA-CIT. Define $\mathbf{X}_0 = (x_{0,1}, \ldots, x_{0,n_0})^\top \in \mathbb{R}^{n_0 \times d_x}$ and $\mathbf{Y}_0 = (y_{0,1}, \ldots, y_{0,n_0})^\top \in \mathbb{R}^{n_0 \times d_y}$. The test statistic computed from the observed data is $\mathbb{T}^{(0)} := \mathbb{T}(\mathbf{X}_0, \mathbf{Y}_0, \mathbf{Z}_0)$, while for each pseudo-sample we compute $\mathbb{T}^{(b)} := \mathbb{T}(\mathbf{X}_0^{(b)}, \mathbf{Y}_0, \mathbf{Z}_0)$, $b = 1, \ldots, B$. The CRT compares the observed statistic $\mathbb{T}^{(0)}$ with the randomized statistics $\{\mathbb{T}^{(b)}\}_{b=1}^B$ and computes the p-value as

$$p = \frac{1 + \sum_{b=1}^B \mathbf{1}\{\mathbb{T}^{(b)} \geq \mathbb{T}^{(0)}\}}{B+1}. \tag{3}$$

Under H_0 in (1), if $\widehat{P}_0(\cdot|Z_0) = P_0(\cdot|Z_0)$, the observed statistic $\mathbb{T}^{(0)}$ and the randomized statistics $\{\mathbb{T}^{(b)}\}_{b=1}^B$ are exchangeable, and the CRT provides finite sample Type I error control, $\mathbb{P}(p \leq \alpha) \leq \alpha$ (Berrett et al., 2020). When $\widehat{P}_0(\cdot|Z_0)$ is an approximation to $P_0(\cdot|Z_0)$, the Type I error is no longer guaranteed to be bounded exactly by α; instead, the excess Type I error can be controlled by the estimation error of the conditional distribution, as shown in (2). Under H_1, $\mathbb{T}^{(0)}$ tends to be larger than its randomized counterparts, leading to smaller p-values and evidence against H_0.

2.2 DIFFUSION MODEL

Diffusion models are a class of generative models that can flexibly approximate complex conditional distributions such as $P_0(\cdot|z)$ (Song et al., 2021; Ho et al., 2020; Fu et al., 2024). A diffusion model consists of a forward diffusion process and a reverse process. Conditioned on $Z_0 = z$, let $X_0(0) \sim P_0(\cdot|z)$. The forward process is defined by the stochastic differential equation (SDE)

$$dX_0(t) = -\frac{1}{2} X_0(t)\, dt + dB(t), \qquad t \in [0, T], \tag{4}$$

where $B(t)$ is a standard Brownian motion and $T > 0$ denotes the terminal time. By the properties of the Ornstein–Uhlenbeck (OU) process, $X_0(t)|X_0(0) \sim \mathcal{N}\left(a_t X_0(0), \sigma_t^2 I_{d_x}\right)$, where $a_t = \exp(-t/2)$ and $\sigma_t = \sqrt{1 - \exp(-t)}$. As T becomes sufficiently large, the marginal distribution of $X_0(T)$ approaches $\mathcal{N}(0, I_{d_x})$.

The reverse process transforms samples from $N(0, I_{d_x})$ back to the data distribution $P_0(\cdot|z)$. Specifically, it is characterized by the reverse-time SDE

$$dX_0(t) = \left[-\frac{1}{2} X_0(t) - \nabla_x \log p_{0,t}\big(X_0(t)|z\big)\right] dt + d\overline{B}(t), \qquad X_0(T) \sim N(0, I_{d_x}), \tag{5}$$

which is simulated backward in time from $t = T$ to an early stopping time $t_0 > 0$. Here, $p_{0,t}(\cdot|z)$ denotes the conditional density of $X_0(t)$ given $Z_0 = z$, and $\overline{B}(t)$ is a Brownian motion associated with the reverse process. The conditional score function $\nabla_x \log p_{0,t}(\cdot|z)$ appearing in (5) is generally unknown and must therefore be estimated from data. We estimate it using denoising score matching (Vincent, 2011; Song et al., 2021). Let $s : \mathbb{R}^{d_x} \times \mathbb{R}^{d_z} \times \mathbb{R} \to \mathbb{R}^{d_x}$ be a score network belonging to a ReLU network class $\mathcal{F}$ (defined in (83)). For a single observation $(x, z) \in \mathcal{X} \times \mathcal{Z}$, define the denoising score-matching loss as

$$\ell(x, z; s) = \frac{1}{T - t_0} \int_{t_0}^T \mathbb{E}_{X_t \sim \mathcal{N}(a_t x, \sigma_t^2 I_{d_x})} \left[\left\| s(X_t, z, t) + (X_t - a_t x)/\sigma_t^2 \right\|_2^2\right] dt. \tag{6}$$

We use the last n_0 samples in D_0, i.e. $\{(x_{0,i}, y_{0,i}, z_{0,i})\}_{i=n_0+1}^{2n_0}$, to estimate the target conditional distribution. The empirical score-matching loss is

$$\widehat{\mathcal{L}}_0(s) = \frac{1}{n_0} \sum_{i=n_0+1}^{2n_0} \ell(x_{0,i}, z_{0,i}; s). \tag{7}$$

The score network is then estimated by empirical risk minimization (ERM):

$$\widehat{s}_0 = \arg\min_{s \in \mathcal{F}} \widehat{\mathcal{L}}_0(s). \tag{8}$$

Replacing the unknown $\nabla_x \log p_{0,t}(X_0(t)|z)$ in (5) with $\widehat{s}_0(X_0(t), z, t)$ yields a learned reverse SDE. Starting from $X_0(T) \sim \mathcal{N}(0, I_{d_x})$ and solving this SDE backward from T to t_0 produces conditional samples from the learned distribution, which approximates $P_0(\cdot|Z_0)$ when T is sufficiently large and t_0 is sufficiently small.

2.3 Domain adaptation through weighted empirical risk minimization (ERM)

We regard the dataset D_0 used for the CRT in Section 2.1 as the target-domain dataset. In addition, we assume access to K source domains, with datasets $D_k = \{(x_{k,i}, z_{k,i})\}_{i=1}^{n_k}, k = 1, \ldots, K$, of sample sizes $n_1, \ldots, n_K$, respectively. Compared with training a model solely on the limited target-domain data, MSDA can exploit informative samples from multiple source domains to improve target-domain performance and potentially accelerate statistical convergence (Ben-David et al., 2006; 2010; Farahani et al., 2021). Among existing domain adaptation (DA) approaches, weighted ERM provides a simple and widely used framework for balancing information from heterogeneous domains (Konstantinov & Lampert, 2019; Vogel et al., 2020; Zhang et al., 2026a).

For each domain $k \in \{0, 1, \ldots, K\}$, let $(X_k, Z_k) \overset{\text{i.i.d.}}{\sim} P_k$ denote random variables, where each P_k is a joint distribution on $\mathcal{X} \times \mathcal{Z}$ with density $p_k(x, z)$ and factorization: $P_k(x, z) = P_k(x|z)P_k(z)$ and $p_k(x, z) = p_k(x|z)p_k(z)$. Here, $P_k(\cdot|z)$ and $p_k(\cdot|z)$ denote the conditional distribution and density of X_k given $Z_k = z$, respectively, while $P_k(z)$ and $p_k(z)$ denote the marginal distribution and density of Z_k. In particular, $P_0(\cdot|z)$ coincides with the target conditional distribution introduced in Section 2.2. Given D_k, we define the empirical diffusion loss on the k-th domain as

$$\widehat{\mathcal{L}}_k(s) := \frac{1}{n_k} \sum_{i=1}^{n_k} \ell(x_{k,i}, z_{k,i}; s), \quad k = 1, \ldots, K. \tag{9}$$

Inspired by Zhang et al. (2026a), we consider the weighted ERM objective

$$\widehat{\mathcal{L}}_w(s) := \frac{n_0 \widehat{\mathcal{L}}_0(s) + \sum_{k=1}^K w_k n_k \widehat{\mathcal{L}}_k(s)}{n_0 + \sum_{k=1}^K w_k n_k}, \tag{10}$$

where $w_k \geq 0$ denotes the weight assigned to the k-th source domain. The factors n_k account explicitly for the quantities of information contributed by each domain, while the weights w_k control the extent to which each source domain is incorporated into training. For a given weight vector $w = (w_1, \ldots, w_K)^\top$, the resulting weighted-ERM estimator is

$$\widehat{s}_w = \arg\min_{s \in \mathcal{F}} \widehat{\mathcal{L}}_w(s). \tag{11}$$

To characterize the target-domain performance of $\widehat{s}_w$, we further define the population loss on domain k as

$$\mathcal{L}_k(s) := \mathbb{E}_{(X_k, Z_k) \sim P_k} \big[\ell(X_k, Z_k; s)\big], \quad k = 0, \ldots, K. \tag{12}$$

Our goal is to choose the source-domain weights so that the resulting estimator achieves the smallest expected population loss on the target domain. Accordingly, we define the optimal weights by

$$(w_1^*, \ldots, w_K^*)^\top = \arg\min_{\{w_k \geq 0\}_{k=1}^K} \mathbb{E}_{D_0, \ldots, D_K} \left[\mathcal{L}_0(\widehat{s}_w)\right]. \tag{13}$$

Let $w^* := (w_1^*, \ldots, w_K^*)^\top$. Substituting w^* into (10) yields the final weighted-ERM estimator

$$\widehat{s}_{w^*} = \arg\min_{s \in \mathcal{F}} \widehat{\mathcal{L}}_{w^*}(s). \tag{14}$$

The role of the weights is to balance the statistical benefit of incorporating additional source samples against the bias induced by discrepancies between the source and target distributions. As shown in the theoretical analysis in Section 3, under suitable conditions, informative source domains can improve the target-domain convergence rate relative to training solely on the target-domain data. The detailed optimization method of (13) can be found in Appendix C.

3 THEORETICAL RESULTS

In this section, we establish theoretical guarantees for DA-Diff and DA-CIT. Theorem 1 bounds the target-domain score-matching risk of DA-Diff, Theorem 2 translates this result into a TV error bound for the learned conditional distribution, and Theorem 3 further establishes asymptotic Type I error control for DA-CIT. Together, these results connect domain-adapted score matching to distributional approximation and, ultimately, valid CI testing. Complete proofs are provided in Appendix K. The complete algorithms of training DA-Diff and DA-CIT can be found in Algorithm 1 and 2.

The proof is organized as follows. We first introduce the Hölder function class and the ReLU network class at the beginning of Appendix K. We then characterize the global population minimizer of the weighted score-matching risk in Appendix K.1, followed by the excess-risk decomposition in Appendix K.2 and the properties of the mixture density $p_{w,t}$ in Appendix K.3. The transfer bias and estimation error are bounded separately in Appendix K.4 and Appendix K.5, respectively, and are combined to prove Theorem 1 in Appendix K.5.6. The TV bound in Theorem 2 is proved in Appendix K.6, and the Type I error result in Theorem 3 is proved in Appendix K.7. The main new ingredient is the extension of diffusion-model convergence analysis to weighted ERM over multiple domains, which requires characterizing the mixture distribution and separately controlling the source-induced transfer bias and the weighted estimation error.

3.1 DOMAIN-ADAPTED DIFFUSION MODEL (DA-DIFF)

In this section, we provide a bound on the score matching risk of DA-Diff trained via weighted ERM, together with the corresponding bound on the TV distance to the target domain.

For each domain $k \in \{0, 1, \ldots, K\}$, we consider the same forward process as in (4), initialized by $X_k(0)|Z_k = z \sim P_k(\cdot|z)$. Let $p_{k,t}(\cdot|z)$ denote the conditional density of $X_k(t)$ given $Z_k = z$, and define the (intractable) conditional score function as:

$$s_k^*(x, z, t) := \nabla_x \log p_{k,t}(x|z), \qquad k = 0, 1, \ldots, K. \tag{15}$$

Define the joint density of $(X_k(t), Z_k)$ by $p_{k,t}(x, z) := p_{k,t}(x|z)\, p_k(z)$. To distinguish from the denoising score matching loss in (6), we define the score matching risk on the k-th domain as:

$$\mathcal{R}_k(s) := \frac{1}{T - t_0} \int_{t_0}^{T} \mathbb{E}_{(X_t, Z) \sim p_{k,t}(\cdot,\cdot)} \Big[\|s(X_t, Z, t) - s_k^*(X_t, Z, t)\|_2^2 \Big]\, dt. \tag{16}$$

According to Vincent (2011), $\mathcal{R}_k(s)$ can be expressed as:

$$\mathcal{R}_k(s) = \mathcal{L}_k(s) + C_{\mathrm{DSM},k}. \tag{17}$$

Therefore, the optimization objective in (13) can be equivalently expressed as $\mathbb{E}_{D_0,\ldots,D_K}\big[\mathcal{R}_0(\widehat{s}_w)\big]$. We next establish an upper bound of the target score matching risk under the following assumptions.

Assumption 1 (Transferability conditions)**.** *Let $W_N := \sum_{k=1}^K w_k n_k$, and define the weighted source marginal density of Z as $p_{w\setminus 0}(z) := \sum_{k=1}^K \frac{w_k n_k}{W_N} p_k(z)$. Assume that there exist constants $C_{\mathrm{dr}} \in (0, 1]$ and $C_{\mathrm{DR}} \geq 1$, such that*

$$C_{\mathrm{dr}} \cdot \max_{1 \leq k \leq K} p_k(z) \leq p_0(z) \leq C_{\mathrm{DR}} \cdot p_{w\setminus 0}(z), \qquad p_0\text{-almost everywhere}. \tag{18}$$

Let $\gamma > 0$ be a convergence rate. For $k = 1, \ldots, K$, assume that

$$\frac{1}{T - t_0} \int_{t_0}^{T} \mathbb{E}_{(X_t, Z) \sim p_{0,t}} \Big[\|s_0^*(X_t, Z, t) - s_k^*(X_t, Z, t)\|_2^2 \Big]\, dt = \mathcal{O}\left(n_0^{-2\gamma} \right). \tag{19}$$

Assumption 2 (Smoothness and tail conditions). *Let $\beta > 0$ be the smoothness parameter. For each $k = 0, \dots, K$, assume that there exists a function*

$$f_k \in \mathcal{H}^\beta \left(\mathbb{R}^{d_x} \times \mathbb{R}^{d_z}, B \right), \tag{20}$$

where the Hölder ball $\mathcal{H}^\beta(\cdot, B)$ is defined in (80), such that the conditional density $p_k(x|z)$ admits the representation

$$p_k(x|z) = \exp\left(-C_x \|x\|_2^2/2\right) f_k(x, z), \qquad (x, z) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_z}, \quad k = 0, \dots, K. \tag{21}$$

Here, $B > 0$ and $C_x > 0$ are constants independent of k. Moreover, there exist constants $0 < C_f < \infty$, independent of k, such that

$$\inf_{k=0,\dots,K} \inf_{(x,z) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_z}} f_k(x, z) \geq C_f. \tag{22}$$

In addition, we assume that each marginal density $p_k(z)$ has a sub-Gaussian tail, i.e., $p_k(z) \lesssim \exp\left(-C_z \|z\|_2^2/2\right)$. We also assume that there exist nonnegative functions q_k for $k = 1, \dots, K$, s.t.

$$q_k \in \mathcal{H}^\beta(\mathbb{R}^{d_z}, L_q) \quad \text{and} \quad p_k(z) = p_0(z) q_k(z) \quad \text{for Lebesgue-almost every } z \in \mathbb{R}^{d_z}, \tag{23}$$

where $L_q > 0$ is a constant. We also define $q_0(z) \equiv 1$. More discussion is in Appendix B.

Theorem 1 (Score matching risk on the target domain of DA-Diff). *Define the effective sample size*

$$N_{\text{eff}} := \frac{(n_0 + W_N)^2}{n_0 + \sum_{k=1}^K w_k^2 n_k}. \tag{24}$$

Suppose that $W_N > 0$, and define the normalized source weight vector $\overline{w} := \left(\frac{w_1 n_1}{W_N}, \dots, \frac{w_K n_K}{W_N}\right)^\top \in \mathbb{R}^K$. Further, define the source discrepancy matrix $G \in \mathbb{R}^{K \times K}$ with entries

$$G_{ij} := \frac{1}{T - t_0} \int_{t_0}^T \mathbb{E}_{(X_t, Z) \sim p_{0,t}} \left[\left\langle s_0^*(X_t, Z, t) - s_i^*(X_t, Z, t), s_0^*(X_t, Z, t) - s_j^*(X_t, Z, t) \right\rangle \right] dt. \tag{25}$$

Suppose that there is at least one $w_k > 0$ and the weights satisfy

$$\max\{1, w_1, \dots, w_K\} \lesssim \frac{n_0 + \sum_{k=1}^K w_k^2 n_k}{n_0 + W_N}. \tag{26}$$

Then, for any weight vector $w = (w_1, \dots, w_K)^\top$ satisfying $w_k \geq 0$ for all k, $W_N > 0$, and (26), the score-matching risk on the target domain can be bounded as

$$\mathbb{E}_{D_0, \dots, D_K} \left[\mathcal{R}_0(\widehat{s}_w) \right] \lesssim \widetilde{\mathcal{O}} \left(N_{\text{eff}}^{-\frac{2\beta}{d_x + d_z + 2\beta}} \right) \tag{27}$$

$$+ \frac{W_N^2}{(n_0 + W_N)^2} \left[\overline{w}^\top G \overline{w} + \operatorname{diag}(G)^\top \overline{w} \right], \tag{28}$$

where $\widetilde{\mathcal{O}}(\cdot)$ suppresses polylogarithmic factors in N_{eff}. The t_0, T and the parameter of $\widehat{s}_w$ is specified in Appendix K.

Remark 1. *The upper bound in Theorem 1 consists of two main components. The first term (27) is the estimation error arising from weighted ERM and decreases as the effective sample size N_{eff} increases. The second term (28) captures the transfer bias induced by discrepancies between the source and target domains.*

Remark 2. *The matrix G in (25) involves the unknown scores $s_0^*, \dots, s_K^*$ and therefore cannot be evaluated directly. For each source domain $k = 1, \dots, K$, we replace s_k^* with a pretrained score network $\widehat{s}_k$ obtained from the corresponding source dataset. For the target domain, we use the weighted ERM estimator $\widehat{s}_w$ as a surrogate for s_0^*, since $\widehat{s}_w$ is trained to achieve a small score matching risk on the target domain. Moreover, the expectation $\mathbb{E}_{(X_t, Z) \sim P_{0,t}}$ (25) can be written as $\mathbb{E}_{(X_0, Z) \sim P_0} \mathbb{E}_{X_t | X_0}$, where the distribution of $X_t | X_0$ is defined in Section 2.2. The outer expectation is approximated using an independent holdout sample from the target distribution. The resulting empirical estimator of G is given in (414), and its estimation consistency is established in Theorem 9.*

Building on Theorem 1, we further establish a convergence bound for the learned conditional distribution in TV distance. Specifically, let $\widehat{P}_w(\cdot|z)$ denote the conditional distribution induced by the reverse process (5) with $\widehat{s}_w$. The following theorem gives the expected TV distance between $\widehat{P}_w(\cdot|z)$ and the target conditional distribution $P_0(\cdot|z)$.

Theorem 2 (Target-domain TV error bound of DA-Diff)**.** *Suppose that the conditions of Theorem 1 hold. Then,*

$$\mathbb{E}_{D_0,\dots,D_K}\mathbb{E}_{Z_0}\left[d_{\mathrm{TV}}\left(P_0(\cdot|Z_0),\widehat{P}_w(\cdot|Z_0)\right)\right] = \widetilde{\mathcal{O}}\left(N_{\mathrm{eff}}^{-\frac{\beta}{d_x+d_z+2\beta}} + \frac{W_N}{n_0+W_N}\sqrt{\overline{w}^\top G\overline{w} + \mathrm{diag}(G)^\top \overline{w}}\right). \tag{29}$$

The two terms in (29) correspond to the estimation error and the transfer bias, respectively. To characterize a sufficient regime for consistency, suppose that the polylogarithmic factor hidden in $\widetilde{\mathcal{O}}(\cdot)$ is of order $(\log N_{\mathrm{eff}})^q$ for some fixed $q > 0$, and let $\Gamma := \frac{\beta}{d_x+d_z+2\beta}$. Under Assumption 1, $\sqrt{\overline{w}^\top G\overline{w} + \mathrm{diag}(G)^\top \overline{w}} = \mathcal{O}(n_0^{-\gamma})$. Hence, the TV error is controlled by $(\log N_{\mathrm{eff}})^q \cdot \left(N_{\mathrm{eff}}^{-\Gamma} + n_0^{-\gamma}\right)$. The estimation error term converges to zero whenever $N_{\mathrm{eff}} \to \infty$. For the transfer bias term, a sufficient condition is $(\log N_{\mathrm{eff}})^q = o(n_0^{\gamma})$. Therefore, a simple sufficient growth regime for the consistency of DA-Diff is

$$1 \ll N_{\mathrm{eff}} \ll \exp(n_0^c), \qquad 0 < c < \frac{\gamma}{q}. \tag{30}$$

This growth regime (30) is rather mild: it only requires $\gamma > 0$ in (19), meaning that the source-to-target score discrepancy may decay at an arbitrarily slow polynomial rate.

3.2 Domain-adapted conditional independence test (DA-CIT)

Theorem 3 (Type I error of DA-CIT)**.** *Suppose that the conditions of Theorem 1 hold. Let p_w denote the p-value returned by DA-CIT. Then, under the null hypothesis H_0 in (1), for any significance level $\alpha \in (0,1)$,*

$$\mathbb{P}\big(p_w \le \alpha\big) \le \alpha + n_0\,\widetilde{\mathcal{O}}\left(N_{\mathrm{eff}}^{-\frac{\beta}{d_x+d_z+2\beta}} + \frac{W_N}{n_0+W_N}\sqrt{\overline{w}^\top G\overline{w} + \mathrm{diag}(G)^\top \overline{w}}\right). \tag{31}$$

The conditions for asymptotic Type I error control are stronger than those required for the TV consistency of DA-Diff in (30) because the TV error is accumulated by n_0 times in (31). Using the same polylogarithmic factor $(\log N_{\mathrm{eff}})^q$ as in (30), the excess Type I error is controlled by $n_0 \cdot (\log N_{\mathrm{eff}})^q \cdot \left(N_{\mathrm{eff}}^{-\Gamma} + n_0^{-\gamma}\right)$. For the estimation term, since $(\log N_{\mathrm{eff}})^q = O(N_{\mathrm{eff}}^{\zeta})$ for any sufficiently small $\zeta > 0$, a sufficient condition is $n_0^{1/(\Gamma-\zeta)} \ll N_{\mathrm{eff}},\ 0 < \zeta < \Gamma$. For the transfer-bias term, the additional factor n_0 strengthens the upper-growth condition in (30) to $(\log N_{\mathrm{eff}})^q = o(n_0^{\gamma-1})$. Thus, when $\gamma > 1$, a simple sufficient growth regime for asymptotic Type I error control is

$$n_0^{1/(\Gamma-\zeta)} \ll N_{\mathrm{eff}} \ll \exp(n_0^c), \qquad 0 < \zeta < \Gamma, \quad 0 < c < \frac{\gamma-1}{q}. \tag{32}$$

Under (32), the excess Type I error vanishes, and hence $\mathbb{P}(p_w \le \alpha) \le \alpha + o(1)$. The lower bound on N_{eff} ensures that the estimation error decays sufficiently fast to offset the factor n_0, whereas the upper bound ensures that the polylogarithmic factor grows sufficiently slowly so that $n_0^{1-\gamma}(\log N_{\mathrm{eff}})^q \to 0$.

By the decomposition in (17), minimizing the objective in (13) with respect to the weights w is equivalent to minimizing $\mathbb{E}_{D_0,\dots,D_K}[\mathcal{R}_0(\widehat{s}_w)]$. In Appendix C, we reformulate this weight optimization problem into a majorization–minimization quadratic program (MM-QP) procedure and solve it accordingly.

4 Synthetic Experiments

Section 4.1 examines the conditional distribution estimation performance of DA-Diff under heterogeneous source domains, while Section 4.2 evaluates the Type I error and power of DA-CIT.

4.1 GAUSSIAN MIXTURE EXAMPLE

We evaluate DA-Diff on a controlled Gaussian mixture setting (Hagemann et al., 2022; Chemseddine et al., 2025; Yang & Fukumizu, 2026). The detailed data-generating mechanism is provided in Appendix D.1. We set $n_0 = n_1 = \cdots = n_K = 500$ and study the effect of varying the number of source domains K. We use the Wasserstein-2 (W_2) distance to evaluate the quality of the estimated conditional distribution. For each Z_0 in the test set, we draw 1,000 samples from the true conditional distribution $P_0(\cdot|Z_0)$ and 1,000 samples from $\widehat{P}_w(\cdot|Z_0)$ estimated by DA-Diff. We then compute the W_2 distance between these two empirical distributions. We repeat this procedure over 200 test values of Z_0 and report the average W_2 distance. Additional ablation studies are provided in Appendix E.

Our baselines include a diffusion model trained only on the target dataset (Target-only), a diffusion model trained on the concatenation of all target and source datasets (Cat-all), and a diffusion model first trained on all available target and source data and then fine-tuned on the target dataset (Finetune). We further compare with diffusion models for transfer learning, namely Domain Guidance (DoG) (Zhong et al., 2025) and the Transfer Guided Diffusion Process (TGDP) (Ouyang et al., 2024). Weighted ERM methods are also included, such as Unified Optimization of Weights and Quantities (UOWQ) (Zhang et al., 2026a) and Robust Learning from Untrusted Sources (Robust) (Konstantinov & Lampert, 2019). The results are reported in Table 1.

Table 1: Performance of proposed DA-Diff, Target-only, Cat-all, Finetune, DoG, TGDP, UOWQ, and Robust under the data-generating mechanism in Appendix D.1. Experiments are repeated over 30 seeds; we report the W_2 distance along with one standard deviation (std). The best, second-best, and third-best results are highlighted in bold, underlined, and †, respectively.

Method	W_2 ($K = 2$)	W_2 ($K = 5$)	W_2 ($K = 10$)	W_2 ($K = 20$)
DA-Diff (Ours)	**0.194**(0.030)	**0.190**(0.052)	<u>0.216</u>(0.060)	<u>0.196</u>(0.033)
Target-only	0.343(0.079)	0.343(0.079)	0.343(0.079)	0.343(0.079)
Cat-all	0.644(0.134)	0.495(0.114)	0.372(0.109)	0.330(0.079)
Finetune	0.384(0.063)	0.283(0.056)	**0.193**(0.043)	**0.168**(0.036)
DoG (2025)	0.311†(0.057)	<u>0.240</u>(0.068)	0.218†(0.062)	0.222†(0.069)
TGDP (2024)	0.663(0.151)	0.466(0.110)	0.372(0.089)	0.331(0.077)
UOWQ (2026a)	0.484(0.108)	0.571(0.284)	0.613(0.182)	0.742(0.237)
Robust (2019)	<u>0.230</u>(0.056)	0.282†(0.093)	0.260(0.078)	0.227(0.066)

Table 1 compares the methods under different K. DA-Diff achieves the lowest mean W_2 error for $K = 2, 5$, and the second-lowest error for $K = 10, 20$, where Finetune performs best. Across all four settings, DA-Diff substantially reduces the error relative to Target-only. In contrast, Cat-all performs worse than Target-only for $K = 2, 5, 10$ and yields only a modest improvement at $K = 20$, illustrating that simply pooling heterogeneous source data can lead to negative transfer. Compared with DoG and TGDP, DA-Diff attains lower W_2 error for every K. TGDP is also worse than Target-only for $K = 2, 5, 10$. UOWQ degrades as K increases, whereas Robust remains relatively stable but is consistently worse than DA-Diff. Additional experiments in Table 2 show a similar pattern at larger sample sizes: when the per-domain sample size is 1000 or 1500, DA-Diff outperforms Target-only, Cat-all, and Finetune for all considered values of K.

4.2 POST-NONLINEAR EXAMPLE

We evaluate the performance of DA-CIT on the post-nonlinear model (Yang et al., 2025a; Scetbon et al., 2022), a commonly used benchmark for nonlinear conditional independence testing. The detailed data-generating procedure is provided in Appendix D.2.

We fix the significance level at $\alpha = 0.05$ and repeat each experiment 100 times, reporting both Type I error and testing power. We consider one target domain and 10 source domains, with $n_0 = \cdots = n_{10} = 500$, and vary the conditioning dimension d_z from 5 to 100. To further assess calibration and its variability, Appendix E compares the empirical distribution of the p-values over the 100 repetitions with the $U(0, 1)$ distribution and reports the corresponding Kolmogorov–Smirnov (KS) statistic. Additional ablation studies are also provided in Appendix E.

We compare DA-CIT with several representative CI tests. KCIT (Zhang et al., 2011) performs CI testing in a reproducing kernel Hilbert space. CDCIT (Yang et al., 2025a) estimates the conditional distribution using half of the target data and applies the CRT to the remaining half. SGMCIT (Ren et al., 2025) estimates the conditional distribution via sliced score matching, while RBPT (Polo et al., 2023) is designed to improve robustness to regression-model misspecification. ECCIT (Pan et al., 2026) calibrates the p-values of the Generalized Covariance Measure (GCM; Shah & Peters, 2020) to mitigate testing miscalibration. CRT* (Zhang et al., 2026c) leverage heterogeneous external and unlabeled data to improve the CRT; however, their approach relies on a sparse linear model for the conditional distribution, whereas our diffusion-based framework accommodates substantially more general nonlinear and non-Gaussian conditional distributions.

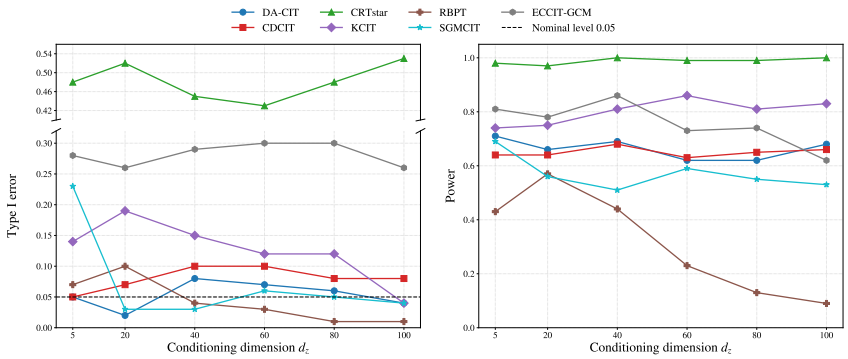


Figure 1: Type I error and power of each CI testing under different d_z.

Figure 1 reports the Type I error and power of different CI testing methods with varying d_z. Across all considered dimensions, DA-CIT consistently achieves a Type I error no larger than that of CDCIT and generally maintains it close to the nominal level, while maintaining comparable power. SGMCIT and RBPT also maintain reasonable Type I error control in higher dimensions, but with lower power than DA-CIT and CDCIT. KCIT maintains relatively high power, but its Type I error is clearly inflated in most settings. ECCIT-GCM and CRT* achieve relatively high power, but both exhibit substantial Type I error inflation across the considered dimensions. In particular, despite incorporating external data, CRT* achieves high power at the cost of severe Type I error inflation.

5 Real Data Analysis

We evaluate 50 CI and 40 non-CI (X_0, Y_0, Z_0) triples in the multi-source Sachs dataset (Sachs et al., 2005), and report the precision, recall, and F1 of each CI testing. Among them, DA-CIT achieves the highest F1. Detailed results are provided in Appendix F.

6 Conclusion

In this work, we address the fundamental difficulty of conditional distribution estimation in the CRT when target-domain data are limited and heterogeneous source-domain data are available. To address this problem, we propose DA-Diff, to the best of our knowledge, the first transfer learning method for diffusion models built on weighted ERM. We establish a target-domain TV error bound for DA-Diff, which explicitly characterizes the trade-off between estimation error and transfer bias. Building on DA-Diff, we further propose DA-CIT for CI testing. Under suitable assumptions, we show that the TV error of DA-Diff converges to zero and that the excess Type I error of DA-CIT also vanishes asymptotically. We evaluate DA-Diff through simulation studies on conditional distribution estimation, and assess DA-CIT on both simulated and real-world data. The empirical results show that DA-Diff can effectively leverage heterogeneous source domains to improve conditional distribution estimation, while DA-CIT provides reliable Type I error control and maintains competitive power. A current limitation is the relatively high computational cost, mainly due to source-model pretraining and repeated weight optimization during weighted ERM training. Improving the computational efficiency of DA-Diff and DA-CIT is therefore an important direction for future work.

Reproducibility Statement

We introduce two new methods, DA-Diff and DA-CIT. The diffusion-model component of DA-Diff is described in Section 2.2, while the domain-adaptation component is presented in Section 2.3 and Appendix C. The complete procedures for training DA-Diff and conducting DA-CIT are summarized in Algorithms 1 and 2, respectively.

The theoretical guarantees are presented in Section 3, including the target-domain total variation error bound for DA-Diff and the asymptotic vanishing of the excess Type I error for DA-CIT. Complete proofs of the theoretical results are provided in Appendix K.

To facilitate reproducibility, we also provide anonymized code through OpenReview, including implementations of DA-Diff, DA-CIT, and the experimental procedures used in this paper.

AI Use Statement

In this work, we used generative AI tools to assist with and polish the writing of the manuscript, identify potentially relevant related work, and assist with software code development. All AI-assisted outputs were independently reviewed and verified by the authors. In particular, potentially relevant literature suggested by generative AI was verified through manual literature searches.

References

Mona Azadkia and Sourav Chatterjee. A simple measure of conditional dependence. *The Annals of Statistics*, 49(6):3070 – 3102, 2021.

Bilol Banerjee, Bhaswar B Bhattacharya, and Anil K Ghosh. A ball divergence-based measure for conditional independence testing with a local wild bootstrap. *Biometrika*, 113(3):asag027, 2026.

Alexis Bellot and Mihaela van der Schaar. Conditional independence testing using generative adversarial networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems*, volume 19, 2006.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1):151–175, 2010.

Thomas B Berrett, Yi Wang, Rina Foygel Barber, and Richard J Samworth. The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):175–197, 2020.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 02 2013.

Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.

Jannis Chemseddine, Paul Hagemann, Gabriele Steidl, and Christian Wald. Conditional wasserstein distances with applications in bayesian ot flow matching. *Journal of Machine Learning Research*, 26(141):1–47, 2025.

Ziheng Cheng, Tianyu Xie, Shiyue Zhang, and Cheng Zhang. Provable sample-efficient transfer learning conditional diffusion models via representation learning. In *Advances in Neural Information Processing Systems*, 2025.

R. Dennis Cook. On the foundational arguments of sufficient dimension reduction. *WIREs Computational Statistics*, 18(2):e70064, 2026. doi: https://doi.org/10.1002/wics.70064. URL `https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wics.70064`. e70064 4107620.

Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R. Arabnia. A brief review of domain adaptation. In *Advances in Data Science and Information Engineering*, pp. 877–894, 2021.

Hengyu Fu, Zhuoran Yang, Mengdi Wang, and Minshuo Chen. Unveil conditional diffusion models with classifier-free guidance: a sharp statistical theory, 2024.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

Paul Hagemann, Johannes Hertrich, and Gabriele Steidl. Stochastic normalizing flows for inverse problems: A markov chains viewpoint. *SIAM/ASA Journal on Uncertainty Quantification*, 10(3): 1162–1190, 2022.

Zheng He, Roman Pogodin, Yazhe Li, Namrata Deka, Arthur Gretton, and Danica J. Sutherland. On the hardness of conditional independence testing in practice. In *Advances in Neural Information Processing Systems*, 2025.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Emiel Hoogeboom, Víctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3D. In *International Conference on Machine Learning*, volume 162, pp. 8867–8887, 2022.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, 2022.

Nikola Konstantinov and Christoph Lampert. Robust learning from untrusted sources. In *International Conference on Machine Learning*, volume 97, pp. 3488–3498, 2019.

Shuai Li, Yingjie Zhang, Hongtu Zhu, Christina Wang, Hai Shu, Ziqi Chen, et al. K-nearest-neighbor local sampling based conditional independence testing. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

Shuangning Li and Molei Liu. Maxway crt: improving the robustness of the model-x inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(5):1441–1470, 11 2023.

Molei Liu, Eugene Katsevich, Lucas Janson, and Aaditya Ramdas. Fast and powerful conditional randomization testing via distillation. *Biometrika*, 109(2):277–293, 2022.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. In *Advances in Neural Information Processing Systems*, volume 21, 2008.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Multiple source adaptation and the rényi divergence. In *Uncertainty in Artificial Intelligence*, pp. 367–374, Arlington, Virginia, USA, 2009. AUAI Press.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan LI. Large language diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 50608–50646, 2025.

Yidong Ouyang, Liyan Xie, Hongyuan Zha, and Guang Cheng. Transfer learning for diffusion models. In *Advances in Neural Information Processing Systems*, 2024.

Milleno Pan, Antoine de Mathelin, and Wesley Tansey. Empirically calibrated conditional independence tests. In *International Conference on Artificial Intelligence and Statistics*, 2026.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Felipe Maia Polo, Yuekai Sun, and Moulinath Banerjee. Conditional independence testing under misspecified inductive biases. In *Advances in Neural Information Processing Systems*, 2023.

Yixin Ren, Chenghou Jin, Yewei Xia, Li Ke, Longtao Huang, Hui Xue, Hao Zhang, Jihong Guan, and Shuigeng Zhou. Score-based generative modeling for conditional independence testing. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2410–2419, 2025.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, 2005.

Subham Sekhar Sahoo, Marianne Arriola, Aaron Gokaslan, Edgar Mariano Marroquin, Alexander M Rush, Yair Schiff, Justin T Chiu, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems*, 2024.

Meyer Scetbon, Laurent Meunier, and Yaniv Romano. An asymptotic test for conditional independence using analytic kernel embeddings. In *International Conference on Machine Learning*, pp. 19328–19346, 2022.

Rajen D Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 2020.

Chengchun Shi, Tianlin Xu, Wicher Bergsma, and Lexin Li. Double generative adversarial networks for conditional independence testing. *Journal of Machine Learning Research*, 22(285):1–32, 2021.

Aliaksandra Shysheya, Cristiana Diaconu, Federico Bergamin, Paris Perdikaris, José Miguel Hernández-Lobato, Richard E. Turner, and Emile Mathieu. On conditional diffusion models for PDE simulations. In *Advances in Neural Information Processing Systems*, 2024.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT Press, 2000.

Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769 – 2794, 2007.

Rosanna Turrisi, Rémi Flamary, Alain Rakotomamonjy, and Massimiliano Pontil. Multi-source domain adaptation via weighted joint distributions optimal transport. In *Uncertainty in Artificial Intelligence*, volume 180, pp. 1970–1980, 2022.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

Robin Vogel, Mastane Achab, Stéphan Clémençon, and Charles Tillier. Weighted empirical risk minimization: Sample selection bias correction based on importance sampling, 2020.

Wenjie Wang, Mingming Gong, Biwei Huang, James Bailey, Bo Han, Kun Zhang, and Feng Liu. Practical kernel selection for kernel-based conditional independence test. In *Advances in Neural Information Processing Systems*, volume 38, pp. 49333–49370, 2025.

Yakun Wang, Leyang Wang, Song Liu, and Taiji Suzuki. Zero-flow encoders. In *International Conference on Machine Learning*, 2026.

Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. In *International Conference on Learning Representations*, 2022.

Yanfeng Yang and Kenji Fukumizu. Diffusion models with double guidance: Generate with aggregated datasets. *arXiv:2505.13213*, 2026.

Yanfeng Yang, Shuai Li, Yingjie Zhang, Zhuoran Sun, Hai Shu, Ziqi Chen, and Renming Zhang. Conditional diffusion models based conditional independence testing. In *AAAI Conference on Artificial Intelligence*, 2025a.

Yanfeng Yang, Lihong Zhang, Ziqi Chen, Miaomiao Yu, and Lei Chen. Confidence interval estimation of mixed oil length with conditional diffusion models. *Journal of Pipeline Science and Engineering*, pp. 100380, 2025b.

Yanfeng Yang, Siwei Chen, Pingping Hu, Zhaotong Shen, Yingjie Zhang, Zhuoran Sun, Shuai Li, Ziqi Chen, and Kenji Fukumizu. Conditionally whitened generative models for probabilistic time series forecasting. In *International Conference on Learning Representations*, 2026.

Weiwei Ye, Zhuopeng Xu, and Ning Gui. Non-stationary diffusion for probabilistic time series forecasting. In *International Conference on Machine Learning*, 2025.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Conference on Uncertainty in Artificial Intelligence*, pp. 804–813, 2011.

Qingyue Zhang, Chang Chu, Haohao Fu, Tianren Peng, Yanru Wu, Guanbo Huang, Yang Li, and Shao-Lun Huang. Unified optimization of source weights and transfer quantities in multi-source transfer learning: An asymptotic framework, 2026a.

Yi Zhang, Linjun Huang, Yun Yang, and Xiaofeng Shao. Doubly robust conditional independence testing with generative neural networks. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 88(1):195–220, 02 2026b.

Yingjie Zhang, Ziqi Chen, and Chenlei Leng. CRT*: Conditional randomization testing with heterogeneous external and unlabeled data, 2026c.

Shunyu Zhao, Yanfeng Yang, Shuai Li, and Kenji Fukumizu. Flow matching based conditional independence tests and causal structure learning. In *ICLR 2026 2nd Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy*, 2026.

Jincheng Zhong, XiangCheng Zhang, Jianmin Wang, and Mingsheng Long. Domain guidance: A simple transfer approach for a pre-trained diffusion model. In *International Conference on Learning Representations*, 2025.

A Related work

We have provided a detailed review of the literature on CI tests in Section 1. In this section, we focus on related work on diffusion models for transfer learning, domain adaptation, and weighted ERM.

A.1 Diffusion models for transfer learning

Diffusion models are a class of generative models that have achieved remarkable success in various fields, including image generation (Ho et al., 2020; Song et al., 2021), scientific discovery (Hoogeboom et al., 2022; Xu et al., 2022; Yang & Fukumizu, 2026), engineering problem (Shysheya et al., 2024; Yang et al., 2025b), time-series forecasting (Ye et al., 2025; Yang et al., 2026), and language modeling (Sahoo et al., 2024; Nie et al., 2025).

These studies demonstrate the broad applicability of diffusion models across a wide range of fields. However, most existing diffusion models are trained on data from a single domain or from pooled datasets, and relatively limited attention has been paid to how heterogeneous source domains can be selectively leveraged to improve generation in a target domain. This motivates our study of multi-source domain adaptation for diffusion models.

Several works have studied transfer learning specifically for diffusion models. Zhong et al. (2025) propose Domain Guidance (DoG), which combines a pre-trained source model and a target-domain fine-tuned model by classifier-free guidance (Ho & Salimans, 2021). Although effective, DoG requires both models to be evaluated at each sampling step, increasing the inference cost, and is primarily designed for transferring a single pre-trained source model rather than selectively exploiting multiple heterogeneous source domains.

Ouyang et al. (2024) propose the Transfer Guided Diffusion Process (TGDP), which expresses the target score as the pre-trained source score augmented by a guidance term estimated from the source–target density ratio. In practice, this density ratio is estimated through a domain classifier. However, reliable density-ratio estimation can be challenging when target data are limited or when the source and target distributions differ substantially, and estimation errors can directly propagate to the guidance term. TGDP therefore introduces additional regularization to improve the training and calibration of the guidance network. Moreover, similar to DoG, TGDP mainly considers transfer from a single source model and does not address source selection and weighting in the multi-source setting.

In contrast, DA-Diff directly incorporates multiple source datasets into diffusion-model training through weighted ERM. It does not require explicit density-ratio estimation and automatically balances the statistical benefit of additional source samples against source–target distribution shifts.

Unlike DoG and TGDP, which mainly focus on transferring knowledge from a single pre-trained source model, Cheng et al. (2025) explicitly consider transfer learning from multiple source domains and provide theoretical guarantees for diffusion-model finetuning. Their approach, however, relies on a different transfer mechanism based on learning a shared representation across domains. Their analysis decomposes the target error into the error of target-domain fine-tuning, the error of learning the shared representation from source domains, and a task-mismatch term. In particular, source data mainly improve the estimation of the shared representation, while the target score model is still fine-tuned using target-domain samples.

In comparison, DA-Diff directly incorporates source samples into the training of the target model through weighted ERM. Consequently, the estimation term in our bound depends on the effective sample size N_{eff}, so that transferable source samples can directly improve the convergence rate of the target conditional distribution estimator. At the same time, our bound in Theorem 2 explicitly characterizes the transfer bias through the source weights, the total source contribution $W_N/(n_0 + W_N)$, and the source discrepancy matrix G, providing a quantitative trade-off between exploiting additional source samples and avoiding negative transfer. Moreover, our analysis does not require the existence of a shared dimension-reducing representation across domains.

A.2 Domain adaptation and Weighted ERM

Weighted ERM is a classical approach to domain adaptation (DA) that learns a model by minimizing a weighted combination of empirical losses across domains (Vogel et al., 2020). Rather than treating all datasets equally, it adjusts their contributions to exploit useful source information while mitigating negative transfer caused by distribution shift.

Two settings closely related to our approach are target-as-mixture adaptation and learning from potentially untrusted sources (Mansour et al., 2008; Konstantinov & Lampert, 2019). In the former, the target distribution is represented or approximated by a convex combination of source distributions. In the latter, a trusted target reference dataset is used to assess source relevance and downweight irrelevant or corrupted sources. Our method is related to both perspectives: it constructs a weighted training mixture and adaptively controls source contributions, without requiring the target distribution to be exactly a mixture of the source distributions.

Foundational theoretical contributions to DA include Ben-David et al. (2006), who establish target-risk bounds highlighting the trade-off between source prediction error and distributional discrepancy in the learned representation. For the multi-source setting, Mansour et al. (2008) provide guarantees for distribution-weighted combinations of source predictors when the target input distribution is a mixture of source distributions. Building on this framework, Mansour et al. (2009) introduce Rényi divergence into the analysis and extend the guarantees to arbitrary target distributions by quantifying their discrepancy from suitable mixtures of the source distributions.

A particularly relevant approach is Robust Learning from Untrusted Sources (Konstantinov & Lampert, 2019). The method uses a small trusted target dataset to evaluate the reliability of multiple source domains, and then assigns larger weights to sources that are more similar to the target domain. The resulting weights are used in a weighted ERM objective to reduce the effect of unreliable or negatively transferable sources. However, its theoretical analysis assumes a uniformly bounded loss, which is not directly satisfied by the denoising score-matching loss in (6). Therefore, extending this framework to diffusion models requires additional treatment of the unbounded loss and a separate analysis of the generated distribution.

Turrisi et al. (2022) propose multi-source domain adaptation via weighted joint distributions optimal transport (MSDA-WJDOT). Their method jointly learns source weights and a target predictor by minimizing an optimal-transport discrepancy between a weighted mixture of source joint distributions and a target joint distribution constructed using predicted labels, thereby accounting for discrepancies in both feature and label spaces.

More recently, Zhang et al. (2026a) propose Unified Optimization of Weights and Quantities (UOWQ), which jointly optimizes source weights and transfer quantities under weighted maximum likelihood estimation. UOWQ is developed in a parametric setting, where source–target discrepancies are characterized through differences between the parameters of the corresponding distributions, rather than directly comparing general nonparametric distributions.

Our work is inspired by the risk-driven weighting strategy of UOWQ (Zhang et al., 2026a), but differs substantially in both its formulation and applicability. UOWQ is developed under a parametric model and weighted maximum likelihood estimation. It represents each domain by a parameter vector and assumes that source and target domains are close in parameter space. While the basic transferability assumption is expressed through the ℓ_2 norm between source and target parameters, the theoretical risk bound and source-weight optimization further use a Fisher-information-weighted version of this ℓ_2 norm. Therefore, UOWQ fundamentally relies on a common parametric representation across domains, rather than directly comparing the underlying nonparametric distributions or learned functions. This parameter-space formulation requires the domain-specific models to share a common and identifiable parameterization, which can be restrictive for modern neural networks. In particular, neural-network parameters are generally non-identifiable: for example, the two ReLU networks $a\,\mathrm{ReLU}(bx + c) + d$ and $\mathrm{ReLU}(abx + ac) + d$ represent exactly the same function for $a > 0$, but have different parameter vectors (a, b, c, d) and $(1, ab, ac, d)$. Consequently, parameter discrepancies may reflect differences in parameterization rather than genuine differences between the learned distributions, and such symmetries may also lead to singular Fisher information. Moreover, the parametric formulation of UOWQ typically requires the models trained on different domains to have the same parameter structures, making it difficult to directly compare models with different neural-network architectures. These assumptions are particularly restrictive for diffusion models, which are commonly trained by denoising score matching rather than regular maximum likelihood estimation. In contrast, DA-Diff defines source discrepancies directly in the function space through differences between the output of the score functions. Therefore, its discrepancy measure is invariant to parameter reparameterization and does not require aligned parameter vectors or identical network architectures across domains; it only requires the models to produce score outputs of the same dimension. This enables DA-Diff to accommodate substantially more flexible and potentially nonparametric generative models while directly measuring the quantities relevant to diffusion-model training.

B Interpretation of the Assumption 1 and 2

Transferability assumptions Assumption 1 characterizes the transferability of the source domains from two complementary perspectives: the **overlap** of the marginal densities of the Z and the **similarity** of the conditional scores of X given Z.

First, (18) requires sufficient overlap between the target and source distributions of Z. The lower bound

$$C_{\mathrm{dr}} \max_{1 \le k \le K} p_k(z) \le p_0(z)$$

prevents any individual source domain from assigning arbitrarily large probability mass to regions that are rarely represented in the target domain. Equivalently, it uniformly controls the source-to-

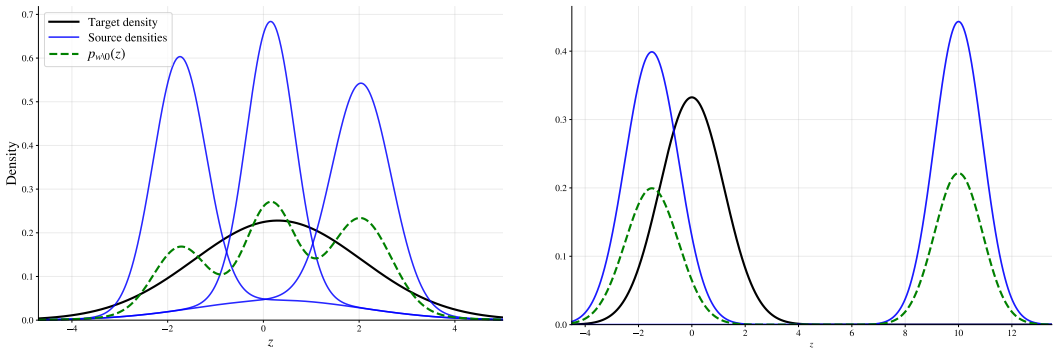


Figure 2: **Left panel:** The source densities are sufficiently covered by the target density ($C_{\mathrm{dr}} \max_{1\le k\le K} p_k(z) \le p_0(z)$), while their weighted mixture $p_{w\backslash 0}$ also covers the target density $\big(p_0(z) \le C_{\mathrm{DR}} p_{w\backslash 0}(z)\big)$; hence both density-ratio conditions can hold uniformly. **Right panel:** One source is located far from the target, while the remaining source does not sufficiently cover the target distribution. Consequently, the weighted source density fails to uniformly dominate p_0, violating the transferability condition.

target density ratios,

$$\frac{p_k(z)}{p_0(z)} \le C_{\mathrm{dr}}^{-1}, \qquad k = 1, \dots, K.$$

The upper bound

$$p_0(z) \le C_{\mathrm{DR}} p_{w\backslash 0}(z)$$

requires the weighted collection of source domains to sufficiently cover the target distribution. In particular, it rules out target regions with substantial probability mass but essentially no source observations. Since $p_{w\backslash 0}(z) \le \max_k p_k(z)$, these two inequalities together also imply

$$C_{\mathrm{dr}} p_{w\backslash 0}(z) \le p_0(z) \le C_{\mathrm{DR}} p_{w\backslash 0}(z),$$

so the target and weighted source marginals of Z are comparable up to constants. This overlap condition allows information learned from the source domains to be transferred to the target domain without introducing instability caused by extreme covariate shift. An illustration showing cases that satisfy or violate the conditions can be found in Figure 2.

Second, (19) controls the discrepancy between the target and source conditional distributions through their diffusion scores. Specifically, (19) requires each transferable source domain to have a conditional score that is sufficiently close to the target score under the target distribution. The parameter γ quantifies the degree of transferability: a larger γ corresponds to a faster decay of the source–target discrepancy and hence a smaller transfer bias.

Smoothness and tail assumptions Assumption 2 imposes standard smoothness and tail regularity conditions on the target and source distributions. It does not assume a specific parametric form for $p_k(X|Z)$ and allows flexible nonlinear and non-Gaussian relationships between X and Z.

Assumption (21) requires only Gaussian-type tail decay in x. The uniform bounds $C_f \le f_k(x, z) \le B$ prevent the density and its logarithmic derivatives from becoming degenerate. Similarly, the sub-Gaussian tail condition on $p_k(z)$ controls the probability of extreme values of Z.

Density-ratio regularity. Condition (23) complements the overlap condition (18). While (18) controls the magnitude of the source-to-target density ratios, it does not control their regularity as functions of z. The Hölder regularity of q_k is used to establish the Hölder regularity of the mixture conditional-density factor in (x, z) in Lemma 1, which is needed for the subsequent score-approximation analysis.

The factorization $p_k(z) = p_0(z) q_k(z)$ avoids explicitly dividing by $p_0(z)$ and therefore allows $p_0(z)$ to vanish. It also requires the source marginal to be absolutely continuous with respect to the

target marginal. Together, (18) and (23) control, respectively, the overlap and the smoothness of the marginal density ratios.

Overall, these smoothness and density-ratio regularity conditions are mainly technical assumptions for approximation and truncation arguments. They exclude highly irregular or extremely heavy-tailed distributions, while still covering a broad class of smooth nonparametric conditional distributions.

C Optimization of the Weighted-ERM Weights

The optimal weighted-ERM weights defined in (13) do not admit a closed-form solution. We therefore use the target score-matching risk bound in Theorem 1 to construct a surrogate weight selection criterion. Define the normalized domain weights

$$\pi_0 := \frac{n_0}{n_0 + W_N}, \qquad \pi_k := \frac{w_k n_k}{n_0 + W_N}, \quad k = 1, \dots, K. \tag{33}$$

Let $\boldsymbol{\pi} := (\pi_0, \pi_1, \dots, \pi_K)^\top \in \Delta^K$, where $\Delta^K := \left\{\boldsymbol{\pi} \in \mathbb{R}^{K+1} : \mathbf{1}^\top \boldsymbol{\pi} = 1,\ \boldsymbol{\pi} \geq 0\right\}$. We further denote $\boldsymbol{\pi}_S := (\pi_1, \dots, \pi_K)^\top$. The original weighted-ERM weights and W_N can be recovered from $\boldsymbol{\pi}$ as

$$w_k = \frac{n_0 \pi_k}{n_k \pi_0}, \quad \overline{w}_k = \frac{\pi_k}{1 - \pi_0}, \quad k = 1, \dots, K, \quad \text{and} \quad W_N = \frac{n_0(1 - \pi_0)}{\pi_0}. \tag{34}$$

C.1 Surrogate objective of (27) and the weight regularity constraint (26)

Reparameterization from w to $\boldsymbol{\pi}$ Define $D_N := \operatorname{diag}\left(n_0^{-1}, n_1^{-1}, \dots, n_K^{-1}\right)$. Then the inverse effective sample size can be written as

$$N_{\text{eff}}^{-1} = \frac{n_0 + \sum_{k=1}^K w_k^2 n_k}{(n_0 + W_N)^2} = \boldsymbol{\pi}^\top D_N \boldsymbol{\pi}. \tag{35}$$

For notational convenience, define $h(\boldsymbol{\pi}) := \boldsymbol{\pi}^\top D_N \boldsymbol{\pi}$. The theoretical estimation term in Theorem 1 involves $N_{\text{eff}}^{-2\Gamma}$, where $\Gamma = \frac{\beta}{d_x + d_z + 2\beta}$. Since the Hölder smoothness parameter β is unknown in practice, we replace 2Γ by a tuning exponent $\rho \in (0, 1)$. The selection of ρ is discussed in Section C.3.

Next, let $g := \operatorname{diag}(G) = (G_{11}, \dots, G_{KK})^\top$. The two transfer-bias terms in (28) can be rewritten as

$$\frac{W_N^2}{(n_0 + W_N)^2} \overline{w}^\top G \overline{w} = \boldsymbol{\pi}_S^\top G \boldsymbol{\pi}_S \quad \text{and} \quad \frac{W_N^2}{(n_0 + W_N)^2} g^\top \overline{w} = \frac{1}{2} \boldsymbol{\pi}_S^\top \left(\mathbf{1} g^\top + g \mathbf{1}^\top\right) \boldsymbol{\pi}_S. \tag{36}$$

Calibration of the Estimation–Bias trade off The bound in (27)–(28) is specified only up to multiplicative constants and polylogarithmic factors. Hence, the relative scales of the estimation term and the two transfer-bias terms are not determined by the theoretical bound. We therefore introduce three positive calibration parameters $\boldsymbol{\lambda} = (\lambda_{\text{est}}, \lambda_G, \lambda_{\text{diag}})^\top$. For a fixed $\rho \in (0, 1)$, consider the surrogate objective

$$F_{\rho, \boldsymbol{\lambda}}(\boldsymbol{\pi}) := \lambda_{\text{est}} h(\boldsymbol{\pi})^\rho + \lambda_G \boldsymbol{\pi}_S^\top G \boldsymbol{\pi}_S + \frac{\lambda_{\text{diag}}}{2} \boldsymbol{\pi}_S^\top \left(\mathbf{1} g^\top + g \mathbf{1}^\top\right) \boldsymbol{\pi}_S. \tag{37}$$

Since multiplying all three calibration parameters by the same positive constant leaves the minimizer of $F_{\rho, \boldsymbol{\lambda}}$ unchanged, only their relative magnitudes matter. Without loss of generality, we normalize them as

$$\lambda_{\text{est}} + \lambda_G + \lambda_{\text{diag}} = 1, \qquad \lambda_{\text{est}}, \lambda_G, \lambda_{\text{diag}} > 0. \tag{38}$$

Thus, although three calibration parameters are introduced, the parameterization has only two degrees of freedom. For notational convenience, define

$$A_G(\boldsymbol{\lambda}) := \lambda_G G + \frac{\lambda_{\text{diag}}}{2} \left(\mathbf{1} g^\top + g \mathbf{1}^\top\right), \tag{39}$$

so that

$$F_{\rho, \boldsymbol{\lambda}}(\boldsymbol{\pi}) = \lambda_{\text{est}} h(\boldsymbol{\pi})^\rho + \boldsymbol{\pi}_S^\top A_G(\boldsymbol{\lambda}) \boldsymbol{\pi}_S. \tag{40}$$

In practice, we select the $\boldsymbol{\lambda}$ over a finite simplex grid. Specifically, for an integer $M_\lambda \geq 3$, define

$$\mathcal{G}_\lambda := \left\{ \left(\frac{j_1}{M_\lambda}, \frac{j_2}{M_\lambda}, \frac{j_3}{M_\lambda} \right) : j_1, j_2, j_3 \in \mathbb{N}_+, \quad j_1 + j_2 + j_3 = M_\lambda \right\}. \tag{41}$$

We then select $(\lambda_{\text{est}}, \lambda_G, \lambda_{\text{diag}}) \in \mathcal{G}_\lambda$ according to the gradient alignment criterion introduced in Section C.3.

Numerical Enforcement of the Regularity Condition The regularity constraint (26) requires $\max\{1, w_1, \ldots, w_K\} \lesssim \frac{n_0 + \sum_{k=1}^K w_k^2 n_k}{n_0 + W_N}$. To enforce this condition numerically, we introduce a fixed constant $C_{\text{reg}} \geq 1$ and impose the constraint

$$\max\{1, w_1, \ldots, w_K\} \leq C_{\text{reg}} \frac{n_0 + \sum_{k=1}^K w_k^2 n_k}{n_0 + W_N}, \tag{42}$$

which is equivalent to

$$\frac{\pi_k}{n_k} \leq C_{\text{reg}} h(\boldsymbol{\pi}), \qquad k = 0, \ldots, K. \tag{43}$$

The value of C_{reg} is selected from a finite candidate grid, with the detailed selection procedure described in Section C.3. Hence, for each fixed ρ, we consider

$$\begin{aligned} \boldsymbol{\pi}^* \in \arg\min_{\boldsymbol{\pi}} \quad & F_{\rho,\boldsymbol{\lambda}}(\boldsymbol{\pi}) \\ \text{s.t.} \quad & \boldsymbol{\pi} \in \Delta^K, \\ & \frac{\pi_k}{n_k} \leq C_{\text{reg}} h(\boldsymbol{\pi}), \quad k = 0, \ldots, K. \end{aligned} \tag{44}$$

In implementation, we additionally impose $\pi_0 \geq \epsilon_\pi$ for a small $\epsilon_\pi > 0$ to ensure that the recovered weights in (34) remain finite.

C.2 Majorization–Minimization quadratic program (MM-QP) update for fixed hyperparameters

The optimization problem (44) is generally nonconvex for three reasons: the fractional-power term h^ρ, the possibly indefinite matrix $A_G(\boldsymbol{\lambda})$, and the quadratic constraints in (43). We handle these terms using a Majorization–Minimization quadratic program (MM-QP) procedure.

Majorization of $h(\boldsymbol{\pi})^\rho$ Suppose that a feasible iterate $\boldsymbol{\pi}^{(\kappa)}$ is available and define

$$h^{(\kappa)} := h(\boldsymbol{\pi}^{(\kappa)}) = (\boldsymbol{\pi}^{(\kappa)})^\top D_N \boldsymbol{\pi}^{(\kappa)}. \tag{45}$$

Since $u \mapsto u^\rho$ is concave for $0 < \rho < 1$,

$$h^\rho \leq (h^{(\kappa)})^\rho + a_\rho^{(\kappa)} (h - h^{(\kappa)}), \tag{46}$$

where

$$a_\rho^{(\kappa)} := \rho (h^{(\kappa)})^{\rho - 1}. \tag{47}$$

Therefore, up to an additive constant independent of $\boldsymbol{\pi}$, the nonconstant part of the majorizer of $h(\boldsymbol{\pi})^\rho$ is

$$a_\rho^{(\kappa)} \boldsymbol{\pi}^\top D_N \boldsymbol{\pi}. \tag{48}$$

Importantly, this construction only uses the concavity of the scalar map $u \mapsto u^\rho$ and does not require the composition $\boldsymbol{\pi} \mapsto h(\boldsymbol{\pi})^\rho$ to be concave.

Majorization of $\boldsymbol{\pi}_S^\top A_G(\boldsymbol{\lambda}) \boldsymbol{\pi}_S$ Consider the spectral decomposition

$$A_G(\boldsymbol{\lambda}) = V \Xi V^\top, \qquad \Xi = \operatorname{diag}(\xi_1, \ldots, \xi_K),$$

where $\xi_1, \ldots, \xi_K$ are the eigenvalues of $A_G(\boldsymbol{\lambda})$. Define

$$\Xi_+ := \operatorname{diag}\big(\max\{\xi_j, 0\}\big), \qquad \Xi_- := \operatorname{diag}\big(\max\{-\xi_j, 0\}\big),$$

and

$$A_G^+(\boldsymbol{\lambda}) := V\Xi_+V^\top \succeq 0, \qquad A_G^-(\boldsymbol{\lambda}) := V\Xi_-V^\top \succeq 0. \tag{49}$$

Then

$$A_G(\boldsymbol{\lambda}) = A_G^+(\boldsymbol{\lambda}) - A_G^-(\boldsymbol{\lambda}), \tag{50}$$

and hence

$$\boldsymbol{\pi}_S^\top A_G(\boldsymbol{\lambda})\boldsymbol{\pi}_S = \boldsymbol{\pi}_S^\top A_G^+(\boldsymbol{\lambda})\boldsymbol{\pi}_S - \boldsymbol{\pi}_S^\top A_G^-(\boldsymbol{\lambda})\boldsymbol{\pi}_S. \tag{51}$$

Since $-\boldsymbol{\pi}_S^\top A_G^-(\boldsymbol{\lambda})\boldsymbol{\pi}_S$ is concave, its first-order Taylor expansion at $\boldsymbol{\pi}_S^{(\kappa)}$ provides the upper bound

$$-\boldsymbol{\pi}_S^\top A_G^-(\boldsymbol{\lambda})\boldsymbol{\pi}_S \le -2(\boldsymbol{\pi}_S^{(\kappa)})^\top A_G^-(\boldsymbol{\lambda})\boldsymbol{\pi}_S + (\boldsymbol{\pi}_S^{(\kappa)})^\top A_G^-(\boldsymbol{\lambda})\boldsymbol{\pi}_S^{(\kappa)}. \tag{52}$$

Therefore, up to an additive constant independent of $\boldsymbol{\pi}_S$, the transfer-bias term $\boldsymbol{\pi}_S^\top A_G(\boldsymbol{\lambda})\boldsymbol{\pi}_S$ is majorized by

$$\boldsymbol{\pi}_S^\top A_G^+(\boldsymbol{\lambda})\boldsymbol{\pi}_S - 2(\boldsymbol{\pi}_S^{(\kappa)})^\top A_G^-(\boldsymbol{\lambda})\boldsymbol{\pi}_S. \tag{53}$$

Inner approximation of the regularity constraint (26). Because $h(\boldsymbol{\pi}) = \boldsymbol{\pi}^\top D_N \boldsymbol{\pi}$ is convex, its first-order Taylor expansion at $\boldsymbol{\pi}^{(\kappa)}$ provides the global lower bound

$$h(\boldsymbol{\pi}) \ge 2(\boldsymbol{\pi}^{(\kappa)})^\top D_N \boldsymbol{\pi} - h^{(\kappa)}. \tag{54}$$

We therefore replace (43) by the stronger inner constraints

$$\frac{\pi_k}{n_k} \le C_{\mathrm{reg}} \left[2(\boldsymbol{\pi}^{(\kappa)})^\top D_N \boldsymbol{\pi} - h^{(\kappa)}\right], \qquad k = 0, \dots, K. \tag{55}$$

Since $2(\boldsymbol{\pi}^{(\kappa)})^\top D_N \boldsymbol{\pi} - h^{(\kappa)} \le h(\boldsymbol{\pi})$, any point satisfying (55) also satisfies the original regularity constraint (43). Moreover, the lower bound is tight at the current iterate, since $2(\boldsymbol{\pi}^{(\kappa)})^\top D_N \boldsymbol{\pi}^{(\kappa)} - h^{(\kappa)} = h^{(\kappa)}$, so the current feasible iterate remains feasible for the inner approximation.

Combining the above bounds, the MM update for a fixed ρ is the convex quadratic program

$$\begin{aligned} \boldsymbol{\pi}^{(\kappa+1)} \in \arg\min_{\boldsymbol{\pi}} \quad & \lambda_{\mathrm{est}} a_\rho^{(\kappa)} \boldsymbol{\pi}^\top D_N \boldsymbol{\pi} + \boldsymbol{\pi}_S^\top A_G^+(\boldsymbol{\lambda})\boldsymbol{\pi}_S - 2(\boldsymbol{\pi}_S^{(\kappa)})^\top A_G^-(\boldsymbol{\lambda})\boldsymbol{\pi}_S \\ \text{s.t.} \quad & \mathbf{1}^\top \boldsymbol{\pi} = 1, \\ & \boldsymbol{\pi} \ge 0, \qquad \pi_0 \ge \epsilon_\pi, \\ & \frac{\pi_k}{n_k} \le C_{\mathrm{reg}} \left[2(\boldsymbol{\pi}^{(\kappa)})^\top D_N \boldsymbol{\pi} - h^{(\kappa)}\right], \quad k = 0, \dots, K. \end{aligned} \tag{56}$$

Since $a_\rho^{(\kappa)} > 0$, $D_N \succ 0$, and $A_G^+ \succeq 0$, the quadratic objective in (56) is strictly convex. Hence, each MM subproblem admits a unique solution.

Let $Q_{\rho,\boldsymbol{\lambda}}^{(\kappa)}(\boldsymbol{\pi})$ denote the objective in (56), i.e.,

$$Q_{\rho,\boldsymbol{\lambda}}^{(\kappa)}(\boldsymbol{\pi}) := \lambda_{\mathrm{est}} a_\rho^{(\kappa)} \boldsymbol{\pi}^\top D_N \boldsymbol{\pi} + \boldsymbol{\pi}_S^\top A_G^+(\boldsymbol{\lambda})\boldsymbol{\pi}_S - 2(\boldsymbol{\pi}_S^{(\kappa)})^\top A_G^-(\boldsymbol{\lambda})\boldsymbol{\pi}_S. \tag{57}$$

The corresponding full MM majorizer differs from $Q_{\rho,\boldsymbol{\lambda}}^{(\kappa)}$ only by the additive constant

$$C_{\rho,\boldsymbol{\lambda}}^{(\kappa)} := \lambda_{\mathrm{est}} \left[\left(h^{(\kappa)}\right)^\rho - a_\rho^{(\kappa)} h^{(\kappa)}\right] + (\boldsymbol{\pi}_S^{(\kappa)})^\top A_G^-(\boldsymbol{\lambda})\boldsymbol{\pi}_S^{(\kappa)}. \tag{58}$$

Hence,

$$F_{\rho,\boldsymbol{\lambda}}(\boldsymbol{\pi}) \le Q_{\rho,\boldsymbol{\lambda}}^{(\kappa)}(\boldsymbol{\pi}) + C_{\rho,\boldsymbol{\lambda}}^{(\kappa)}, \tag{59}$$

with equality at $\boldsymbol{\pi} = \boldsymbol{\pi}^{(\kappa)}$. Since $C_{\rho,\boldsymbol{\lambda}}^{(\kappa)}$ is independent of $\boldsymbol{\pi}$, minimizing the full majorizer is equivalent to solving (56). Moreover, the inner approximation of the regularity constraint contains the current feasible iterate $\boldsymbol{\pi}^{(\kappa)}$. Therefore, the construction guarantees monotonic descent of $F_{\rho,\boldsymbol{\lambda}}$:

$$\begin{aligned} F_{\rho,\boldsymbol{\lambda}}(\boldsymbol{\pi}^{(\kappa+1)}) &\le Q_{\rho,\boldsymbol{\lambda}}^{(\kappa)}(\boldsymbol{\pi}^{(\kappa+1)}) + C_{\rho,\boldsymbol{\lambda}}^{(\kappa)} \\ &\le Q_{\rho,\boldsymbol{\lambda}}^{(\kappa)}(\boldsymbol{\pi}^{(\kappa)}) + C_{\rho,\boldsymbol{\lambda}}^{(\kappa)} \\ &= F_{\rho,\boldsymbol{\lambda}}(\boldsymbol{\pi}^{(\kappa)}). \end{aligned} \tag{60}$$

Initialization. A guaranteed feasible initialization is obtained by setting $w_k^{(0)} = 1$ for all source domains, which gives

$$\pi_k^{(0)} = \frac{n_k}{\sum_{j=0}^{K} n_j}, \qquad k = 0, \dots, K. \tag{61}$$

For this initialization,

$$\frac{\pi_k^{(0)}}{n_k} = \frac{1}{\sum_{j=0}^{K} n_j} = h(\boldsymbol{\pi}^{(0)}),$$

and hence the original regularity constraint is satisfied for every $C_{\mathrm{reg}} \geq 1$.

C.3 Selection of tuning parameters by gradient alignment

The theoretical exponent $2\beta/(d_x + d_z + 2\beta)$ in (27) depends on the unknown Hölder smoothness parameter β. As mentioned in Section C.1, we therefore treat ρ as a tuning parameter and select it from a finite candidate set

$$\mathcal{G}_\rho = \{\rho_1, \dots, \rho_J\} \subset (0, 1),$$

where the upper bound 1 corresponds to the limiting case $\beta \to \infty$. Similarly, the constant C_{reg} in the weight regularity constraint is unknown in practice. We therefore select it from

$$\mathcal{G}_C = \{C_1, \dots, C_L\} \subset (1, C_{\max}],$$

where $C_{\max} < \infty$ is a fixed constant independent of the sample sizes. The calibration parameters $\boldsymbol{\lambda} = (\lambda_{\mathrm{est}}, \lambda_G, \lambda_{\mathrm{diag}})^\top$ introduced in (37) are also selected from the finite simplex grid $\mathcal{G}_\lambda$ defined in (41). Write

$$\mathcal{G}_\lambda = \{\boldsymbol{\lambda}^{(1)}, \dots, \boldsymbol{\lambda}^{(R)}\}, \qquad R := |\mathcal{G}_\lambda| = \binom{M_\lambda - 1}{2}.$$

Although $\boldsymbol{\lambda}$ contains three components, the normalization in (38) leaves only two degrees of freedom.

We do not select these tuning parameters by directly comparing the values of the surrogate objective $F_{\rho,\boldsymbol{\lambda}}$ across different candidates. In particular, changing ρ changes the scale of $h(\boldsymbol{\pi})^\rho$, while changing $\boldsymbol{\lambda}$ changes the relative scales of the estimation and transfer-bias terms. Instead, we jointly select ρ, C_{reg}, and $\boldsymbol{\lambda}$ according to the predicted first-order improvement of the target validation loss.

Let θ denote the parameters of the current score network. For domain k, denote its empirical training loss by $\widehat{\mathcal{L}}_k(\theta)$. Define their gradient by

$$\boldsymbol{u}_k := \nabla_\theta \widehat{\mathcal{L}}_k(\theta), \qquad k = 0, \dots, K. \tag{62}$$

Let

$$\boldsymbol{u}_0^{\mathrm{val}} := \nabla_\theta \widehat{\mathcal{L}}_{0,\mathrm{val}}(\theta) \tag{63}$$

denote the gradient of the target validation loss.

For each candidate triple $\left(\rho_j, C_l, \boldsymbol{\lambda}^{(m)}\right) \in \mathcal{G}_\rho \times \mathcal{G}_C \times \mathcal{G}_\lambda$, we solve (56) with $\rho = \rho_j$, $C_{\mathrm{reg}} = C_l$, and $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{(m)}$ to obtain

$$\widehat{\boldsymbol{\pi}}_{j,l,m} = \left(\widehat{\pi}_{0,j,l,m}, \dots, \widehat{\pi}_{K,j,l,m}\right)^\top.$$

The resulting weighted training loss is

$$\widehat{\mathcal{L}}_{j,l,m}(\theta) := \sum_{k=0}^{K} \widehat{\pi}_{k,j,l,m} \widehat{\mathcal{L}}_k(\theta), \tag{64}$$

whose gradient is

$$\boldsymbol{u}_{j,l,m} := \nabla_\theta \widehat{\mathcal{L}}_{j,l,m}(\theta) = \sum_{k=0}^{K} \widehat{\pi}_{k,j,l,m} \boldsymbol{u}_k. \tag{65}$$

To motivate the selection criterion, consider a single gradient step with learning rate $\eta > 0$,

$$\theta_{j,l,m}^{+} = \theta - \eta \boldsymbol{u}_{j,l,m}.$$

A first-order Taylor expansion of the target validation loss gives

$$\begin{aligned}\widehat{\mathcal{L}}_{0,\mathrm{val}}(\theta^{+}_{j,l,m}) &= \widehat{\mathcal{L}}_{0,\mathrm{val}}(\theta - \eta \boldsymbol{u}_{j,l,m}) \\ &= \widehat{\mathcal{L}}_{0,\mathrm{val}}(\theta) - \eta \left\langle \boldsymbol{u}_0^{\mathrm{val}}, \boldsymbol{u}_{j,l,m} \right\rangle + O(\eta^2). \end{aligned} \tag{66}$$

Hence, to first order, a larger positive inner product

$$\left\langle \boldsymbol{u}_0^{\mathrm{val}}, \boldsymbol{u}_{j,l,m} \right\rangle$$

corresponds to a larger predicted reduction in the target validation loss. We therefore jointly select the tuning parameters according to

$$\left(\widehat{\rho}, \widehat{C}_{\mathrm{reg}}, \widehat{\boldsymbol{\lambda}}\right) \in \underset{\rho_j \in \mathcal{G}_\rho, C_l \in \mathcal{G}_C, \boldsymbol{\lambda}^{(m)} \in \mathcal{G}_\lambda}{\arg\max} \left\langle \boldsymbol{u}_0^{\mathrm{val}}, \boldsymbol{u}_{j,l,m} \right\rangle. \tag{67}$$

Because the weighted gradient is linear in the domain weights, (67) can be evaluated without performing an additional backward pass for every candidate triple. Define the domain-wise target-alignment scores

$$c_k := \left\langle \boldsymbol{u}_0^{\mathrm{val}}, \boldsymbol{u}_k \right\rangle, \qquad k = 0, \ldots, K, \tag{68}$$

and let

$$c := (c_0, \ldots, c_K)^\top.$$

Then

$$\begin{aligned}\left\langle \boldsymbol{u}_0^{\mathrm{val}}, \boldsymbol{u}_{j,l,m} \right\rangle &= \left\langle \boldsymbol{u}_0^{\mathrm{val}}, \sum_{k=0}^{K} \widehat{\pi}_{k,j,l,m} \boldsymbol{u}_k \right\rangle \\ &= \sum_{k=0}^{K} \widehat{\pi}_{k,j,l,m} \left\langle \boldsymbol{u}_0^{\mathrm{val}}, \boldsymbol{u}_k \right\rangle \\ &= c^\top \widehat{\boldsymbol{\pi}}_{j,l,m}. \end{aligned} \tag{69}$$

Consequently, the tuning parameters can equivalently be selected as

$$\boxed{\left(\widehat{\rho}, \widehat{C}_{\mathrm{reg}}, \widehat{\boldsymbol{\lambda}}\right) \in \underset{\rho_j \in \mathcal{G}_\rho, C_l \in \mathcal{G}_C, \boldsymbol{\lambda}^{(m)} \in \mathcal{G}_\lambda}{\arg\max} c^\top \widehat{\boldsymbol{\pi}}_{j,l,m}.} \tag{70}$$

This criterion has a direct domain-adaptation interpretation. If

$$c_k = \left\langle \boldsymbol{u}_0^{\mathrm{val}}, \boldsymbol{u}_k \right\rangle > 0,$$

then a gradient step using domain k is locally aligned with a descent direction of the target validation loss. Conversely, $c_k < 0$ indicates a locally conflicting training direction and potential negative transfer. Hence, (70) favors tuning-parameter configurations whose MM-QP solutions place larger weights on domains whose training gradients are beneficial to the target domain.

Since $\widehat{C}_{\mathrm{reg}} \leq C_{\max}$ and $C_{\max}$ is fixed independently of the sample sizes, the selected weights continue to satisfy the weight regularity condition with a sample-size-independent constant.

After selecting $\left(\widehat{\rho}, \widehat{C}_{\mathrm{reg}}, \widehat{\boldsymbol{\lambda}}\right)$, we use the corresponding MM-QP solution

$$\widehat{\boldsymbol{\pi}} := \widehat{\boldsymbol{\pi}}_{\widehat{\rho}, \widehat{C}_{\mathrm{reg}}, \widehat{\boldsymbol{\lambda}}}$$

for weighted-ERM training, and recover the original source weights via

$$\widehat{w}_k = \frac{n_0 \widehat{\pi}_k}{n_k \widehat{\pi}_0}, \qquad k = 1, \ldots, K. \tag{71}$$

In practice, neither the tuning parameters nor the MM-QP weights are updated at every epoch. We first warm up the score model using only target-domain data, then jointly select $(\widehat{\rho}, \widehat{C}_{\mathrm{reg}}, \widehat{\boldsymbol{\lambda}})$ by gradient alignment and solve the corresponding MM-QP for $\widehat{\boldsymbol{\pi}}$. The model is subsequently trained for several epochs with these weights fixed, after which the tuning parameters and weights are updated again.

Computational complexity. For a fixed triple $(\rho, C_{\mathrm{reg}}, \boldsymbol{\lambda})$, each MM iteration requires the eigendecomposition in (49) and the solution of a convex quadratic program of dimension $K+1$ in (56). Both operations have computational complexity $\mathcal{O}(K^3)$, so the overall complexity of each MM iteration remains $\mathcal{O}(K^3)$.

Let

$$J = |\mathcal{G}_\rho|, \qquad L = |\mathcal{G}_C|, \qquad R = |\mathcal{G}_\lambda|.$$

The MM-QP weight selection across all candidate triples therefore costs $\mathcal{O}(JLRK^3)$ per MM iteration. If the number of MM iterations is treated as fixed, the overall cost of each tuning-parameter selection step is also $\mathcal{O}(JLRK^3)$. Moreover, since $A_G(\boldsymbol{\lambda})$ depends only on $\boldsymbol{\lambda}$, its eigendecomposition can be reused across all (ρ, C_{reg}) pairs associated with the same $\boldsymbol{\lambda}$. Since K is typically much smaller than the number of score-network parameters, this optimization cost is negligible relative to score-network training.

Moreover, the gradient-alignment criterion does not require one backward pass for each candidate triple $(\rho, C_{\mathrm{reg}}, \boldsymbol{\lambda})$. The domain gradients $\boldsymbol{u}_0, \ldots, \boldsymbol{u}_K$ and the target validation gradient $\boldsymbol{u}_0^{\mathrm{val}}$ are computed only once. The alignment score for each candidate triple is then obtained from the inexpensive inner product $c^\top \widehat{\boldsymbol{\pi}}_{j,l,m}$.

Detailed training behavior of DA-Diff and MM-QP can be found in Appendix H.

D Data generation mechanism of the synthetic experiments in Section 4

D.1 Gaussian mixture data

We first introduce the data-generating mechanism for the Gaussian mixture data used to evaluate DA-Diff and other transfer-learning methods in Section 4.1.

Let $X_0 \in \mathbb{R}^5$ follow a Gaussian mixture $X_0 \sim \sum_{i=1}^{10} 0.1 \cdot \mathcal{N}(\mu_i, \Sigma_i)$, where $\mathcal{N}$ is the density of the normal distribution. We consider the linear transform:

$$Z_0 = AX_0 + \sigma\,\varepsilon, \qquad \varepsilon \sim N(0, I_5),$$

where $A \in \mathbb{R}^{5\times 5}$ is fixed and $\sigma > 0$. Conditioning on $Z_0 = z$, the posterior remains a Gaussian mixture with analytically available parameters:

$$P_0(\cdot|Z_0 = z) \propto \textstyle\sum_{i=1}^{10} \widetilde{\pi}_i \cdot \mathcal{N}(\widetilde{\mu}_i, \widetilde{\Sigma}_i), \tag{72}$$

where $\widetilde{\Sigma}_i = \left(\Sigma_i^{-1} + A^\top A/\sigma^2\right)^{-1}$, $\widetilde{\mu}_i = \widetilde{\Sigma}_i\left(\Sigma_i^{-1}\mu_i + A^\top z/\sigma^2\right)$, and the mixture weights satisfy $\widetilde{\pi}_i = 0.1 \cdot \exp[0.5(\widetilde{\mu}_i^\top \widetilde{\Sigma}_i^{-1}\widetilde{\mu}_i - \mu_i^\top \Sigma_i^{-1}\mu_i)]/\sqrt{|\Sigma_i|}$.

We draw $\mu_i \sim U([-1.5, 1.5]^5)$ and set $\Sigma_i = 0.01 I_5$ for $i = 1, \ldots, 10$, $A = 0.25\,\mathbf{1}\mathbf{1}^\top + 0.25 I_5$, and $\sigma = 0.5$. We generate three K domains. The sample sizes are set to $n_0 = n_1 = \cdots = n_K = 500$. Let $d^{(k)}$ denote the shift strength of the k-th source. We consider five near-source domains, with shift strengths lying on the linear interpolation segment from 0.1 toward 0.2, and five far-source domains, with shift strengths lying on the segment from 0.2 toward 0.5. For each source, we introduce a controlled distribution shift by perturbing the mixture parameters as:

$$\begin{aligned}
\mu_i^{(k)} &= \mu_i + d^{(k)}\,\xi_i^{(k)}, \qquad \xi_i^{(k)} \sim N(0, I_5),\\
\Sigma_i^{(k)} &= \Sigma_i + \left[d^{(k)}\,\mathrm{diag}(u_i^{(k)})\right]^2, \qquad u_i^{(k)} \sim U([0,1]^5),\\
A_{ij}^{(k)} &= A_{ij} + d^{(k)}\,\psi_{ij}^{(k)}, \qquad \psi_{ij}^{(k)} \sim N(0,1),\\
\sigma^{(k)} &= \max\{\sigma + d^{(k)}\gamma^{(k)}, 10^{-4}\}, \qquad \gamma^{(k)} \sim N(0,1).
\end{aligned} \tag{73}$$

D.2 Post-nonlinear model

In this subsection, we describe the data-generating mechanism for the synthetic experiments used to evaluate DA-CIT and the competing conditional independence tests. Under the null hypothesis,

the data-generating mechanism follows the fork structure $X \leftarrow Z \rightarrow Y$. Under the alternative, conditional dependence is introduced through a direct effect of Y on X. We construct multiple source domains by perturbing the covariate distribution and the conditional mechanisms relating Z to X and Y, with the transferability of each source controlled by its shift strength $d^{(k)}$.

For the target domain, we generate $Z_0 \sim N(0, I_{d_z})$, and independently draw $\beta_X, \beta_Y \sim N(0, I_{d_z})$. Given Z_0, we generate

$$Y_0 = g_Y\left(Z_0^\top \beta_Y / \|\beta_Y\|_2 + 0.5 \cdot \epsilon_Y\right), \quad X_0 = g_X\left(Z_0^\top \beta_X / \|\beta_X\|_2 + 0.5 \cdot \epsilon_X\right) + h_1 \cdot Y_0. \tag{74}$$

Here, $h_1 \geq 0$ controls the strength of the conditional dependence between X_0 and Y_0 given Z_0. When $h_1 = 0$, X_0 and Y_0 are CI given Z_0. As h_1 increases, the conditional dependence between X_0 and Y_0 becomes stronger, resulting in a larger departure from conditional independence. Moreover, $\epsilon_X, \epsilon_Y \overset{\text{i.i.d.}}{\sim} N(0,1)$, and the functions g_X and g_Y are independently and uniformly sampled from $\texttt{func_set} = \left\{x,\, x^2,\, \cos(x),\, \sin(x),\, \tanh(x),\, \exp(-|x|)\right\}$.

For source domain k, let $d^{(k)}$ denote its shift strength. We perturb the data generating mechanism through the regression mechanism, nonlinear functions, noise scale, and marginal distribution of Z. Specifically, the regression mechanism is shifted as

$$\beta_j^{(k)} = \beta_j + d^{(k)} \xi_j^{(k)}, \qquad \xi_j^{(k)} \sim N(0, I_{d_z}), \qquad j \in \{X, Y\}.$$

The noise scale is perturbed according to

$$\sigma^{(k)} = 0.5 + d^{(k)} u^{(k)}, \qquad u^{(k)} \sim U(0,1). \tag{75}$$

For the nonlinear functions, with probability $1 - \exp(-d^{(k)})$, both $g_X^{(k)}$ and $g_Y^{(k)}$ are independently resampled from $\texttt{func_set}$; otherwise, the target functions g_X and g_Y are retained.

We also include a covariate shift in Z. For each source domain,

$$Z_k \sim N\left(\mu^{(k)}, \operatorname{diag}\left((s_Z^{(k)})^2\right)\right), \tag{76}$$

where

$$\begin{aligned} \mu^{(k)} &= 0.1 d^{(k)} \cdot \eta^{(k)}, \quad \eta^{(k)} \sim N(0, I_{d_z}), \\ s_Z^{(k)} &= \exp(0.1 d^{(k)} \cdot \nu^{(k)}), \quad \nu^{(k)} \sim N(0, I_{d_z}), \end{aligned} \tag{77}$$

with the exponential applied element-wise. The source variables are then generated as

$$Y_k = g_Y^{(k)}\left(Z_k^\top \beta_Y^{(k)} / \|\beta_Y^{(k)}\|_2 + \sigma^{(k)} \cdot \epsilon_{Y,k}\right), \quad X_k = g_X^{(k)}\left(Z_k^\top \beta_X^{(k)} / \|\beta_X^{(k)}\|_2 + \sigma^{(k)} \cdot \epsilon_{X,k}\right) + h_1 \cdot Y_k, \tag{78}$$

where $\epsilon_{X,k}, \epsilon_{Y,k} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. We use seven near sources and three far sources. The shift strengths of the near sources are linearly interpolated from 0.01 to 0.1, while those of the far sources are linearly interpolated from 0.2 to 0.5.

E Ablation and sensitivity studies

In this section, we conduct ablation and sensitivity analyses of DA-Diff and DA-CIT using the data-generation mechanisms described in Appendices D.1 and D.2. UOWQ has apparent negative transfer performance in Table 1, thus we skip the experiments of UOWQ.

E.1 Sensitivity analysis on Gaussian mixture data

We study the sensitivity of DA-Diff to the target and source sample sizes, the number of source domains, and several key hyperparameters.

Table 2 reports the performance of DA-Diff across different sample sizes and numbers of source domains. DA-Diff achieves the best performance for all considered values of K when the sample size is 1000 or 1500, and remains among the top-performing methods when the sample size is 2000.

Table 2: Performance of the proposed DA-Diff, Target-only, Cat-all, Finetune, DoG (Zhong et al., 2025), TGDP (Ouyang et al., 2024), and Robust (Konstantinov & Lampert, 2019) under the data-generating mechanism in Section 4.1, with sample size $n_0 = \cdots = n_K = 1000, 1500, 2000$. We report the W_2 distance along with one standard deviation (std).

Sample size	Method	W_2 ($K = 2$)	W_2 ($K = 5$)	W_2 ($K = 10$)	W_2 ($K = 20$)
1000	DA-Diff (Ours)	**0.142**(0.039)	**0.140**(0.037)	**0.133**(0.030)	**0.134**(0.027)
	Target-only	0.376(0.070)	0.376(0.070)	0.376(0.070)	0.376(0.070)
	Cat-all	0.440(0.112)	0.371(0.087)	0.334(0.088)	0.308(0.063)
	Finetune	0.216†(0.056)	0.203(0.057)	0.173(0.045)	0.164(0.040)
	DoG (2025)	0.285(0.103)	0.231(0.066)	0.216(0.064)	0.181(0.046)
	TGDP (2024)	0.562(0.135)	0.413(0.116)	0.389(0.104)	0.328(0.074)
	Robust (2019)	0.175(0.055)	0.219†(0.062)	0.208†(0.070)	0.176†(0.047)
1500	DA-Diff (Ours)	**0.111**(0.025)	**0.113**(0.022)	**0.106**(0.024)	**0.110**(0.027)
	Target-only	0.259(0.047)	0.259(0.047)	0.259(0.047)	0.259(0.047)
	Cat-all	0.336(0.098)	0.339(0.073)	0.322(0.081)	0.316(0.071)
	Finetune	0.226(0.052)	0.154(0.037)	0.135(0.035)	0.121(0.032)
	DoG (2025)	0.197†(0.058)	0.164†(0.048)	0.147†(0.037)	0.138†(0.032)
	TGDP (2024)	0.558(0.131)	0.428(0.118)	0.370(0.096)	0.337(0.083)
	Robust (2019)	0.150(0.050)	0.200(0.062)	0.176(0.053)	0.178(0.068)
2000	DA-Diff (Ours)	**0.095**(0.022)	**0.098**(0.027)	**0.090**(0.022)	0.097(0.022)
	Target-only	0.123†(0.018)	0.123(0.018)	0.123(0.018)	0.123(0.018)
	Cat-all	0.291(0.075)	0.335(0.114)	0.330(0.080)	0.310(0.069)
	Finetune	0.105(0.017)	0.103(0.041)	0.092(0.019)	**0.085**(0.020)
	DoG (2025)	0.165(0.057)	0.121†(0.035)	0.113†(0.028)	0.109†(0.031)
	TGDP (2024)	0.482(0.128)	0.421(0.121)	0.370(0.093)	0.328(0.081)
	Robust (2019)	0.123(0.037)	0.202(0.060)	0.162(0.043)	0.161(0.054)

These results suggest that DA-Diff is relatively robust to changes in sample size and the number of source domains.

Figure 3 further examines the sensitivity of DA-Diff to four hyperparameters: the number of time bins B_T in Section H, the weight/tuning update interval `update_every` in Algorithm 1, the grid resolution of $\mathcal{G}_\lambda$ in (41), and the number of warmup epochs e_w in Algorithm 1. The W_2 distance varies only moderately across the considered hyperparameter values, with largely overlapping error bars. Overall, DA-Diff shows limited sensitivity to these hyperparameter choices.

E.2 Sensitivity analysis on the post-nonlinear model

We further investigate the sensitivity of DA-CIT to the number of source domains, source–target similarity, the target and source sample sizes, and the strength of conditional dependence under H_1. Following the setting in Section 4.2, we fix $d_z = 20$ throughout and vary one factor at a time while keeping the remaining settings at their default values. Unless otherwise specified, we use $K = 10$ source domains, consisting of seven near sources and three far sources, and set $n_0 = n_1 = \cdots = n_K = 500$. The construction of near/far sources is introduced in Appendix D.2. The parameter h_1 controls the strength of conditional dependence under H_1, with larger values corresponding to stronger departures from conditional independence. Specifically, we consider

$$
\begin{aligned}
&K \in \{2, 5, 10, 20\},\\
&\text{Near sources: Far sources} \in \{10:0, 7:3, 3:7, 0:10\},\\
&n_0 = n_1 = \cdots = n_K \in \{500, 1000, 1500, 2000\},\\
&n_1 = \cdots = n_K = 500, \quad n_0 \in \{400, 500, 1000, 1500\},\\
&n_0 = 500, \quad n_1 = \cdots = n_K \in \{500, 1000, 2000\},\\
&h_1 \in \{0.05, 0.1, 0.2, 0.3, 0.5\}.
\end{aligned}
$$

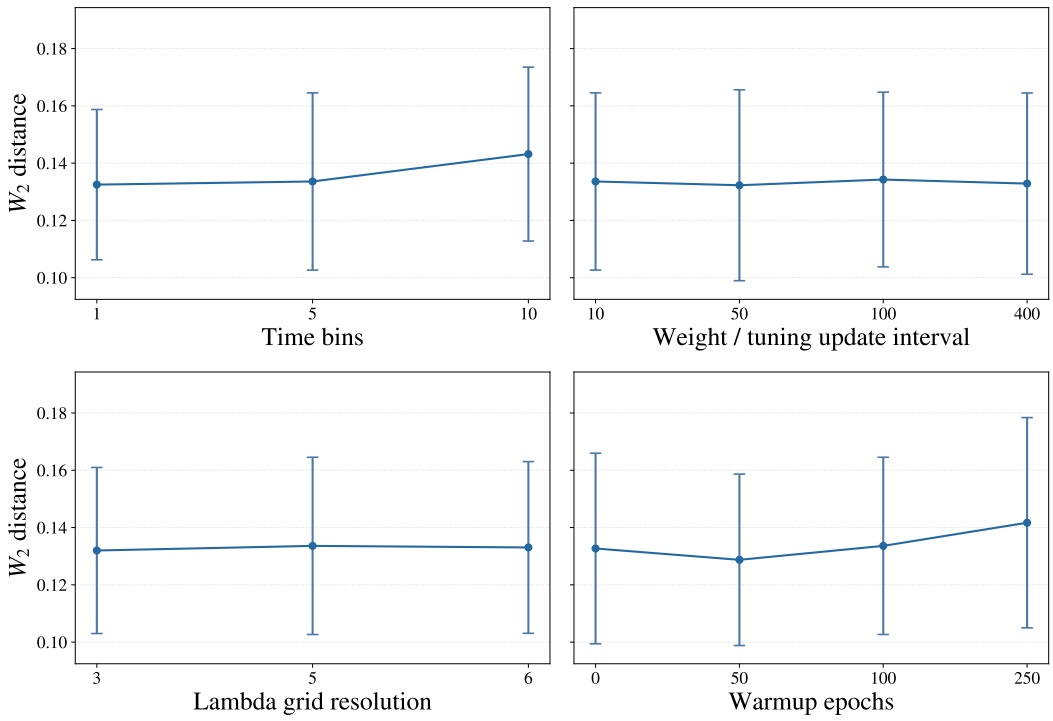


Figure 3: Sensitivity study of DA-Diff on the Gaussian mixture data. We vary the number of time bins B_T, the weight/tuning update interval `update_every`, the resolution of the $\mathcal{G}_\lambda$, and the number of warmup epochs e_w, and evaluate the performance using the W_2 distance.

Figure 4 summarizes the sensitivity results. Overall, DA-CIT remains stable across changes in the number of source domains, the near–far source composition, and the target and source sample sizes. In most settings, DA-CIT maintains Type I error at or below the nominal level and consistently yields lower Type I error than CDCIT. At the same time, DA-CIT achieves higher power than CDCIT and SGMCIT in most settings. As h_1 increases, the power of DA-CIT increases, indicating that stronger conditional dependence under H_1 is more readily tested. These results suggest that DA-CIT can effectively leverage heterogeneous source-domain information while maintaining reliable Type I error control.

For comparison, we also consider a naive pooling strategy in which the target and source samples are directly concatenated before CI testing. As shown in Figure 5, this strategy generally leads to substantial Type I error inflation for the competing methods. In contrast, DA-CIT remains substantially better calibrated, highlighting the importance of accounting for source–target heterogeneity rather than naively pooling observations across domains.

E.3 Empirical cumulative distribution functions of CI-test p-values

We additionally examine the empirical distribution of the p-values produced by different CI testing methods under H_0. As a diagnostic of null calibration, we compare the empirical cumulative distribution function (ECDF) of the p-values with the CDF of $U(0,1)$. We also report the p-value of the Kolmogorov–Smirnov (KS) test, denoted by p_{KS}, based on the 100 p-values obtained across repeated experiments.

As shown in Figure 6, the ECDFs produced by DA-CIT closely follow the CDF of $U(0,1)$ across all considered conditioning dimensions d_z. Moreover, none of the corresponding KS tests rejects the null of uniformity at the 0.05 significance level.

Figure 7 further shows that this behavior remains stable across the sensitivity settings. The ECDFs of DA-CIT remain close to the CDF of $U(0,1)$ as we vary the number of source domains, the number of far-source domains, and the target and source sample sizes. None of the KS tests rejects

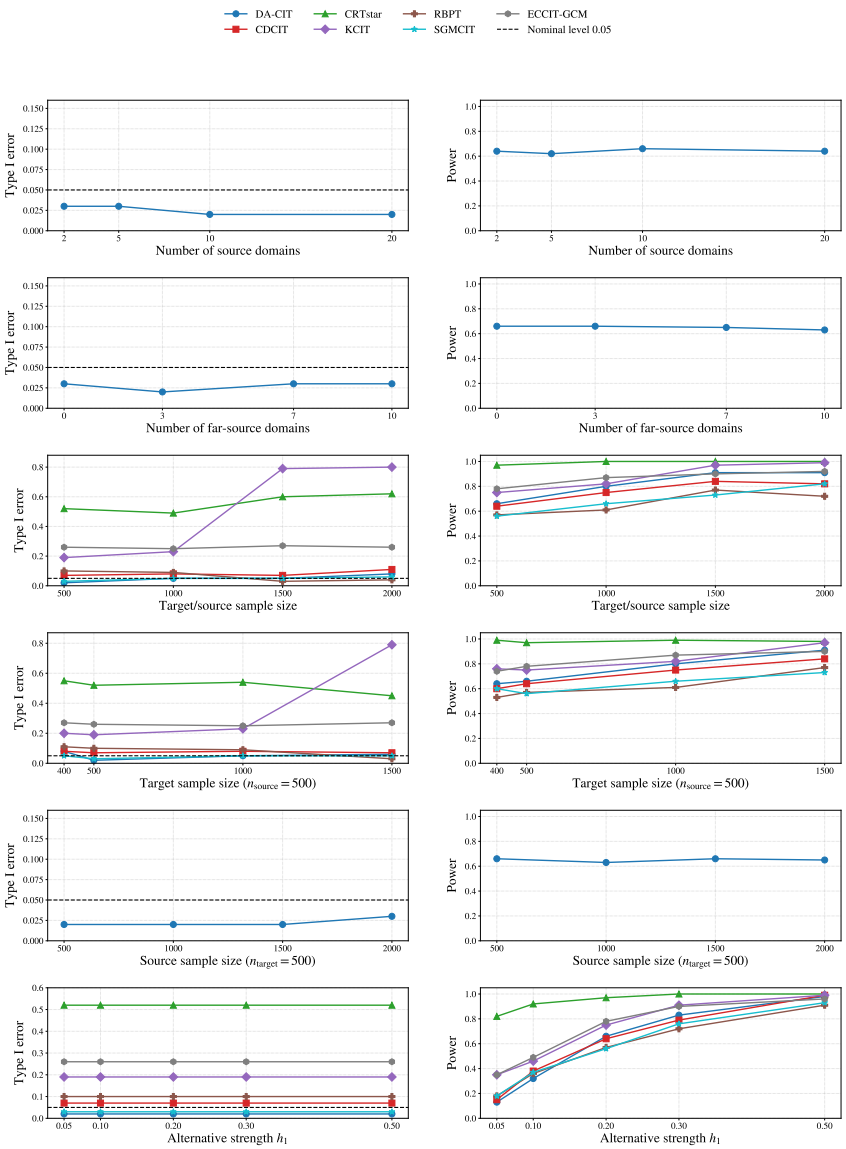


Figure 4: Sensitivity results of DA-CIT on the post-nonlinear data with $d_z = 20$. We vary the number of source domains, the proportion of near and far sources, the target and source sample sizes, and the alternative strength h_1.

uniformity at the 0.05 significance level in these settings. These results provide additional evidence that DA-CIT maintains stable null calibration across heterogeneous multi-source settings.

F Real data analysis

We evaluate our method on the Sachs flow-cytometry dataset (Sachs et al., 2005), a widely used real-world benchmark for causal discovery. The dataset contains simultaneous single-cell measurements of 11 phosphorylated proteins and phospholipids involved in human T-cell signaling: Raf, Mek, Plcγ, PIP2, PIP3, Erk, Akt, PKA, PKC, P38, and Jnk.

The Sachs dataset consists of 14 sub-datasets collected under different experimental conditions involving cellular stimulation and molecular interventions. Each sub-dataset contains the same 11 variables and 707–927 single-cell observations. We treat each experimental condition as a sepa-

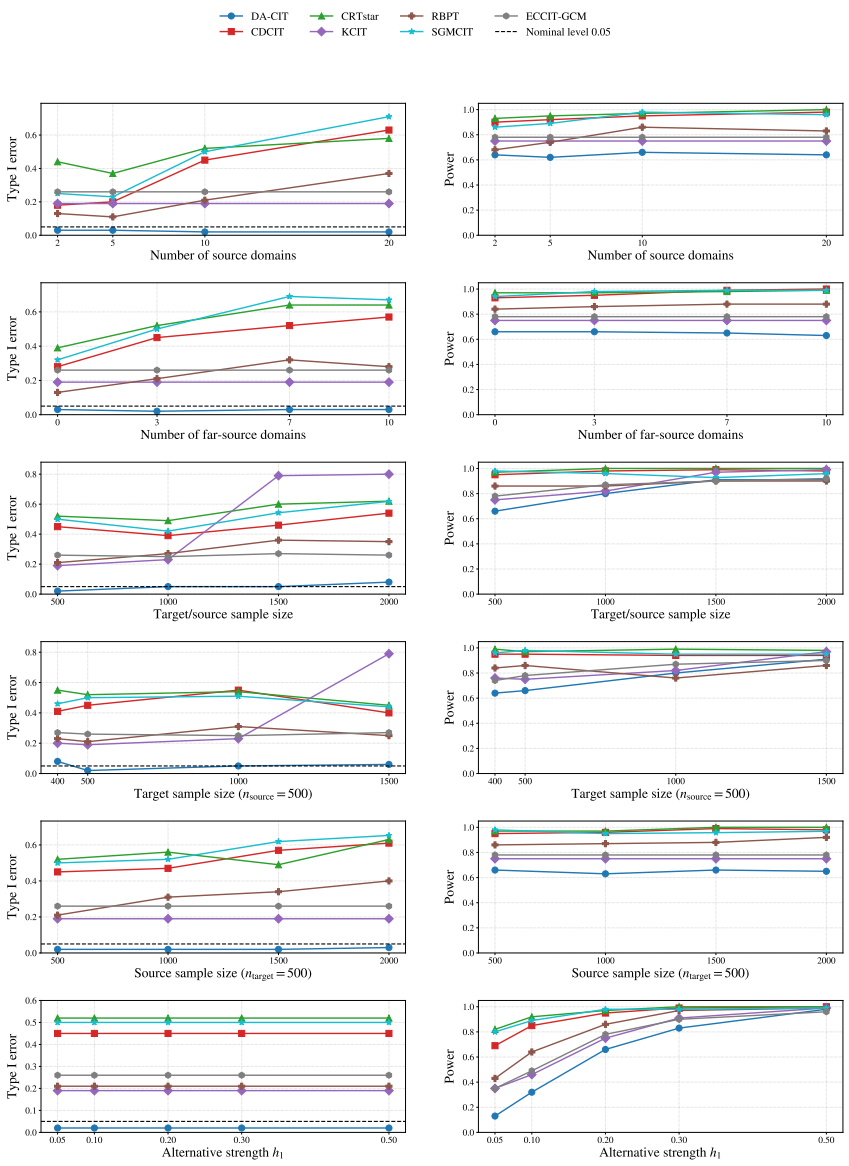


Figure 5: Ablation results on the post-nonlinear data with $d_z = 20$ under the Cat-all setting. We vary the number of source domains, the proportion of near and far sources, the target and source sample sizes, and the alternative strength h_1.

rate domain rather than as an independent replicate from a common distribution, yielding a natural multi-source domain adaptation setting with heterogeneous distributions across domains.

Specifically, we use the `cd3cd28` condition as the target domain. This condition corresponds to T-cell activation through CD3 and CD28 stimulation without an additional inhibitor or ICAM-2 stimulation. The remaining 13 conditions are treated as source domains. These source domains include: (i) CD3/CD28 stimulation combined with one additional molecular perturbation; (ii) CD3/CD28 stimulation together with ICAM-2 stimulation; (iii) PMA or β_2cAMP stimulation; and (iv) additional CD3/CD28+ICAM-2 conditions combined with molecular perturbations. Table 3 summarizes the target and source domains used in our experiments.

To visualize the distributional heterogeneity across domains, we project samples from all domains into two dimensions using t-SNE. Figure 8 shows the resulting representation, where each color corresponds to one experimental domain and label 0 denotes the target domain. The target sam-

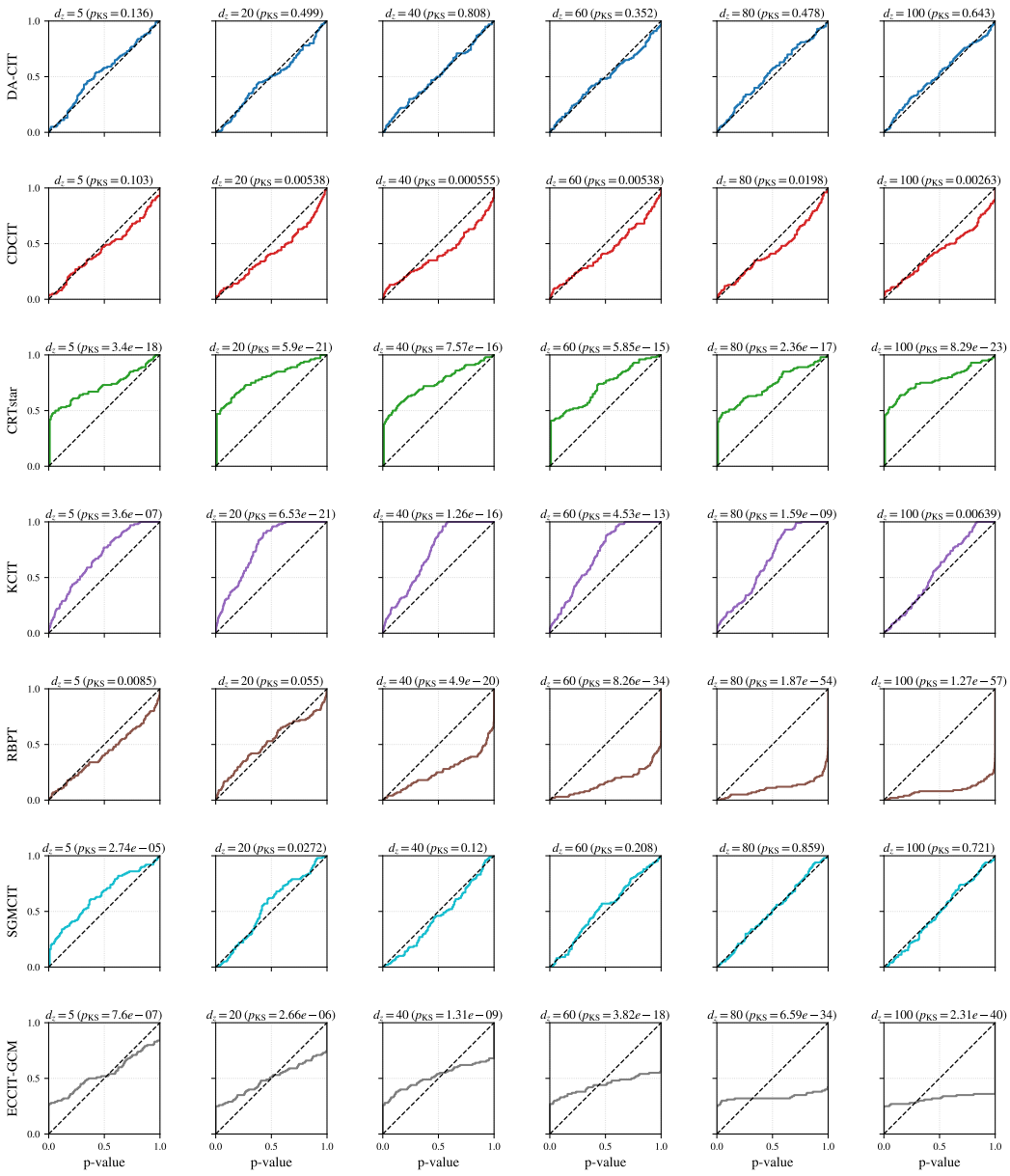


Figure 6: ECDFs of the p-values under H_0 for the CI testing methods considered in Figure 1. The dashed diagonal line represents the CDF of $U(0,1)$, and p_{KS} denotes the p-value of the KS test based on 100 null p-values.

ples are broadly distributed over the main high-density region of the embedding and substantially overlap with samples from several source domains, suggesting that a subset of the source domains shares similar distributional characteristics with the target. At the same time, several source domains exhibit visibly shifted structures from the main sample population. This motivates selectively exploiting source information rather than naively pooling all source observations.

Following Li et al. (2024), we select 50 CI and 40 non-CI (X_0, Y_0, Z_0) triples from the reference causal graph of the Sachs dataset (Sachs et al., 2005). In all triples, X and Y are one-dimensional, while the dimension of Z ranges from 1 to 9. Treating CI as the positive class, we evaluate each method using precision, recall, and F1 score, defined as $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$, $\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$, and $2\,\mathrm{Precision}\,\mathrm{Recall}/(\mathrm{Precision}+\mathrm{Recall})$, respectively.

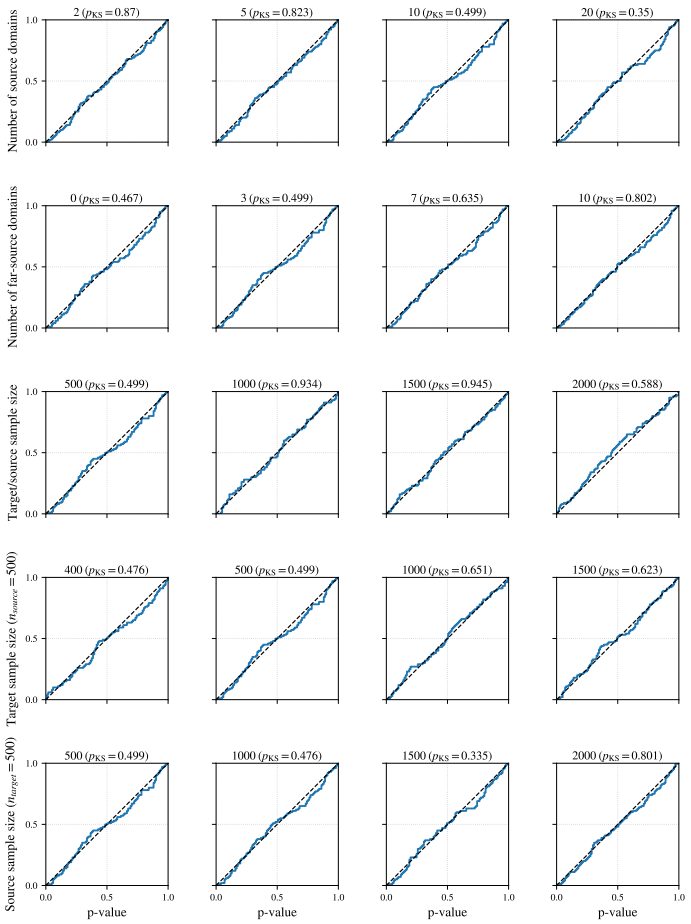


Figure 7: ECDFs of the DA-CIT p-values under H_0 for the ablation settings considered in Figures 4 and 5, with $d_z = 20$. The dashed diagonal line represents the CDF of $U(0, 1)$, and p_{KS} denotes the p-value of the KS test based on 100 null p-values.

We compare DA-CIT with the six baselines introduced in Section 4.2 on the Sachs dataset. Except for DA-CIT and CRT*, which are able to leverage external source data directly, we evaluate each baseline under both the Target-only and Cat-all settings. To improve the robustness of the results, we repeat the experiment five times with different random seeds for each CI or non-CI triple, and determine the final prediction by majority vote over the five runs.

Table 4 summarizes the results. DA-CIT achieves the highest F1 score among the compared methods, while maintaining both high precision and high recall. In comparison, CRT* achieves high precision but suffers from substantially lower recall, while methods such as CDCIT and RBPT attain high recall at the cost of lower precision. KCIT also performs competitively under the Target-only setting, but its performance deteriorates markedly when all domains are concatenated.

For most competing methods, Cat-all performs worse than the corresponding Target-only setting, particularly in terms of recall and F1 score. This observation is consistent with our motivation that naively pooling heterogeneous source and target data may introduce substantial transfer bias. In contrast, DA-CIT selectively exploits external source information and achieves the best F1 score while maintaining both high precision and high recall.

G Algorithms of DA-Diff and DA-CIT

The algorithms of DA-Diff and DA-CIT are provided in ALgorithm 1 and 2.

Table 3: The 14 domains used in the Sachs experiment. The `cd3cd28` condition is used as the target domain, while the remaining 13 conditions are treated as source domains. The domain labels are consistent with those used in our implementation and visualization.

Label	Role	Dataset	Stimulation / intervention condition	sample size
0	Target	`cd3cd28`	CD3 + CD28	853
1	Source	`b2camp`	β_2cAMP	707
2	Source	`cd3cd28_aktinhib`	CD3 + CD28 + Akt inhibitor	911
3	Source	`cd3cd28_g0076`	CD3 + CD28 + Gö6976	723
4	Source	`cd3cd28_ly`	CD3 + CD28 + LY294002	848
5	Source	`cd3cd28_psitect`	CD3 + CD28 + Psitectorigenin	810
6	Source	`cd3cd28_u0126`	CD3 + CD28 + U0126	799
7	Source	`cd3cd28icam2`	CD3 + CD28 + ICAM-2	902
8	Source	`cd3cd28icam2_aktinhib`	CD3 + CD28 + ICAM-2 + Akt inhibitor	899
9	Source	`cd3cd28icam2_g0076`	CD3 + CD28 + ICAM-2 + Gö6976	753
10	Source	`cd3cd28icam2_ly`	CD3 + CD28 + ICAM-2 + LY294002	927
11	Source	`cd3cd28icam2_psit`	CD3 + CD28 + ICAM-2 + Psitectorigenin	868
12	Source	`cd3cd28icam2_u0126`	CD3 + CD28 + ICAM-2 + U0126	759
13	Source	`pma`	PMA	913

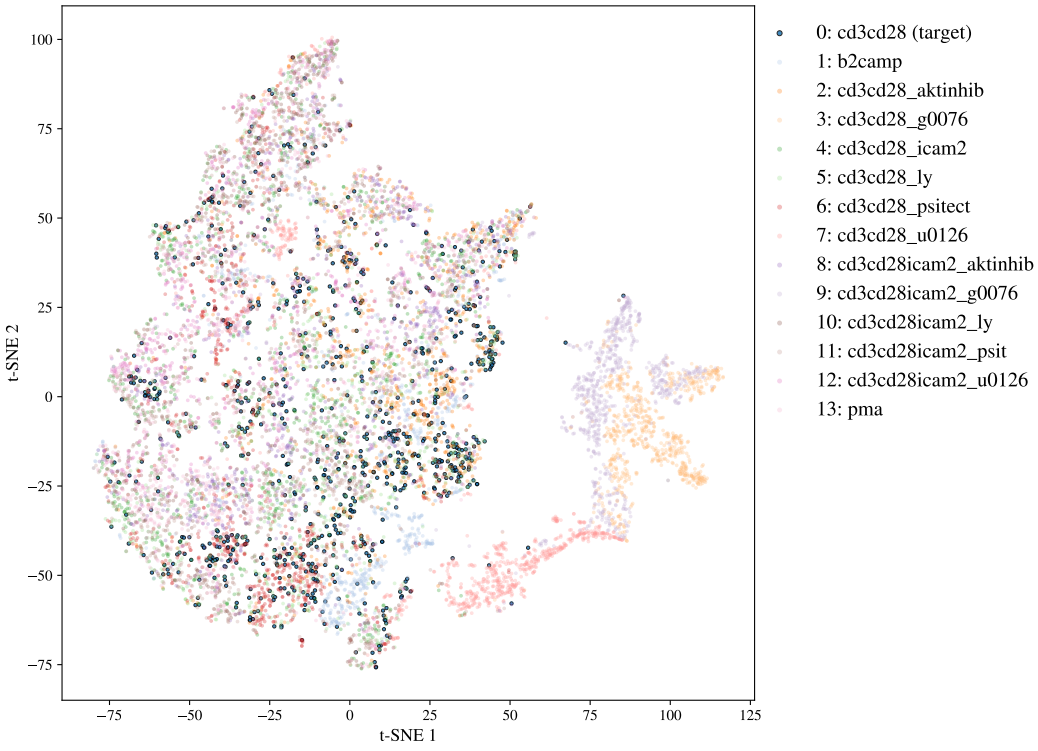


Figure 8: t-SNE visualization of the 14 domains in the Sachs dataset. Each color represents a different stimulation or intervention condition. The `cd3cd28` dataset is used as the target domain, and the remaining 13 datasets are used as source domains.

H Training behavior of DA-Diff and MM-QP

Although our theoretical results and weighted ERM formulation are based on the entire diffusion-time interval $t \in [t_0, T]$, in practice, we can partition $[t_0, T]$ into B_T time bins and perform weighted ERM separately within each bin. The intuition behind this strategy is that the degree of distribution shift across domains varies with the diffusion time. At $t = 0$, the distribution shifts among different source domains are fully preserved. As the forward diffusion process progressively adds noise to the variables, these shifts are gradually attenuated, and as $t \to \infty$, the distributions across do-

Table 4: Performance of different CI testing methods on the Sachs dataset. Each entry reports *Target-only / Cat-all*. DA-CIT and CRT* report a single result because both methods directly leverage external source data, making the Target-only/Cat-all distinction unnecessary. The best, second-best, and third-best results are highlighted in bold, underlined, and †, respectively.

Method	Precision	Recall	F-score
DA-CIT (Ours)	0.738†	<u>0.960</u>	**0.834**
CRT* (2026c)	**0.764**	0.260	0.388
CDCIT (2025a)	0.671 / 0.692	**0.980** / 0.540	0.796† / 0.606
KCIT (2011)	0.725 / <u>0.750</u>	0.900† / 0.060	<u>0.803</u> / 0.111
RBPT (2023)	0.620 / 0.625	**0.980** / 0.600	0.759 / 0.612
ECCIT-GCM (2026)	**0.764** / 0.733	0.520 / 0.220	0.619 / 0.338
SGMCIT (2025)	0.721 / 0.285	0.880 / 0.040	0.792 / 0.070

Algorithm 1 Training the Domain-adapted diffusion model (DA-Diff)

Input: Dataset from the target domain $D_0 = \{(x_{0,i}, z_{0,i})\}_{i=1}^{n_0}$, datasets from K source domains $D_k = \{(x_{k,i}, z_{k,i})\}_{i=1}^{n_k}$ and the corresponding pre-trained source models $\widehat{s}_k$ trained on D_k (for $k = 1, \ldots, K$), the grid $\mathcal{G}_M$ of M, the grid $\mathcal{G}_\rho$ of ρ, the number of the warm-up epochs e_w, the weight-update interval `update_every`, the total number all epochs E.
Output: A neural network $\widehat{s}_w$ trained by weighted ERM.

Initialize a neural network $\widehat{s}_w(x, z, t)$
for e in $1, \ldots, E$ **do**
  (warm-up phase)
  if $e \leq e_w$ **then**
    Compute $\widehat{\mathcal{L}}_0(\widehat{s}_w) = \frac{1}{n_0} \sum_{i=1}^{n_0} \ell(x_{0,i}, z_{0,i}; \widehat{s}_w)$
    Take optimization step on $\nabla \widehat{\mathcal{L}}_0(\widehat{s}_w)$ and update the parameters of $\widehat{s}_w$
  (weighted ERM phase)
  else
    if $(e - e_w - 1) \bmod \texttt{update_every} = 0$ **then**
      Calculate $\widehat{G}_{ij}$ for $i, j = 1, \ldots, K$ by (437) and (438) in parallel and obtain $\widehat{G}$
      For each $(\rho, C_{\text{reg}}, \boldsymbol{\lambda}) \in \mathcal{G}_\rho \times \mathcal{G}_C \times \mathcal{G}_\lambda$, iteratively solve $\widehat{\boldsymbol{\pi}}_{j,l,m}$ using the updates in (56), executed in parallel
      Select $(\rho^*, C^*_{\text{reg}}, \boldsymbol{\lambda}^*)$ that maxmizes the objective in (70) and obtain the corresponding $\boldsymbol{\pi}^*$
      Solve $w_1^*, \ldots, w_K^*$ and W_N^* from $\boldsymbol{\pi}^*$ by (34)
    end if
    Compute $\widehat{\mathcal{L}}_w(\widehat{s}_w) = \left[\sum_{i=1}^{n_0} \ell(x_{0,i}, z_{0,i}; \widehat{s}_w) + \sum_{k=1}^{K} w_k^* \sum_{i=1}^{n_k} \ell(x_{k,i}, z_{k,i}; \widehat{s}_w)\right] / (n_0 + W_N^*)$
    Take optimization step on $\nabla \widehat{\mathcal{L}}_w(\widehat{s}_w)$ and update the parameters of $\widehat{s}_w$
  end if
end for
Return $\widehat{s}_w$

mains converge to the same limiting distribution, so that the domain shifts asymptotically vanish. Therefore, estimating separate weights $w^*(t)$ for different time bins and performing weighted ERM accordingly can better account for the time-varying transferability of the source domains than using a single set of weights over the entire diffusion-time interval. Ablation studies on the number of time bins are provided in Figure 3, Appendix E.

We next examine the training behavior of DA-Diff trained using the MM-QP procedure proposed in Appendix C on the Gaussian mixture data introduced in Appendix D.1. For training stability, we consider only three source domains and set the sample sizes of the target and source domains to $n_0 = n_1 = n_2 = n_3 = 5000$. The shift strengths $d^{(k)}$ defined in Appendix D.1 are set to $d^{(1)} = 0.01, d^{(2)} = 0.1, d^{(3)} = 0.2$. The histograms of the target and source data are shown in Figure 9.

Algorithm 2 Domain-adapted conditional independence test (DA-CIT)

Input: Target dataset $D_0 = \{(x_{0,i}, y_{0,i}, z_{0,i})\}_{i=1}^{n_0}$, trained neural network $\widehat{s}_w$, the number of repetitions B.

Output: The p-value p.

1: Let $\mathbf{X}_0 = (x_{0,1}, \ldots, x_{0,n_0})^\top, \mathbf{Y}_0 = (y_{0,1}, \ldots, y_{0,n_0})^\top, \mathbf{Z}_0 = (z_{0,1}, \ldots, z_{0,n_0})^\top$

2: Compute $\mathbb{T} = \mathbb{T}(\mathbf{X}_0, \mathbf{Y}_0, \mathbf{Z}_0)$

3: Concatenate $\mathbf{Z}_0$ for B times and use the reverse process in (5) with the estimated neural network $\widehat{s}_w$ to generate $\mathbf{X}_0^{(b)}$ for $b = 1, \ldots, B$ in one shot

4: Compute $\mathbb{T}^{(b)} = \mathbb{T}(\mathbf{X}_0^{(b)}, \mathbf{Y}_0, \mathbf{Z}_0)$ for $b = 1, \ldots, B$ in parallel

5: Compute p-value:

$$p = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\{\mathbb{T}^{(b)} \geq \mathbb{T}\}}{1 + B}.$$

6: **Return** p

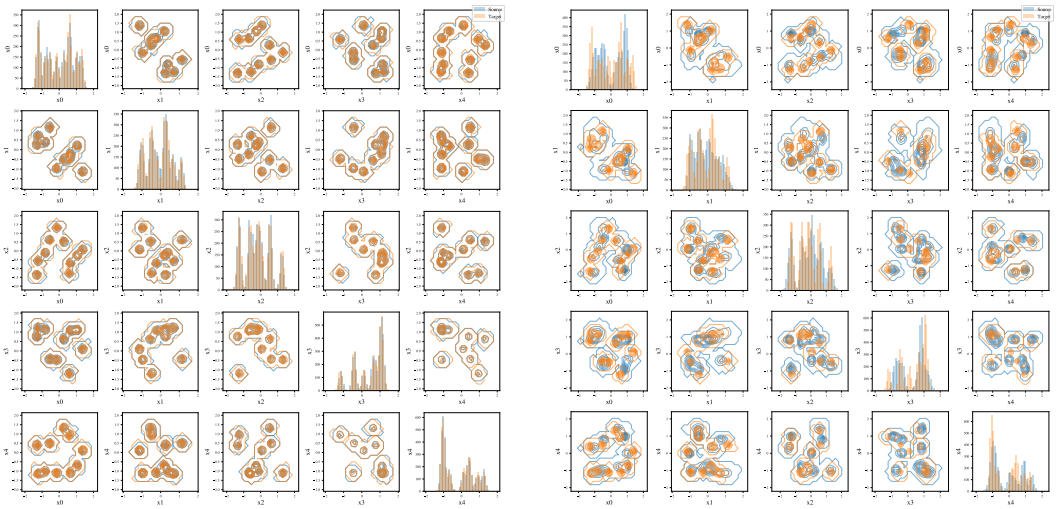


Figure 9: Histograms of the target and source variable X in the Gaussian mixture experiment. The orange histograms represent the target-domain data, while the blue histograms represent the source-domain data. The left panel compares the target domain with the first source domain, whose shift strength is $d^{(1)} = 0.01$, whereas the right panel compares the target domain with the third source domain, whose shift strength is $d^{(3)} = 0.2$. As the shift strength increases to $d^{(3)} = 0.2$, a clear distributional discrepancy between the target and source domains can be observed.

We train DA-Diff according to Algorithm 1 and report the validation losses of DA-Diff and the target-only diffusion model in Figure 10. We set the warm-up epoch and the total epoch number to 100 and 500, respectively. We dynamically update the weights of the weighted ERM and hyperparameters every 10 epochs by MM-QP proposed in Appendix C.

As shown in Figure 10, the validation-loss curve of DA-Diff exhibits a similar level of smoothness to that of the target-only diffusion model.

We next visualize the weights obtained by our optimization procedure. Figure 11 shows the solution $\boldsymbol{\pi}$ obtained by the MM-QP procedure in Appendix C, together with the corresponding source weights $w_1, \ldots, w_K$ and W_N recovered using (34). The hyperparameters ρ, C_{reg}, and $\boldsymbol{\lambda}$ selected by the gradient-alignment criterion in Appendix C.3 are shown in Figure 12. We use the candidate grids

$$\mathcal{G}_\rho = \{0.2, 0.4, 0.6, 0.8\}, \qquad \mathcal{G}_C = \{2^1, \ldots, 2^7\},$$

and set $M_\lambda = 5$ for the construction of $\mathcal{G}_\lambda$.

As shown in Figure 11, the trajectories of $\boldsymbol{\pi}$, w, and W_N exhibit a zig-zag pattern over epochs. This is expected, since the hyperparameters, $\boldsymbol{\pi}$, w, and W_N are dynamically updated every 10 epochs. Importantly, Figure 10 shows that these periodic updates do not noticeably increase the fluctuations of the validation loss, and the overall training process remains stable.

From the left column of Figure 11, we observe that the mixture distribution used in weighted ERM consistently assigns the largest proportion to the target domain. Moreover, the MM-QP procedure is able to reflect the degree of distribution shift between the source and target domains and assign the source-domain weights accordingly. The behavior of $\boldsymbol{\pi}$ across different time bins is also consistent with our motivation for partitioning the diffusion-time interval: when the diffusion time is close to t_0, the distributional discrepancies across domains are relatively large, whereas these discrepancies become weaker as the diffusion time approaches T. In contrast, Figure 14 shows that the weights obtained by UOWQ (Zhang et al., 2026a) for diffusion-model training exhibit little differentiation across source domains, suggesting that it has difficulty identifying which sources should contribute more and which should be downweighted. This behavior is also reflected in Table 1, where the performance of UOWQ deteriorates as the number of source domains increases, indicating increasingly severe negative transfer. We emphasize, however, that UOWQ was originally developed for regular parametric models rather than specifically for diffusion models, so its assumptions and weighting mechanism are not directly tailored to the score-matching setting considered here.

The source weights w shown in the middle column of Figure 11 exhibit the same pattern. The right column reports the total weighted source sample size W_N for each time bin. From the scale of the y-axis, we can see that relatively few source samples are effectively used in the early time bins, while substantially more source information is incorporated in the later time bins. This is consistent with the fact that the forward diffusion process progressively reduces the distributional discrepancies across domains.

Figure 12 shows that the hyperparameters selected by the gradient-alignment criterion vary throughout the training process. In particular, ρ switches among several candidate values as training progresses, while C_{reg} remains relatively small for most epochs but occasionally takes substantially larger values. The three components of $\boldsymbol{\lambda}$ also change across epochs, indicating that the relative importance assigned to the corresponding terms in the selection criterion is adjusted dynamically during training. These results suggest that using a single fixed set of hyperparameters throughout training may be overly restrictive, whereas the proposed gradient-alignment procedure can adapt the hyperparameters to the current state of the model. Such changes in the selected hyperparameters also help explain the piecewise-varying behavior of $\boldsymbol{\pi}$, w, and W_N observed in Figure 11.

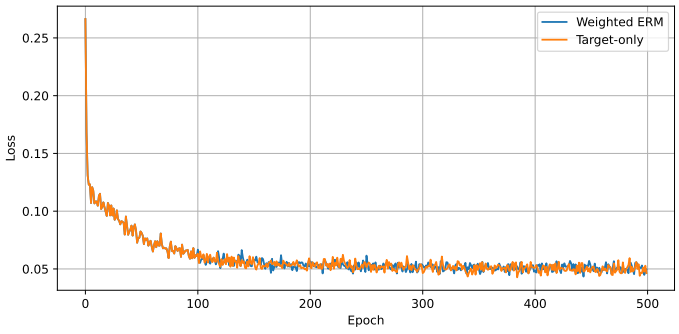


Figure 10: Validation loss curves of DA-Diff and the target-only diffusion model during training.

I Computational Efficiency

We additionally provide a detailed analysis of the wall-clock runtime of the competing CI testing methods and DA-CIT.

As shown in Figure 16, the computational cost of DA-CIT consists of five main components: source-model pretraining, MM-QP optimization for the domain weights, adaptation excluding MM-QP, CRT sampling, and statistic computation. Among these, the adaptation stage and MM-QP optimiza-

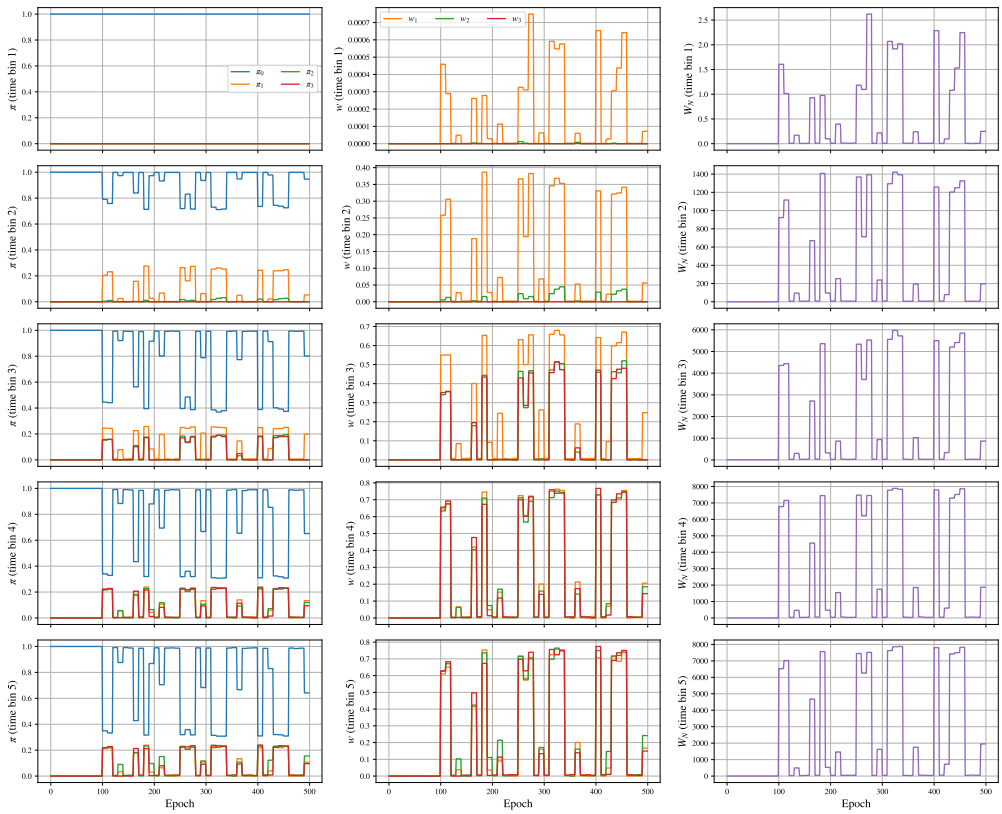


Figure 11: Evolution of the MM-QP solution $\boldsymbol{\pi}$ and the corresponding $w_1, \ldots, w_K$ and W_N recovered by (34) during the training of DA-Diff.

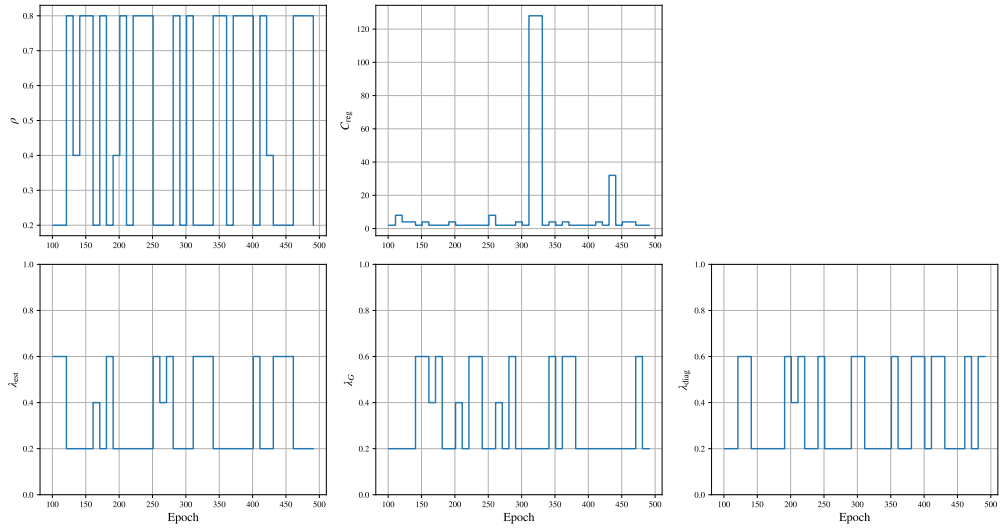


Figure 12: Hyperparameters ρ, C_{reg}, and $\boldsymbol{\lambda}$ selected by the gradient-alignment criterion during the training of DA-Diff.

tion account for most of the runtime, while CRT sampling and statistic computation contribute only a small fraction. In particular, the MM-QP cost increases when weight optimization is performed more frequently, when more time bins are used, or when a finer λ grid is adopted.

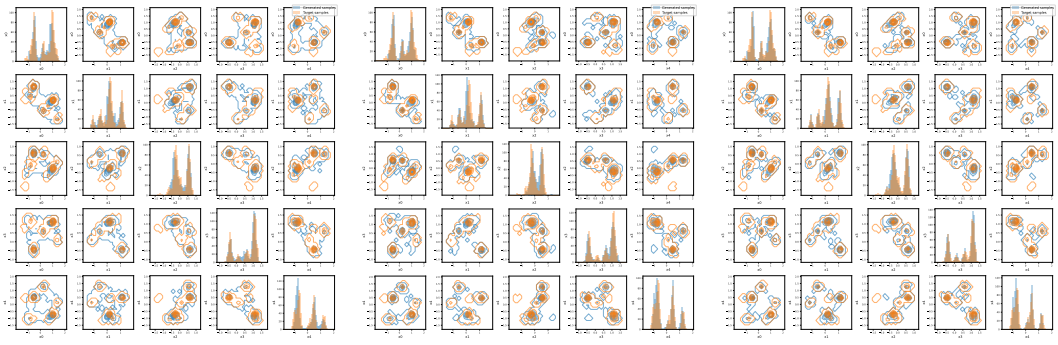


Figure 13: Generated samples in the Gaussian mixture experiment. From left to right, the methods are Target only, Cat-all, and DA-Diff. Their corresponding W_2 distances are 0.040, 0.047, and 0.023, respectively.



Figure 14: Source proportions $\alpha_1, \ldots, \alpha_K$, source weights $w_1, \ldots, w_K$, and the total weighted source sample size W_N solved by UOWQ (Zhang et al., 2026a).

Although DA-CIT achieves strong CI testing performance, this improvement comes at the cost of substantially higher computational time, as shown in Figure 15. Compared with CDCIT, the additional computational overhead of DA-CIT mainly arises from pretraining the source-specific score models and repeatedly solving the MM-QP problems during weighted-ERM training. Therefore, improving the computational efficiency of source pretraining and weight optimization is an important direction for future work.

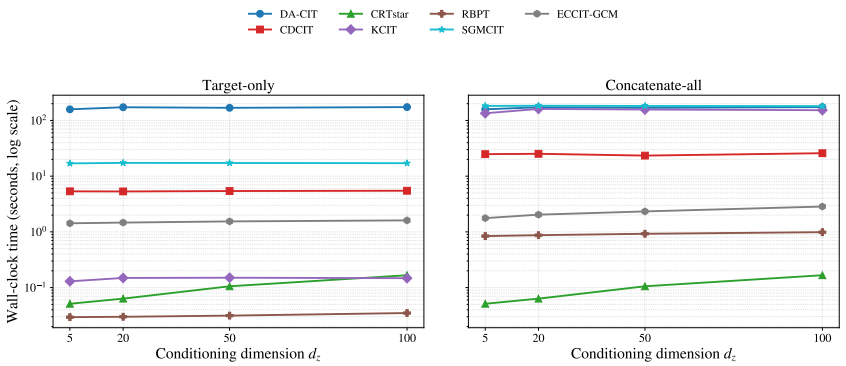


Figure 15: Wall-clock runtime of the CI testing methods considered in Figure 1 across different conditioning dimensions d_z. We report the runtime under both the Target-only and Concatenate-all settings. The vertical axis is shown on a logarithmic scale.

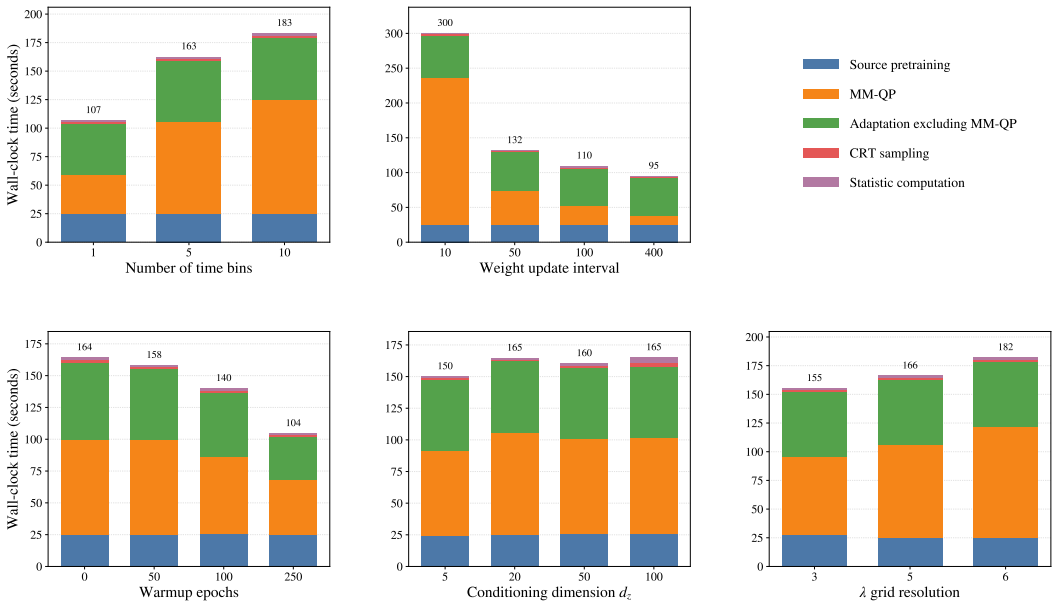


Figure 16: Detailed runtime decomposition of DA-CIT under the ablation settings considered in Figures 4 and 5. The total runtime is decomposed into source pretraining, MM-QP optimization, adaptation excluding MM-QP, CRT sampling, and statistic computation.

J Implementation Details

Unless otherwise specified, we use the following implementation for DA-Diff throughout the experiments. The score network is a fully connected neural network with hidden dimension 256 and three hidden layers.

For each source domain, we first pretrain a source-specific score network for 500 epochs with learning rate 10^{-3}. The domain-adaptation stage is then trained for another 500 epochs with learning rate 10^{-3}, with a warm-up period of 100 epochs. We use five batches per epoch.

To allow the transferability of source domains to vary across diffusion time, we divide the diffusion interval into 5 time bins and learn separate domain weights for each bin. The weights are updated every 10 epochs. The hyperparameters used for weight optimization are also re-selected every 100 epochs using the target validation set. We use 30% of the target training data for validation.

For hyperparameter selection, we fix

$$\rho \in \{0.2, 0.4, 0.6, 0.8\}, \qquad C_{\mathrm{reg}} \in \{2, 4, 8, 16, 32, 64, 128\},$$

and discretize the λ search space with a grid resolution of 5. The minimum target-domain mixture weight is set to $\epsilon_\pi = 10^{-6}$.

We solve the weight optimization problem using the MM-QP procedure described in Appendix C. For each optimization, we use both the pooling and target-only solutions as initializations and retain the solution with the smaller objective value. The maximum number of MM iterations is 50 with tolerance 10^{-6}. Each convex QP subproblem is solved with at most 2000 iterations and tolerance 10^{-9}.

All experiments were conducted on a machine equipped with an NVIDIA RTX A6000 GPU and an Intel(R) Xeon(R) Gold 6430 CPU with 128 cores.

K Proof of Theorems

Although the first n_0 observations in D_0 are designated as the samples used for the CRT in (7), the target-domain observations are i.i.d., so using the first n_0 samples or the last n_0 samples of D_0 for training is theoretically equivalent. Therefore, without loss of generality, throughout the theoretical analysis of the diffusion model, we take the first n_0 observations in D_0 as the target-domain training samples.

Before we begin the proof, we first define the Hölder function class and the ReLU neural network class $\mathcal{F}$ used throughout the analysis.

Hölder function class. Let $\beta > 0$ denote the degree of smoothness, and write:

$$\beta = m + \gamma, \qquad m = \lfloor \beta \rfloor, \qquad \gamma \in [0, 1).$$

For a multi-index $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_d) \in \mathbb{N}_0^d$, let:

$$|\boldsymbol{\nu}|_1 := \sum_{j=1}^{d} \nu_j, \qquad \partial^{\boldsymbol{\nu}} := \frac{\partial^{|\boldsymbol{\nu}|_1}}{\partial x_1^{\nu_1} \cdots \partial x_d^{\nu_d}}.$$

For a function $f : \mathbb{R}^d \to \mathbb{R}$, its β-Hölder norm is defined as:

$$\|f\|_{\mathcal{H}^\beta(\mathbb{R}^d)} := \max_{|\boldsymbol{\nu}|_1 \le m} \sup_{x \in \mathbb{R}^d} |\partial^{\boldsymbol{\nu}} f(x)| + \max_{|\boldsymbol{\nu}|_1 = m} \sup_{x \neq x'} \frac{|\partial^{\boldsymbol{\nu}} f(x) - \partial^{\boldsymbol{\nu}} f(x')|}{\|x - x'\|_\infty^\gamma}. \tag{79}$$

where the second term is omitted when $\gamma = 0$. A function f is said to be β-Hölder continuous if $\|f\|_{\mathcal{H}^\beta(\mathbb{R}^d)} < \infty$.

For any $B > 0$, define the Hölder ball of smoothness β and radius B as:

$$\mathcal{H}^\beta(\mathbb{R}^d, B) := \left\{ f : \mathbb{R}^d \to \mathbb{R} : \|f\|_{\mathcal{H}^\beta(\mathbb{R}^d)} \le B \right\}. \tag{80}$$

ReLU neural network class. We use neural networks to parameterize the score functions. Let:

$$\sigma(u) := \max\{u, 0\}$$

denote the ReLU activation function, applied componentwise. For the input:

$$u_0 := \left[x^\top \; z^\top \; t \right]^\top, \tag{81}$$

consider an L-layer feed-forward ReLU network defined recursively by:

$$\begin{aligned} u_\ell &= \sigma\left(A_\ell u_{\ell-1} + b_\ell\right), \qquad \ell = 1, \ldots, L-1, \\ s(x, z, t) &= A_L u_{L-1} + b_L. \end{aligned} \tag{82}$$

Here, $d_0 = d_x + d_z + 1, d_L = d_x$, and $A_\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}, b_\ell \in \mathbb{R}^{d_\ell}, \ell = 1, \ldots, L$.

For $M_t, W, \kappa, L, S > 0$, define the ReLU neural network class:

$$\mathcal{F}(M_t, W, \kappa, L, S) := \Bigg\{ s \text{ of the form (82)} : \begin{array}{l} \max\limits_{1\le \ell \le L-1} d_\ell \le W, \\ \sup\limits_{x,z} \|s(x,z,t)\|_\infty \le M_t, \\ \max\limits_{1\le \ell \le L} \left(\|A_\ell\|_\infty \vee \|b_\ell\|_\infty\right) \le \kappa, \\ \sum\limits_{\ell=1}^{L} \left(\|A_\ell\|_0 + \|b_\ell\|_0\right) \le S \end{array} \Bigg\}. \tag{83}$$

Here, W controls the maximum width of the network, L denotes the network depth, κ bounds the magnitude of the network parameters, and S controls the total number of nonzero parameters. Moreover, $\|\cdot\|_\infty$ denotes the entrywise maximum norm and $\|\cdot\|_0$ denotes the number of nonzero entries. The quantity M_t controls the magnitude of the network output and is allowed to depend on the diffusion time t.

K.1 Global population minimizer of the weighted score matching risk

We define the weighted score matching risk:

$$\mathcal{R}_w(s) := n_0\, \mathcal{R}_0(s) + \sum_{k=1}^{K} w_k n_k\, \mathcal{R}_k(s). \tag{84}$$

Theorem 4 (Global population minimizer of the weighted score matching risk (84))**.** *Recall the π_k defined in (33):*

$$\pi_k := \begin{cases} \frac{n_0}{n_0 + W_N}, & k = 0, \\ \frac{w_k n_k}{n_0 + W_N}, & k \in \{1, \dots, K\}. \end{cases} \tag{85}$$

For each $t \in [t_0, T]$, define the mixture density:

$$p_{w,t}(x,z) := \sum_{k=0}^{K} \pi_k\, p_{k,t}(x,z), \tag{86}$$

and the posterior domain probability:

$$\eta_k(x,z,t) := \frac{\pi_k\, p_{k,t}(x,z)}{p_{w,t}(x,z)}. \tag{87}$$

Then the (global) population minimizer over all measurable functions s:

$$s_w^o := \arg\min_s \mathcal{R}_w(s) = \arg\min_s \sum_{k=0}^{K} \pi_k \mathcal{R}_k(s) \tag{88}$$

satisfies, for all (x,z,t) with $p_{w,t}(x,z) > 0$,

$$s_w^o(x,z,t) = \sum_{k=0}^{K} \eta_k(x,z,t)\, s_k^*(x,z,t). \tag{89}$$

Proof. For notational convenience, we introduce the following inner products and associated norms on the relevant function spaces. For any measurable vector-valued functions $g_1(x,z,t)$ and $g_2(x,z,t)$, we define the inner product induced by the k-th domain as

$$\langle g_1, g_2 \rangle_{\mathcal{H}_k} := \frac{1}{T - t_0} \int_{t_0}^{T} \mathbb{E}_{(X_t,Z)\sim p_{k,t}(\cdot,\cdot)} \Big[\langle g_1(X_t,Z,t), g_2(X_t,Z,t) \rangle \Big]\, dt, \tag{90}$$

with the associated squared norm $\|g_1\|^2_{\mathcal{H}_k} := \langle g_1, g_1\rangle_{\mathcal{H}_k}$. Under this definition and (16), the population risk on the k-th domain can be equivalently written as $\mathcal{R}_k(s) = \|s - s_k^*\|^2_{\mathcal{H}_k}$. Similarly, define the inner product under the mixture density $p_{w,t}$ as

$$\langle g_1, g_2\rangle_{\mathcal{H}_w} := \frac{1}{T-t_0}\int_{t_0}^{T} \mathbb{E}_{(X_t,Z)\sim p_{w,t}(\cdot,\cdot)}\Big[\langle g_1(X_t,Z,t), g_2(X_t,Z,t)\rangle\Big]\,dt, \tag{91}$$

with the associated squared norm $\|g_1\|^2_{\mathcal{H}_w} := \langle g_1, g_1\rangle_{\mathcal{H}_w}$.

Define $\overline{\mathcal{R}}_w(s) := \sum_{k=0}^K \pi_k \mathcal{R}_k(s)$, we know:

$$\begin{aligned}
\overline{\mathcal{R}}_w(s) &= \frac{1}{T-t_0}\int_{t_0}^{T}\sum_{k=0}^{K}\pi_k \mathbb{E}_{(X_t,Z)\sim p_{k,t}(\cdot,\cdot)}\|s - s_k^*\|_2^2\,dt \\
&= \frac{1}{T-t_0}\int_{t_0}^{T}\int_{\mathcal{X}\times\mathcal{Z}}\sum_{k=0}^{K}\pi_k p_{k,t}\|s - s_k^*\|_2^2\,dxdzdt \\
&= \frac{1}{T-t_0}\int_{t_0}^{T}\int_{\{(x,z)\in\mathcal{X}\times\mathcal{Z}|p_{w,t}(x,z)>0\}} p_{w,t}\sum_{k=0}^{K}\frac{\pi_k p_{k,t}}{p_{w,t}}\|s - s_k^*\|_2^2\,dxdzdt \\
&= \frac{1}{T-t_0}\int_{t_0}^{T}\int_{\{(x,z)\in\mathcal{X}\times\mathcal{Z}|p_{w,t}(x,z)>0\}} p_{w,t}\sum_{k=0}^{K}\eta_k\|s - s_k^*\|_2^2\,dxdzdt \\
&= \frac{1}{T-t_0}\int_{t_0}^{T}\mathbb{E}_{(X_t,Z)\sim p_{w,t}(\cdot,\cdot)}\left[\sum_{k=0}^{K}\eta_k\|s - s_k^*\|_2^2\right]dt.
\end{aligned} \tag{92}$$

We next characterize the global minimizer of $\overline{\mathcal{R}}_w$ over the space of measurable functions. To derive its minimizer, let h be an arbitrary admissible perturbation and consider the perturbed function $s_\epsilon = s + \epsilon h$. Assuming that differentiation with respect to ϵ can be interchanged with the time integral and expectation, the first variation of $\overline{\mathcal{R}}_w(s)$ at s in the direction h is given by

$$\begin{aligned}
D\overline{\mathcal{R}}_w(s)[h] &:= \frac{d}{d\epsilon}\overline{\mathcal{R}}_w(s+\epsilon h)\Big|_{\epsilon=0} \\
&= \frac{2}{T-t_0}\int_{t_0}^{T}\mathbb{E}_{(X_t,Z)\sim p_{w,t}(\cdot,\cdot)}\left[\left\langle\sum_{k=0}^{K}\eta_k\big(s - s_k^*\big), h\right\rangle\right]dt.
\end{aligned} \tag{93}$$

Since h is arbitrary, the first-order optimality condition $D\overline{\mathcal{R}}_w(s)[h] = 0$ for every admissible h implies:

$$\sum_{k=0}^{K}\eta_k\big(s - s_k^*\big) = 0 \tag{94}$$

for almost every $t \in [t_0, T]$ and Lebesgue-almost every $(x_t, z) \in \{(x,z) \in \mathcal{X}\times\mathcal{Z}|p_{w,t}(x,z) > 0\}$. Using the truth $\sum_{k=0}^K \eta_k = 1$ ((86) and (87)), we obtain:

$$\boxed{s_w^o(x_t,z,t) = \sum_{k=0}^{K}\eta_k(x_t,z,t)s_k^*(x_t,z,t).}$$

Moreover, the second variation in the direction h is

$$\begin{aligned}
D^2\overline{\mathcal{R}}_w(s)[h,h] &= \frac{2}{T-t_0}\int_{t_0}^{T}\mathbb{E}_{(X_t,Z)\sim p_{w,t}}\left[\|h(x_t,z,t)\|_2^2\right]\,dt \\
&= 2\|h\|^2_{\mathcal{H}_w} \geq 0.
\end{aligned} \tag{95}$$

Therefore, $\overline{\mathcal{R}}_w(s)$ is strictly convex in s, and s_w^o is its unique global minimizer almost everywhere.

□

K.2 EXCESS-RISK DECOMPOSITION

Knowing the global minimizer s_w^o, we can further decompose $\overline{\mathcal{R}}_w(s)$ based on (92) as follows

$$\begin{aligned}\overline{\mathcal{R}}_w(s) &= \frac{1}{T-t_0}\int_{t_0}^{T}\mathbb{E}_{(X_t,Z)\sim p_{w,t}(\cdot,\cdot)}\left[\sum_{k=0}^{K}\eta_k\|s-s_k^*\|_2^2\right]dt\\ &= \frac{1}{T-t_0}\int_{t_0}^{T}\mathbb{E}_{(X_t,Z)\sim p_{w,t}(\cdot,\cdot)}\left[\sum_{k=0}^{K}\eta_k\|s-s_w^o+s_w^o-s_k^*\|_2^2\right]dt\\ &= \frac{1}{T-t_0}\int_{t_0}^{T}\mathbb{E}_{(X_t,Z)\sim p_{w,t}(\cdot,\cdot)}\Bigg[\sum_{k=0}^{K}\eta_k\|s-s_w^o\|_2^2+\sum_{k=0}^{K}\eta_k\|s_w^o-s_k^*\|_2^2\\ &\qquad\qquad +2\sum_{k=0}^{K}\eta_k\left\langle s-s_w^o,s_w^o-s_k^*\right\rangle\Bigg]dt. \end{aligned}\tag{96}$$

By the truth that $\sum_{k=0}^{K}\eta_k=1$ and (89), we have $\sum_{k=0}^{K}\eta_k\|s-s_w^o\|_2^2=\|s-s_w^o\|_2^2$ and $\sum_{k=0}^{K}\eta_k\left\langle s-s_w^o,s_w^o-s_k^*\right\rangle=\left\langle s-s_w^o,\sum_{k=0}^{K}\eta_k(s_w^o-s_k^*)\right\rangle=\left\langle s-s_w^o,s_w^o-\sum_{k=0}^{K}\eta_k s_k^*\right\rangle=0$. Therefore, $\overline{\mathcal{R}}_w(s)$ can be rewritten as follows

$$\begin{aligned}\overline{\mathcal{R}}_w(s) &= \frac{1}{T-t_0}\int_{t_0}^{T}\mathbb{E}_{(X_t,Z)\sim p_{w,t}(\cdot,\cdot)}\|s-s_w^o\|_2^2\,dt\\ &\quad+\frac{1}{T-t_0}\int_{t_0}^{T}\mathbb{E}_{(X_t,Z)\sim p_{w,t}(\cdot,\cdot)}\left[\sum_{k=0}^{K}\eta_k\|s_w^o-s_k^*\|_2^2\right]dt,\end{aligned}\tag{97}$$

where the first term depends on s, while the second term is irrelevant to s. Moreover, it holds that

$$\overline{\mathcal{R}}_w(s_w^o)=\frac{1}{T-t_0}\int_{t_0}^{T}\mathbb{E}_{(X_t,Z)\sim p_{w,t}(\cdot,\cdot)}\left[\sum_{k=0}^{K}\eta_k\|s_w^o-s_k^*\|_2^2\right]dt,\tag{98}$$

and the excessive score matching risk is defined as

$$\overline{\mathcal{E}}_w(s):=\overline{\mathcal{R}}_w(s)-\overline{\mathcal{R}}_w(s_w^o)=\frac{1}{T-t_0}\int_{t_0}^{T}\mathbb{E}_{(X_t,Z)\sim p_{w,t}(\cdot,\cdot)}\|s-s_w^o\|_2^2\,dt=\|s-s_w^o\|_{\mathcal{H}_w}^2.\tag{99}$$

With the definitions in (90) and (91), we can bound the score matching risk $\mathbb{E}_{D_0,\ldots,D_K}\left[\mathcal{R}_0(\widehat{s}_w)\right]$ in Theorem 1 as follows

$$\boxed{\begin{aligned}\mathbb{E}_{D_0,\ldots,D_K}\left[\mathcal{R}_0(\widehat{s}_w)\right] &= \mathbb{E}_{D_0,\ldots,D_K}\left[\|\widehat{s}_w-s_0^*\|_{\mathcal{H}_0}^2\right]\\ &\lesssim \mathbb{E}_{D_0,\ldots,D_K}\left[\|\widehat{s}_w-s_w^o\|_{\mathcal{H}_0}^2\right]+\|s_w^o-s_0^*\|_{\mathcal{H}_0}^2,\end{aligned}}\tag{100}$$

where the first term is the estimation error under the density $p_{0,t}$, and the second term is the transfer bias between the optimal minimizer s_w^o of $\overline{\mathcal{R}}_w(s)$ and the true score in the target domain.

K.3 PROPERTIES OF $p_{w,t}$

In this section, we unveil some properties of the mixture density $p_{w,t}$, including smoothness, its conditional density given z, and its score function.

Lemma 1 (Conditional-density structure and score representation of the mixture distribution)**.** *Let*

$$P_w(x,z):=\sum_{k=0}^{K}\pi_k P_k(x,z),\qquad \pi_k\geq 0,\qquad \sum_{k=0}^{K}\pi_k=1.$$

and its joint density is $p_{w,0}(x,z)$ defined in (86) when the diffusion time $t=0$. Under Assumption 2, the following statements hold.

Define the marginal density of Z under P_w as

$$p_w(z) := \sum_{k=0}^{K} \pi_k p_k(z), \tag{101}$$

and, for $p_w(z) > 0$, define

$$\overline{\eta}_k(z) := \frac{\pi_k p_k(z)}{p_w(z)}, \qquad k = 0, \dots, K. \tag{102}$$

Then,

$$\overline{\eta}_k(z) \geq 0, \qquad \sum_{k=0}^{K} \overline{\eta}_k(z) = 1. \tag{103}$$

The conditional density of X given $Z = z$ under P_w satisfies

$$p_{w,0}(x|z) = \sum_{k=0}^{K} \overline{\eta}_k(z) p_k(x|z) = \exp\left(-C_x \|x\|_2^2/2\right) \widetilde{f}_w(x, z), \tag{104}$$

where

$$\widetilde{f}_w(x, z) := \sum_{k=0}^{K} \overline{\eta}_k(z) f_k(x, z) \tag{105}$$

satisfies

$$\widetilde{f}_w \in \mathcal{H}^{\beta}\left(\mathbb{R}^{d_x} \times \mathbb{R}^{d_z}, B_{\mathrm{mix}}\right), \quad \sup_{x,z} |f_k(x, z)| \leq B \text{ for all } k, \quad \text{and} \quad \widetilde{f}_w(x, z) \geq C_f, \tag{106}$$

where $B_{\mathrm{mix}} > 0$ depends only on β, d_x, d_z, B, L_q, and C_{DR}. Moreover, the marginal density $p_w(z)$ has a sub-Gaussian tail:

$$p_w(z) \lesssim \exp\left(-C_z \|z\|_2^2/2\right). \tag{107}$$

For any $t \in [t_0, T]$, the conditional density under the forward diffusion satisfies

$$p_{w,t}(x|z) = \sum_{k=0}^{K} \overline{\eta}_k(z) p_{k,t}(x|z). \tag{108}$$

Furthermore, the population minimizer s_w^o of the weighted score-matching risk satisfies

$$s_w^o(x, z, t) = \nabla_x \log p_{w,t}(x|z) \tag{109}$$

for almost every (x, z, t).

Proof. By definition,

$$p_{w,0}(x, z) = \sum_{k=0}^{K} \pi_k p_k(x, z) = \sum_{k=0}^{K} \pi_k p_k(z) p_k(x|z). \tag{110}$$

Integrating with respect to x gives

$$p_w(z) = \sum_{k=0}^{K} \pi_k p_k(z). \tag{111}$$

Therefore, for every z such that $p_w(z) > 0$,

$$\begin{aligned} p_{w,0}(x|z) &= \frac{p_{w,0}(x, z)}{p_w(z)} \\ &= \sum_{k=0}^{K} \frac{\pi_k p_k(z)}{p_w(z)} p_k(x|z) \\ &= \sum_{k=0}^{K} \overline{\eta}_k(z) p_k(x|z). \end{aligned} \tag{112}$$

By Assumption 2,

$$p_k(x|z) = \exp\left(-C_x\|x\|_2^2/2\right) f_k(x,z).$$

Hence,

$$\begin{aligned} p_{w,0}(x|z) &= \exp\left(-C_x\|x\|_2^2/2\right) \sum_{k=0}^{K} \overline{\eta}_k(z) f_k(x,z) \\ &= \exp\left(-C_x\|x\|_2^2/2\right) \widetilde{f}_w(x,z), \end{aligned} \tag{113}$$

which proves (104). A direct result of assumption (20) and definition in (79) is that $\sup_{x,z} |f_k(x,z)| \le B$ for all k. Since $\overline{\eta}_k(z) \ge 0$, $\sum_{k=0}^{K} \overline{\eta}_k(z) = 1$, and $C_f \le f_k(x,z) \le B$, we have

$$C_f \le \widetilde{f}_w(x,z) \le B, \tag{114}$$

which proves the second and third statements of (106).

We next establish the Hölder regularity of $\widetilde{f}_w(x,z)$ in (106). Recall the definition of $q_k, k = 0,\dots,K$ in (23). We define

$$q_w(z) := \sum_{k=0}^{K} \pi_k q_k(z). \tag{115}$$

Then, for Lebesgue-almost every z,

$$\begin{aligned} p_w(z) &= \sum_{k=0}^{K} \pi_k p_k(z) \\ &= p_0(z) \sum_{k=0}^{K} \pi_k q_k(z) \\ &= p_0(z) q_w(z). \end{aligned} \tag{116}$$

In particular, this identity remains meaningful even at points where $p_0(z) = 0$, since no density ratio is formed there. Further, define

$$F_w(x,z) := \sum_{k=0}^{K} \pi_k q_k(z) f_k(x,z). \tag{117}$$

Whenever $p_w(z) > 0$, we necessarily have $p_0(z) > 0$ by (116), and hence

$$\begin{aligned} \widetilde{f}_w(x,z) &= \sum_{k=0}^{K} \frac{\pi_k p_k(z)}{p_w(z)} f_k(x,z) \\ &= \sum_{k=0}^{K} \frac{\pi_k q_k(z)}{q_w(z)} f_k(x,z) \\ &= \frac{F_w(x,z)}{q_w(z)}. \end{aligned} \tag{118}$$

We first control the Hölder regularity of q_w. By the subadditivity of the Hölder norm, (23) and $\sum_{k=0}^{K} \pi_k = 1$,

$$\begin{aligned} \|q_w\|_{\mathcal{H}^\beta(\mathbb{R}^{d_z})} &\le \sum_{k=0}^{K} \pi_k \|q_k\|_{\mathcal{H}^\beta(\mathbb{R}^{d_z})} \\ &\le L_q. \end{aligned} \tag{119}$$

We next derive a lower bound for q_w on the support of the target marginal distribution. By the definitions in (101) and $p_{w\setminus 0}(z)$ in Assumption 1, we have

$$p_w(z) = \pi_0 p_0(z) + (1-\pi_0) p_{w\setminus 0}(z). \tag{120}$$

By the marginal overlap condition (18),

$$p_{w\setminus 0}(z) \geq C_{\mathrm{DR}}^{-1} p_0(z) \tag{121}$$

for p_0-almost every z. Consequently,

$$\begin{aligned} p_w(z) &\geq \left[\pi_0 + \frac{1-\pi_0}{C_{\mathrm{DR}}}\right] p_0(z) \\ &\geq C_{\mathrm{DR}}^{-1} p_0(z). \end{aligned} \tag{122}$$

Combining (116) and (122) yields

$$q_w(z) \geq C_{\mathrm{DR}}^{-1}, \tag{123}$$

for p_0-almost every z.

We next control the Hölder regularity of F_w. Regarding $q_k(z)$ as a function of (x, z) that is constant in x, the standard product inequality for Hölder spaces gives

$$\|gh\|_{\mathcal{H}^\beta(\mathbb{R}^{d_x+d_z})} \leq C_{\beta,d_x,d_z} \|g\|_{\mathcal{H}^\beta(\mathbb{R}^{d_z})} \|h\|_{\mathcal{H}^\beta(\mathbb{R}^{d_x+d_z})}, \tag{124}$$

where $C_{\beta,d_x,d_z} > 0$ is a constant. Therefore, by the definition of F_w in (117),

$$\begin{aligned} \|F_w\|_{\mathcal{H}^\beta(\mathbb{R}^{d_x+d_z})} &\leq \sum_{k=0}^{K} \pi_k \|q_k f_k\|_{\mathcal{H}^\beta(\mathbb{R}^{d_x+d_z})} \\ &\leq C_{\beta,d_x,d_z} \sum_{k=0}^{K} \pi_k \|q_k\|_{\mathcal{H}^\beta(\mathbb{R}^{d_z})} \|f_k\|_{\mathcal{H}^\beta(\mathbb{R}^{d_x+d_z})} \\ &\leq C_{\beta,d_x,d_z} L_q B. \end{aligned} \tag{125}$$

The representation in (118) is only required on the support of p_w. At points where $p_w(z) = 0$, the conditional distribution is not uniquely determined, and therefore it can be modified without changing the joint distribution. Let $\chi : \mathbb{R} \to [0, 1]$ be a fixed infinitely differentiable function satisfying

$$\chi(u) = 0 \quad \text{for } u \leq C_{\mathrm{DR}}^{-1}/2, \qquad \chi(u) = 1 \quad \text{for } u \geq C_{\mathrm{DR}}^{-1}. \tag{126}$$

Define

$$h(u) := \begin{cases} \chi(u)/u, & u > 0, \\ 0, & u \leq 0. \end{cases} \tag{127}$$

Since χ vanishes on a neighborhood of zero, h is infinitely differentiable with uniformly bounded derivatives of every finite order. We then define, for every $(x, z) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_z}$,

$$\widetilde{f}_w^{\mathrm{ext}}(x, z) := h(q_w(z)) F_w(x, z) + \big[1 - \chi(q_w(z))\big] f_0(x, z). \tag{128}$$

For p_w-almost every z, we have $p_0(z) > 0$ and, by (123), $q_w(z) \geq C_{\mathrm{DR}}^{-1}$. Hence

$$\chi(q_w(z)) = 1, \qquad h(q_w(z)) = \frac{1}{q_w(z)}.$$

Therefore,

$$\begin{aligned} \widetilde{f}_w^{\mathrm{ext}}(x, z) &= \frac{F_w(x, z)}{q_w(z)} \\ &= \widetilde{f}_w(x, z) \end{aligned} \tag{129}$$

for p_w-almost every z. Thus $\widetilde{f}_w^{\mathrm{ext}}$ defines a valid version of the conditional-density factor. For simplicity, we henceforth denote this version again by $\widetilde{f}_w$.

We next verify that the uniform lower and upper bounds are preserved. When $q_w(z) > 0$, (128) can be rewritten as

$$\begin{aligned} \widetilde{f}_w^{\mathrm{ext}}(x, z) = \sum_{k=0}^{K} \chi(q_w(z)) \frac{\pi_k q_k(z)}{q_w(z)} f_k(x, z) \\ + \big[1 - \chi(q_w(z))\big] f_0(x, z). \end{aligned} \tag{130}$$

All coefficients in (130) are nonnegative and their sum equals

$$\chi(q_w(z))\sum_{k=0}^{K}\frac{\pi_k q_k(z)}{q_w(z)} + 1 - \chi(q_w(z)) = 1. \tag{131}$$

Hence $\widetilde{f}_w^{\mathrm{ext}}$ is a convex combination of $f_0, \dots, f_K$. When $q_w(z) = 0$, (128) reduces to

$$\widetilde{f}_w^{\mathrm{ext}}(x, z) = f_0(x, z).$$

Since

$$C_f \le f_k(x, z) \le B, \qquad k = 0, \dots, K,$$

we conclude that

$$C_f \le \widetilde{f}_w^{\mathrm{ext}}(x, z) \le B \tag{132}$$

for all (x, z).

It remains to establish the joint Hölder regularity. By (119), $q_w \in \mathcal{H}^\beta(\mathbb{R}^{d_z}, L_q)$. Since χ and h are fixed smooth functions with bounded derivatives up to every finite order, the standard composition inequality for Hölder spaces implies

$$\|\chi \circ q_w\|_{\mathcal{H}^\beta(\mathbb{R}^{d_z})} \le C_\chi, \tag{133}$$

$$\|h \circ q_w\|_{\mathcal{H}^\beta(\mathbb{R}^{d_z})} \le C_h, \tag{134}$$

where C_χ and C_h depend only on β, d_z, L_q, and C_{DR}. Combining (125), (133), (134), and the Hölder product inequality yields

$$\begin{aligned}\|\widetilde{f}_w^{\mathrm{ext}}\|_{\mathcal{H}^\beta(\mathbb{R}^{d_x+d_z})} &\le C_{\beta,d_x,d_z}\Big(\|h \circ q_w\|_{\mathcal{H}^\beta}\|F_w\|_{\mathcal{H}^\beta} + \|1 - \chi \circ q_w\|_{\mathcal{H}^\beta}\|f_0\|_{\mathcal{H}^\beta}\Big) \\ &\le B_{\mathrm{mix}},\end{aligned} \tag{135}$$

where $B_{\mathrm{mix}} = C\left(\beta, d_x, d_z, B, L_q, C_{\mathrm{DR}}\right) < \infty$. Importantly, B_{mix} does not depend on K, the sample sizes, or the particular mixture weights. Therefore,

$$\widetilde{f}_w^{\mathrm{ext}} \in \mathcal{H}^\beta\left(\mathbb{R}^{d_x} \times \mathbb{R}^{d_z}, B_{\mathrm{mix}}\right). \tag{136}$$

Since $\widetilde{f}_w^{\mathrm{ext}}$ coincides with the original conditional-density factor p_w-almost everywhere, the first statement of (106) is proven.

The sub-Gaussian tail of $p_w(z)$ follows directly from Assumption 2:

$$\begin{aligned}p_w(z) &= \sum_{k=0}^{K}\pi_k p_k(z) \\ &\lesssim \sum_{k=0}^{K}\pi_k \exp\left(-C_z\|z\|_2^2/2\right) \\ &\lesssim \exp\left(-C_z\|z\|_2^2/2\right),\end{aligned} \tag{137}$$

which proves (107). We next consider the forward diffusion. Since the diffusion process is applied only to X and leaves Z unchanged,

$$p_{k,t}(x, z) = p_k(z)p_{k,t}(x|z), \qquad k = 0, \dots, K. \tag{138}$$

Consequently,

$$\begin{aligned}p_{w,t}(x, z) &= \sum_{k=0}^{K}\pi_k p_{k,t}(x, z) \\ &= \sum_{k=0}^{K}\pi_k p_k(z)p_{k,t}(x|z).\end{aligned} \tag{139}$$

Since the marginal distribution of Z is unchanged by the forward diffusion,

$$\int_{\mathbb{R}^{d_x}} p_{w,t}(x, z)\,dx = p_w(z).$$

Therefore,

$$\begin{aligned} p_{w,t}(x|z) &= \frac{p_{w,t}(x,z)}{p_w(z)} \\ &= \sum_{k=0}^{K} \frac{\pi_k p_k(z)}{p_w(z)} p_{k,t}(x|z) \\ &= \sum_{k=0}^{K} \overline{\eta}_k(z) p_{k,t}(x|z), \end{aligned} \tag{140}$$

which proves (108). Recall that

$$s_k^*(x,z,t) := \nabla_x \log p_{k,t}(x|z). \tag{141}$$

Since $\overline{\eta}_k(z)$ does not depend on x, differentiating (108) with respect to x gives

$$\begin{aligned} \nabla_x p_{w,t}(x|z) &= \sum_{k=0}^{K} \overline{\eta}_k(z) \nabla_x p_{k,t}(x|z) \\ &= \sum_{k=0}^{K} \overline{\eta}_k(z) p_{k,t}(x|z) s_k^*(x,z,t). \end{aligned} \tag{142}$$

Therefore,

$$\begin{aligned} \nabla_x \log p_{w,t}(x|z) &= \frac{\nabla_x p_{w,t}(x|z)}{p_{w,t}(x|z)} \\ &= \frac{\sum_{k=0}^{K} \overline{\eta}_k(z) \nabla_x p_{k,t}(x|z)}{p_{w,t}(x|z)} \\ &= \frac{\sum_{k=0}^{K} \overline{\eta}_k(z) p_{k,t}(x|z) s_k^*(x,z,t)}{p_{w,t}(x|z)} \\ &= \sum_{k=0}^{K} \frac{\overline{\eta}_k(z) p_{k,t}(x|z)}{p_{w,t}(x|z)} s_k^*(x,z,t) \\ &= \sum_{k=0}^{K} \eta_{k,w,t}(x,z) s_k^*(x,z,t), \end{aligned} \tag{143}$$

where

$$\begin{aligned} \eta_{k,w,t}(x,z) &:= \frac{\overline{\eta}_k(z) p_{k,t}(x|z)}{p_{w,t}(x|z)} \\ &= \frac{\pi_k p_k(z) p_{k,t}(x|z)}{p_w(z) p_{w,t}(x|z)} \\ &= \frac{\pi_k p_{k,t}(x,z)}{p_{w,t}(x,z)} \\ &= \eta_k(x,z,t). \end{aligned} \tag{144}$$

It follows from (89) that

$$\nabla_x \log p_{w,t}(x|z) = \sum_{k=0}^{K} \eta_k(x,z,t) s_k^*(x,z,t) = s_w^o(x,z,t), \tag{145}$$

which proves (109).

□

K.4 BOUNDING THE TRANSFER BIAS $\|s_w^o - s_0^*\|_{\mathcal{H}_0}^2$

Theorem 5 (The bound of transfer bias)**.** *Under Assumption 1 and 2, the transfer bias $\|s_w^o - s_0^*\|_{\mathcal{H}_0}^2$ can be bounded as*

$$\|s_w^o - s_0^*\|_{\mathcal{H}_0}^2 \lesssim \frac{W_N^2}{(n_0 + W_N)^2} \left[\overline{w}^\top G \overline{w} + \mathrm{diag}(G)^\top \overline{w}\right]. \tag{146}$$

Proof. Recall that in Assumption 1 and Theorem 1, we defined

$$W_N := \sum_{k=1}^K w_k n_k, \qquad \overline{w}_k := \frac{w_k n_k}{W_N}, \qquad \overline{w} := (\overline{w}_1, \ldots, \overline{w}_K)^\top,$$

so that

$$\overline{w}_k \geq 0, \qquad \sum_{k=1}^K \overline{w}_k = 1.$$

Define

$$\Delta_{k,t}(x_t, z) := s_k^*(x_t, z, t) - s_0^*(x_t, z, t), \qquad k = 1, \ldots, K. \tag{147}$$

According to the expression of the population-optimal weighted score s_w^o in (89), we have

$$s_w^o = \frac{n_0 p_{0,t} s_0^* + \sum_{k=1}^K w_k n_k p_{k,t} s_k^*}{n_0 p_{0,t} + \sum_{k=1}^K w_k n_k p_{k,t}}. \tag{148}$$

For $p_{0,t}(x_t, z) > 0$, define the joint density ratio

$$R_{k,t}(x_t, z) := \frac{p_{k,t}(x_t, z)}{p_{0,t}(x_t, z)} = \frac{p_k(z)}{p_0(z)} \frac{p_{k,t}(x_t|z)}{p_{0,t}(x_t|z)}. \tag{149}$$

Then,

$$s_w^o - s_0^* = \frac{W_N \sum_{k=1}^K \overline{w}_k R_{k,t} \Delta_{k,t}}{n_0 + W_N \sum_{k=1}^K \overline{w}_k R_{k,t}}. \tag{150}$$

We first derive a uniform lower bound for the denominator. By Assumption 2 and Lemma 1,

$$p_k(x|z) = \exp\left(-C_x \|x\|_2^2/2\right) f_k(x, z),$$

where $C_f \leq f_k(x, z) \leq B$. Therefore,

$$\frac{C_f}{B} p_0(x|z) \leq p_k(x|z) \leq \frac{B}{C_f} p_0(x|z). \tag{151}$$

Since the forward diffusion is applied only to X, integrating (151) against the nonnegative transition kernel preserves the inequalities. Hence, for every $t \in [t_0, T]$,

$$\frac{C_f}{B} p_{0,t}(x_t|z) \leq p_{k,t}(x_t|z) \leq \frac{B}{C_f} p_{0,t}(x_t|z), \text{ i.e.}$$

$$\frac{C_f}{B} \leq \frac{p_{k,t}(x_t|z)}{p_{0,t}(x_t|z)} \leq \frac{B}{C_f}. \tag{152}$$

Using (149) and (152), we obtain

$$\begin{aligned}
\sum_{k=1}^K \overline{w}_k R_{k,t}(x_t, z) &= \sum_{k=1}^K \overline{w}_k \frac{p_k(z)}{p_0(z)} \frac{p_{k,t}(x_t|z)}{p_{0,t}(x_t|z)} \\
&\geq \frac{C_f}{B} \sum_{k=1}^K \overline{w}_k \frac{p_k(z)}{p_0(z)} \\
&= \frac{C_f}{B} \frac{p_{w\setminus 0}(z)}{p_0(z)} \\
&\geq \frac{C_f}{B C_{\mathrm{DR}}},
\end{aligned} \tag{153}$$

where the last inequality follows from Assumption 1. Consequently,

$$n_0 + W_N \sum_{k=1}^{K} \overline{w}_k R_{k,t} \geq n_0 + \frac{C_f}{BC_{\mathrm{DR}}} W_N \gtrsim n_0 + W_N, \tag{154}$$

where the hidden constant depends only on C_{DR}, C_f, and B. It follows from (150) that

$$\|s_w^o - s_0^*\|_{\mathcal{H}_0}^2 \lesssim \frac{W_N^2}{(n_0 + W_N)^2} \left\| \sum_{k=1}^{K} \overline{w}_k R_{k,t} \Delta_{k,t} \right\|_{\mathcal{H}_0}^2. \tag{155}$$

We next bound the term involving the joint density ratios. By (18), (149), and (152),

$$0 \leq R_{k,t}(x_t, z) \leq \frac{B}{C_f C_{\mathrm{dr}}}, \tag{156}$$

uniformly over k, x_t, z, and t. Since all constants on the right-hand side are independent of the sample sizes and the weights, and $B/C_f C_{\mathrm{dr}}$ is larger than 1, we have

$$|R_{k,t}(x_t, z) - 1| \lesssim 1. \tag{157}$$

Now decompose

$$\sum_{k=1}^{K} \overline{w}_k R_{k,t} \Delta_{k,t} = \sum_{k=1}^{K} \overline{w}_k \Delta_{k,t} + \sum_{k=1}^{K} \overline{w}_k (R_{k,t} - 1) \Delta_{k,t}. \tag{158}$$

Using $\|a + b\|_2^2 \leq 2\|a\|_2^2 + 2\|b\|_2^2$, we obtain

$$\left\| \sum_{k=1}^{K} \overline{w}_k R_{k,t} \Delta_{k,t} \right\|_{\mathcal{H}_0}^2 \lesssim \left\| \sum_{k=1}^{K} \overline{w}_k \Delta_{k,t} \right\|_{\mathcal{H}_0}^2 + \left\| \sum_{k=1}^{K} \overline{w}_k (R_{k,t} - 1) \Delta_{k,t} \right\|_{\mathcal{H}_0}^2. \tag{159}$$

By the definition of the source discrepancy matrix G in (25),

$$\left\| \sum_{k=1}^{K} \overline{w}_k \Delta_{k,t} \right\|_{\mathcal{H}_0}^2 = \overline{w}^\top G \overline{w}. \tag{160}$$

For the second term, since $\overline{w}_k \geq 0$ and $\sum_{k=1}^{K} \overline{w}_k = 1$, Jensen's inequality gives, pointwise,

$$\left\| \sum_{k=1}^{K} \overline{w}_k (R_{k,t} - 1) \Delta_{k,t} \right\|_2^2 \leq \sum_{k=1}^{K} \overline{w}_k (R_{k,t} - 1)^2 \|\Delta_{k,t}\|_2^2 \lesssim \sum_{k=1}^{K} \overline{w}_k \|\Delta_{k,t}\|_2^2, \tag{161}$$

where the last inequality follows from (157). Taking expectation under $p_{0,t}$ and integrating over $t \in [t_0, T]$ yield

$$\left\| \sum_{k=1}^{K} \overline{w}_k (R_{k,t} - 1) \Delta_{k,t} \right\|_{\mathcal{H}_0}^2 \lesssim \sum_{k=1}^{K} \overline{w}_k \|\Delta_{k,t}\|_{\mathcal{H}_0}^2 = \sum_{k=1}^{K} \overline{w}_k G_{kk} = \mathrm{diag}(G)^\top \overline{w}. \tag{162}$$

Combining (159), (160), and (162), we obtain

$$\left\| \sum_{k=1}^{K} \overline{w}_k R_{k,t} \Delta_{k,t} \right\|_{\mathcal{H}_0}^2 \lesssim \overline{w}^\top G \overline{w} + \mathrm{diag}(G)^\top \overline{w}. \tag{163}$$

Substituting (163) into (155) gives

$$\boxed{\|s_w^o - s_0^*\|_{\mathcal{H}_0}^2 \lesssim \frac{W_N^2}{(n_0+W_N)^2}\left[\overline{w}^\top G \overline{w} + \mathrm{diag}(G)^\top \overline{w}\right].} \tag{164}$$

Here, the hidden constant depends only on $C_{\mathrm{dr}}, C_{\mathrm{DR}}, C_f$, and B, and is independent of n_0, W_N, and $\overline{w}$.

□

K.5 Bounding the estimation error $\mathbb{E}_{D_0,\ldots,D_K}\left[\|\widehat{s}_w - s_w^o\|_{\mathcal{H}_0}^2\right]$

K.5.1 Turn $\mathcal{H}_0$ to $\mathcal{H}_w$

In this section, we relate $\mathbb{E}_{D_0,\ldots,D_K}\left[\|\widehat{s}_w - s_w^o\|_{\mathcal{H}_0}^2\right]$ to the corresponding error measured under $\mathcal{H}_w$.

Lemma 2 (Change of measure from $\mathcal{H}_0$ to $\mathcal{H}_w$)**.** *Under Assumption 1 and 2, we have*

$$\mathbb{E}_{D_0,\ldots,D_K}\left[\|\widehat{s}_w - s_w^o\|_{\mathcal{H}_0}^2\right] \lesssim \mathbb{E}_{D_0,\ldots,D_K}\left[\|\widehat{s}_w - s_w^o\|_{\mathcal{H}_w}^2\right]. \tag{165}$$

Proof. First, Assumption 2 implies a uniform overlap between the conditional distributions across domains. In particular, since

$$p_k(x|z) = \exp\left(-C_x\|x\|_2^2/2\right) f_k(x,z)$$

and $C_f \le f_k(x,z) \le B$, we have

$$p_k(x|z) \ge \frac{C_f}{B} p_0(x|z), \qquad k = 1,\ldots,K. \tag{166}$$

Since the forward diffusion is applied only to X, integrating (166) against the transition kernel preserves the inequality. Hence, for every $t \in [t_0, T]$,

$$p_{k,t}(x_t|z) \ge \frac{C_f}{B} p_{0,t}(x_t|z). \tag{167}$$

Recall that in Assumption 1, we defined

$$p_{w\setminus 0}(z) = \sum_{k=1}^K \frac{w_k n_k}{W_N} p_k(z).$$

Combining (167) with Assumption 1, we obtain

$$\begin{aligned}
\sum_{k=1}^K \frac{w_k n_k}{W_N} p_{k,t}(x_t,z) &= \sum_{k=1}^K \frac{w_k n_k}{W_N} p_k(z) p_{k,t}(x_t|z) \\
&\ge \frac{C_f}{B} p_{0,t}(x_t|z) \sum_{k=1}^K \frac{w_k n_k}{W_N} p_k(z) \\
&= \frac{C_f}{B} p_{0,t}(x_t|z) p_{w\setminus 0}(z) \\
&\ge \frac{C_f}{BC_{\mathrm{DR}}} p_{0,t}(x_t,z).
\end{aligned} \tag{168}$$

Recall that

$$p_{w,t}(x_t,z) = \frac{n_0 p_{0,t}(x_t,z) + \sum_{k=1}^K w_k n_k p_{k,t}(x_t,z)}{n_0 + W_N}.$$

Using (168), we have

$$\begin{aligned}
p_{w,t}(x_t,z) &\ge \frac{n_0 + W_N \cdot C_f/C_{\mathrm{DR}}B}{n_0+W_N} p_{0,t}(x_t,z) \\
&\ge \frac{C_f}{C_{\mathrm{DR}}B} p_{0,t}(x_t,z),
\end{aligned} \tag{169}$$

where the last inequality follows from $C_f/C_{\mathrm{DR}}B \le 1$. Therefore,

$$\frac{p_{0,t}(x_t,z)}{p_{w,t}(x_t,z)} \le \frac{C_{\mathrm{DR}}B}{C_f}. \tag{170}$$

For any measurable function g, by the definitions of $\mathcal{H}_0$ and $\mathcal{H}_w$, we have

$$\begin{aligned}
\|g\|_{\mathcal{H}_0}^2 &= \frac{1}{T-t_0}\int_{t_0}^{T}\int \|g(x_t,z,t)\|_2^2 p_{0,t}(x_t,z)\,dx_t\,dz\,dt \\
&= \frac{1}{T-t_0}\int_{t_0}^{T}\int \|g(x_t,z,t)\|_2^2 \frac{p_{0,t}(x_t,z)}{p_{w,t}(x_t,z)} p_{w,t}(x_t,z)\,dx_t\,dz\,dt \\
&\le \frac{C_{\mathrm{DR}}B}{C_f}\|g\|_{\mathcal{H}_w}^2.
\end{aligned} \tag{171}$$

Applying (171) to $g = \widehat{s}_w - s_w^o$ and taking expectation with respect to $D_0,\dots,D_K$ yield

$$\boxed{\mathbb{E}_{D_0,\dots,D_K}\left[\|\widehat{s}_w - s_w^o\|_{\mathcal{H}_0}^2\right] \le \frac{C_{\mathrm{DR}}B}{C_f}\mathbb{E}_{D_0,\dots,D_K}\left[\|\widehat{s}_w - s_w^o\|_{\mathcal{H}_w}^2\right] \lesssim \mathbb{E}_{D_0,\dots,D_K}\left[\|\widehat{s}_w - s_w^o\|_{\mathcal{H}_w}^2\right].} \tag{172}$$

Hence, any convergence rate established for the estimation error under $\mathcal{H}_w$ is preserved under $\mathcal{H}_0$.

□

K.5.2 Decomposition of the expected excess risk $\mathbb{E}_{D_0,\dots,D_K}\left[\overline{\mathcal{E}}_w(\widehat{s}_w)\right]$

Recall that in (99), the excess risk is defined as

$$\overline{\mathcal{E}}_w(s) = \overline{\mathcal{R}}_w(s) - \overline{\mathcal{R}}_w(s_w^o).$$

By the definition of $\overline{\mathcal{R}}_w(s)$ defined above (92), the relationship between $\mathcal{R}_k(s)$ and $\mathcal{L}_k(s)$ in (17), and the definition of $\mathcal{L}_k(s)$ (12), we have:

$$\begin{aligned}
\overline{\mathcal{E}}_w(s) &= \sum_{k=0}^{K}\pi_k\left[\mathcal{R}_k(s) - \mathcal{R}_k(s_w^o)\right] \\
&= \sum_{k=0}^{K}\pi_k\left[\mathcal{L}_k(s) - \mathcal{L}_k(s_w^o)\right] \\
&= \sum_{k=0}^{K}\pi_k\mathbb{E}_{(x,z)\sim P_k}\left[\ell(x,z;s) - \ell(x,z;s_w^o)\right] \\
&= \sum_{k=0}^{K}\pi_k\mathbb{E}_{D_k}\left[\frac{1}{n_k}\sum_{i=1}^{n_k}\ell(x_{k,i},z_{k,i};s) - \frac{1}{n_k}\sum_{i=1}^{n_k}\ell(x_{k,i},z_{k,i};s_w^o)\right].
\end{aligned} \tag{173}$$

We introduce an independent ghost sample for each domain. Specifically, let

$$D_k' := \left\{(x_{k,i}',z_{k,i}')\right\}_{i=1}^{n_k}, \qquad (x_{k,i}',z_{k,i}') \overset{\text{i.i.d.}}{\sim} P_k, \qquad k=0,\dots,K, \tag{174}$$

where D_k' is independent of D_k, and the collection $D' = (D_0',\dots,D_K')$ is independent of the training data $D = (D_0,\dots,D_K)$.

For any s, define the excess loss

$$h_s(x,z) := \ell(x,z;s) - \ell(x,z;s_w^o). \tag{175}$$

Then, conditional on the training data D, $\widehat{s}_w$ is fixed, and

$$\begin{aligned}
\overline{\mathcal{E}}_w(\widehat{s}_w) &= \sum_{k=0}^{K}\pi_k\mathbb{E}_{(X,Z)\sim P_k}\left[h_{\widehat{s}_w}(X,Z)\right] \\
&= \mathbb{E}_{D'}\left[\sum_{k=0}^{K}\pi_k\frac{1}{n_k}\sum_{i=1}^{n_k}h_{\widehat{s}_w}(x_{k,i}',z_{k,i}')\,\middle|\,D\right].
\end{aligned} \tag{176}$$

Consequently,

$$\mathbb{E}_D\left[\overline{\mathcal{E}}_w(\widehat{s}_w)\right] = \mathbb{E}_D\mathbb{E}_{D'}\left[\sum_{k=0}^{K}\pi_k\frac{1}{n_k}\sum_{i=1}^{n_k}h_{\widehat{s}_w}(x'_{k,i},z'_{k,i})\right]. \tag{177}$$

To control the contribution from the tail region, for some truncation level $R>0$, define the truncated loss as:

$$\ell^{\mathrm{tr}}(x,z;s) := \ell(x,z;s)\mathbf{1}\left\{\|x\|_\infty \le R,\ \|z\|_\infty \le R\right\}, \tag{178}$$

and the corresponding truncated excess loss as

$$\begin{aligned} h_s^{\mathrm{tr}}(x,z) &:= \ell^{\mathrm{tr}}(x,z;s) - \ell^{\mathrm{tr}}(x,z;s_w^o) \\ &= h_s(x,z)\mathbf{1}\left\{\|x\|_\infty \le R,\ \|z\|_\infty \le R\right\}. \end{aligned} \tag{179}$$

We next define the empirical excess losses evaluated on the training samples and the ghost samples, respectively. Let:

$$H_1 := \sum_{k=0}^{K}\pi_k\frac{1}{n_k}\sum_{i=1}^{n_k}h_{\widehat{s}_w}(x_{k,i},z_{k,i}), \tag{180}$$

and

$$H_1^{\mathrm{tr}} := \sum_{k=0}^{K}\pi_k\frac{1}{n_k}\sum_{i=1}^{n_k}h^{\mathrm{tr}}_{\widehat{s}_w}(x_{k,i},z_{k,i}). \tag{181}$$

Similarly, define

$$H_2 := \sum_{k=0}^{K}\pi_k\frac{1}{n_k}\sum_{i=1}^{n_k}h_{\widehat{s}_w}(x'_{k,i},z'_{k,i}), \tag{182}$$

and

$$H_2^{\mathrm{tr}} := \sum_{k=0}^{K}\pi_k\frac{1}{n_k}\sum_{i=1}^{n_k}h^{\mathrm{tr}}_{\widehat{s}_w}(x'_{k,i},z'_{k,i}). \tag{183}$$

By (177), we have

$$\mathbb{E}_D\left[\overline{\mathcal{E}}_w(\widehat{s}_w)\right] = \mathbb{E}_D\mathbb{E}_{D'}\left[H_2\right]. \tag{184}$$

Adding and subtracting H_2^{tr}, H_1^{tr}, and H_1 yields the decomposition

$$\begin{aligned} \mathbb{E}_D\left[\overline{\mathcal{E}}_w(\widehat{s}_w)\right] &= \mathbb{E}_D\mathbb{E}_{D'}\left[H_2 - H_2^{\mathrm{tr}}\right] \\ &\quad + \mathbb{E}_D\left[\mathbb{E}_{D'}\left[H_2^{\mathrm{tr}}\right] - H_1^{\mathrm{tr}}\right] \\ &\quad + \mathbb{E}_D\left[H_1^{\mathrm{tr}} - H_1\right] \\ &\quad + \mathbb{E}_D\left[H_1\right] \\ &:= A_2 + B + A_1 + C, \end{aligned} \tag{185}$$

where

$$A_2 := \mathbb{E}_D\mathbb{E}_{D'}\left[H_2 - H_2^{\mathrm{tr}}\right], \tag{186}$$

$$B := \mathbb{E}_D\left[\mathbb{E}_{D'}\left[H_2^{\mathrm{tr}}\right] - H_1^{\mathrm{tr}}\right], \tag{187}$$

$$A_1 := \mathbb{E}_D\left[H_1^{\mathrm{tr}} - H_1\right], \tag{188}$$

$$C := \mathbb{E}_D\left[H_1\right]. \tag{189}$$

K.5.3 Bounding A_1 and A_2

Uniform bound on the score-matching loss. We first state the following uniform bound on the score-matching loss.

Lemma 3 (Uniform bound of $\ell(\cdot,\cdot,s)$; Fu et al. (2024), Lemma D.1 and Appendix D.6.5)**.** *Suppose that the network parameters M_t, W, κ, L, S are configured according to Proposition C.4 of Fu et al. (2024), and let*

$$m_t := \frac{M_t}{\sqrt{\log N}}, \qquad M_\ell := \int_{t_0}^{T} m_t^2\,dt, \tag{190}$$

where N is a sufficiently large number. Then, uniformly over $s \in \mathcal{F}(M_t, W, \kappa, L, S)$ defined in (83) and $(x, z) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_z}$,

$$0 \le \ell^{\mathrm{tr}}(x, z; s) \le \ell(x, z; s) \lesssim M_\ell. \tag{191}$$

Moreover, under the network configuration of Fu et al. (2024)'s Proposition C.4,

$$m_t \lesssim \frac{1}{\sigma_t}.$$

Hence, if $t_0 = N^{-C_\sigma}$ and $T = C_\alpha \log N$, then

$$M_\ell \lesssim \int_{t_0}^{T} \frac{1}{\sigma_t^2}\, dt = \mathcal{O}\left(\log \frac{1}{t_0}\right) = \mathcal{O}(\log N). \tag{192}$$

Lemma 4 (Truncation bound of ReLU network $s \in \mathcal{F}$; Fu et al. (2024), (D.33) and Appendix D.6.5)**.** *Suppose that Assumption 2 holds for $P_0, \dots, P_K$, and let the jointly truncated loss be defined as in (178), namely,*

$$\ell^{\mathrm{tr}}(x, z; s) := \ell(x, z; s)\mathbf{1}\left\{\|x\|_\infty \le R,\ \|z\|_\infty \le R\right\}.$$

Then, uniformly over $k = 0, \dots, K$ and $s \in \mathcal{F}$,

$$\begin{aligned}\mathbb{E}_{(X,Z)\sim P_k}\left[\left|\ell(X, Z; s) - \ell^{\mathrm{tr}}(X, Z; s)\right|\right] &\lesssim \left[\exp(-C_x R^2) + \exp(-C_z R^2)\right] R M_\ell \\ &\lesssim \exp(-C_{\mathrm{sg}} R^2) R M_\ell,\end{aligned} \tag{193}$$

where

$$C_{\mathrm{sg}} := \min\{C_x, C_z\} > 0,$$

up to a universal multiplicative constant, and M_ℓ is the uniform bound on the score-matching loss given in Lemma 3. The constant C_{sg} is independent of k, N, and R.

Theorem 6 (Uniform bound of $\ell(\cdot, \cdot, s_w^o)$ and truncation bound of the mixture score s_w^o)**.** *Suppose that the conditions of Lemma 1 and 3 hold, and let ℓ^{tr} be defined as in (178). Suppose further that the network envelope is chosen such that*

$$m_t = \frac{M_t}{\sqrt{\log N}} \asymp \frac{1}{\sigma_t}, \qquad M_\ell = \int_{t_0}^{T} m_t^2\, dt. \tag{194}$$

Then, we have

$$\ell(x, z; s_w^o) \lesssim \|x\|_2^2 + M_\ell, \tag{195}$$

and for sufficiently large R, it holds that

$$\mathbb{E}_{(X,Z)\sim P_w}\left[\left|\ell(X, Z; s_w^o) - \ell^{\mathrm{tr}}(X, Z; s_w^o)\right|\right] \lesssim \exp(-C_{\mathrm{sg}} R^2) R M_\ell, \quad \textit{and} \tag{196}$$

$$\mathbb{E}_{(X,Z)\sim P_k}\left[\left|\ell(X, Z; s_w^o) - \ell^{\mathrm{tr}}(X, Z; s_w^o)\right|\right] \lesssim \exp(-C_{\mathrm{sg}} R^2) R M_\ell, \quad k = 0, \dots, K. \tag{197}$$

where $C_{\mathrm{sg}} > 0$ is independent of N, R, and the mixture weights.

Proof. By Lemma 1, (104),

$$p_{w,0}(x|z) = \exp\left(-C_x \|x\|_2^2/2\right) \widetilde{f}_w(x, z), \tag{198}$$

where

$$\widetilde{f}_w(x, z) \in \mathcal{H}^\beta\left(\mathbb{R}^{d_x} \times \mathbb{R}^{d_z}, B\right)$$

and

$$0 < C_f \le \widetilde{f}_w(x, z) \le B < \infty$$

uniformly over (x, z) and the mixture weights.

Define the Gaussian reference density

$$g(x) := C_G^{-1} \exp\left(-\frac{C_x \|x\|_2^2}{2}\right), \qquad C_G = \left(\frac{2\pi}{C_x}\right)^{d_x/2}. \tag{199}$$

Combining (104) and (199), we have

$$C_f C_G \le \frac{p_{w,0}(x|z)}{g(x)} \le C_G B, \qquad (x, z) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_z}. \tag{200}$$

Thus, g is the density of $\mathcal{N}(0, C_x^{-1} I_{d_x})$ and has the same Gaussian tail factor in x as $p_{w,0}(x|z)$.

Define the transition kernel of the forward process of diffusion model in Section 2.2 as

$$q_t(x_t|x) = \frac{1}{(2\pi\sigma_t^2)^{d_x/2}} \exp\left(-\frac{\|x_t - a_t x\|_2^2}{2\sigma_t^2}\right). \tag{201}$$

Let $\widetilde{p}_t$ denote the diffused density obtained by taking g as the initial density. By the forward process defined in Section 2.2 we obtain

$$p_{w,t}(x_t|z) = \int_{\mathbb{R}^{d_x}} q_t(x_t|x) p_{w,0}(x|z)\, dx, \tag{202}$$

$$\widetilde{p}_t(x_t) = \int_{\mathbb{R}^{d_x}} q_t(x_t|x) g(x)\, dx. \tag{203}$$

Since the Gaussian kernel is nonnegative, (200) implies

$$C_f C_G \widetilde{p}_t(x_t) \le p_{w,t}(x_t|z) \le C_G B \widetilde{p}_t(x_t). \tag{204}$$

Therefore,

$$C_f C_G \le \frac{p_{w,t}(x_t|z)}{\widetilde{p}_t(x_t)} \le C_G B. \tag{205}$$

By Bayes' rule, (200) (205), and the truth that $X_t \perp\!\!\!\perp Z|X_0$,

$$\begin{aligned} p_w(x|x_t, z) &= \frac{q_t(x_t|x) p_{w,0}(x|z)}{p_{w,t}(x_t|z)} \\ &\le \frac{B}{C_f} \frac{q_t(x_t|x) g(x)}{\widetilde{p}_t(x_t)}. \end{aligned} \tag{206}$$

Hence,

$$p_w(x|x_t, z) \le \frac{B}{C_f} \widetilde{p}(x|x_t), \tag{207}$$

where $\widetilde{p}(x|x_t)$ denotes the posterior distribution under the Gaussian reference model (199), i.e. $\widetilde{p}(x|x_t) = q_t(x_t|x)g(x)/\widetilde{p}_t(x_t)$. The posterior comparison in (207) allows us to control posterior moments under the mixture distribution by the corresponding moments under the Gaussian reference model. In particular, for any nonnegative measurable function h,

$$\mathbb{E}_{X_0 \sim p_w(\cdot|x_t, z)} [h(X_0)] \le \frac{B}{C_f} \mathbb{E}_{X_0 \sim \widetilde{p}(\cdot|x_t)} [h(X_0)]. \tag{208}$$

Under the reference model (199),

$$X_0 \sim \mathcal{N}\left(0, C_x^{-1} I_{d_x}\right),$$

and

$$X_t = a_t X_0 + \sigma_t \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I_{d_x}).$$

Therefore,

$$X_0|X_t = x_t \sim \mathcal{N}\left(\frac{a_t}{a_t^2 + C_x\sigma_t^2} x_t,\ \frac{\sigma_t^2}{a_t^2 + C_x\sigma_t^2} I_{d_x}\right). \tag{209}$$

It follows that

$$\mathbb{E}_{X_0 \sim \widetilde{p}(\cdot|x_t)} \left[\left\| \frac{x_t - a_t X_0}{\sigma_t^2} \right\|_2^2\right] = \frac{C_x^2}{(a_t^2 + C_x\sigma_t^2)^2} \|x_t\|_2^2 + \frac{d_x a_t^2}{\sigma_t^2 (a_t^2 + C_x\sigma_t^2)}. \tag{210}$$

For the Ornstein–Uhlenbeck forward process,

$$a_t^2 + \sigma_t^2 = 1,$$

and hence

$$a_t^2 + C_x\sigma_t^2 \ge \min\{1, C_x\} > 0.$$

Thus,

$$\mathbb{E}_{X_0\sim\widetilde{p}(\cdot|x_t)}\left[\left\|\frac{x_t-a_tX_0}{\sigma_t^2}\right\|_2^2\right]\lesssim\|x_t\|_2^2+\frac{1}{\sigma_t^2}. \tag{211}$$

By (109) in Lemma 1 and the denoising score-matching identity (Vincent, 2011),

$$s_w^o(x_t,z,t)=-\mathbb{E}_{X_0\sim p_w(\cdot|x_t,z)}\left[\frac{x_t-a_tX_0}{\sigma_t^2}\right]. \tag{212}$$

Therefore, Jensen's inequality, (207) and (211) imply

$$\begin{aligned}\|s_w^o(x_t,z,t)\|_2^2&\le\mathbb{E}_{X_0\sim p_w(\cdot|x_t,z)}\left[\left\|\frac{x_t-a_tX_0}{\sigma_t^2}\right\|_2^2\right]\\&\lesssim\mathbb{E}_{X_0\sim\widetilde{p}(\cdot|x_t)}\left[\left\|\frac{x_t-a_tX_0}{\sigma_t^2}\right\|_2^2\right]\\&\lesssim\|x_t\|_2^2+\frac{1}{\sigma_t^2}.\end{aligned} \tag{213}$$

Importantly, this bound is uniform over $z\in\mathbb{R}^{d_z}$ and the mixture weights.

We next derive an envelope for the score-matching loss. For fixed (x,z),

$$\begin{aligned}&\mathbb{E}_{X_t\sim\mathcal{N}(a_tx,\sigma_t^2I_{d_x})}\left[\left\|s_w^o(X_t,z,t)+\frac{X_t-a_tx}{\sigma_t^2}\right\|_2^2\right]\\&\le 2\mathbb{E}_{X_t\sim\mathcal{N}(a_tx,\sigma_t^2I_{d_x})}\left[\|s_w^o(X_t,z,t)\|_2^2\right]+2\mathbb{E}_{X_t\sim\mathcal{N}(a_tx,\sigma_t^2I_{d_x})}\left[\left\|\frac{X_t-a_tx}{\sigma_t^2}\right\|_2^2\right]\\&\lesssim\mathbb{E}_{X_t\sim\mathcal{N}(a_tx,\sigma_t^2I_{d_x})}\left[\|X_t\|_2^2\right]+\frac{1}{\sigma_t^2}.\end{aligned} \tag{214}$$

Since

$$\mathbb{E}_{X_t\sim\mathcal{N}(a_tx,\sigma_t^2I_{d_x})}\left[\|X_t\|_2^2\right]=a_t^2\|x\|_2^2+d_x\sigma_t^2\lesssim\|x\|_2^2+1, \tag{215}$$

we obtain

$$\mathbb{E}_{X_t\sim\mathcal{N}(a_tx,\sigma_t^2I_{d_x})}\left[\left\|s_w^o(X_t,z,t)+\frac{X_t-a_tx}{\sigma_t^2}\right\|_2^2\right]\lesssim\|x\|_2^2+\frac{1}{\sigma_t^2}. \tag{216}$$

Integrating over $t\in[t_0,T]$ and using (194),

$$\boxed{\ell(x,z;s_w^o)\lesssim\|x\|_2^2+M_\ell,} \tag{217}$$

which proves (195) and this envelope is uniform in z. Now define the truncation region

$$\mathcal{A}_R:=\{(x,z):\|x\|_\infty\le R,\ \|z\|_\infty\le R\}. \tag{218}$$

Since

$$\ell^{\mathrm{tr}}(x,z;s_w^o)=\ell(x,z;s_w^o)\mathbf{1}_{\mathcal{A}_R}(x,z),$$

we have

$$\begin{aligned}\mathbb{E}_{(X,Z)\sim P_w}\left[\left|\ell(X,Z;s_w^o)-\ell^{\mathrm{tr}}(X,Z;s_w^o)\right|\right]&=\mathbb{E}_{P_w}\left[\ell(X,Z;s_w^o)\mathbf{1}_{\mathcal{A}_R^c}(X,Z)\right]\\&=\underbrace{\mathbb{E}_{P_w}\left[\ell(X,Z;s_w^o)\mathbf{1}_{\{\|X\|_\infty>R\}}\right]}_{=:I_x}\\&\quad+\underbrace{\mathbb{E}_{P_w}\left[\ell(X,Z;s_w^o)\mathbf{1}_{\{\|X\|_\infty\le R,\,\|Z\|_\infty>R\}}\right]}_{=:I_z}.\end{aligned} \tag{219}$$

By Lemma 1,

$$p_{w,0}(x,z)\lesssim B\exp\left(-C_x\|x\|_2^2/2-C_z\|z\|_2^2/2\right). \tag{220}$$

Combining this with (195),

$$\begin{aligned} I_x &\lesssim \int_{\|x\|_\infty > R} \int_{\mathbb{R}^{d_z}} \left(\|x\|_2^2 + M_\ell\right) \exp\left(-C_x\|x\|_2^2/2 - C_z\|z\|_2^2/2\right) dz\, dx \\ &\lesssim \exp(-C_x' R^2) R M_\ell \end{aligned} \tag{221}$$

for some $C_x' > 0$.

Similarly,

$$\begin{aligned} I_z &\le \mathbb{E}_{P_w}\left[\ell(X, Z; s_w^o)\mathbf{1}_{\{\|Z\|_\infty > R\}}\right] \\ &\lesssim \int_{\|z\|_\infty > R} \int_{\mathbb{R}^{d_x}} \left(\|x\|_2^2 + M_\ell\right) \exp\left(-C_x\|x\|_2^2/2 - C_z\|z\|_2^2/2\right) dx\, dz \\ &\lesssim \exp(-C_z' R^2) R M_\ell \end{aligned} \tag{222}$$

for some $C_z' > 0$.

The last inequalities follow from the Gaussian tail-moment bound; any polynomial factor in R can be absorbed into the exponential term after decreasing the corresponding exponential constant.

Combining (219), (221), and (222),

$$\boxed{\begin{aligned} E_{(X,Z)\sim P_w}\left[\left|\ell(X, Z; s_w^o) - \ell^{\mathrm{tr}}(X, Z; s_w^o)\right|\right] &\lesssim \left[\exp(-C_x' R^2) + \exp(-C_z' R^2)\right] R M_\ell \\ &\lesssim \exp(-C_{\mathrm{sg}} R^2) R M_\ell, \end{aligned}} \tag{223}$$

where

$$C_{\mathrm{sg}} := \min\{C_x', C_z'\} > 0. \tag{224}$$

The constants are independent of the mixture weights because B, C_f, C_F, C_x, and C_z are uniform over $k = 0, \ldots, K$. This completes the proof.

The same truncation bound also holds uniformly under each P_k, $k = 0, \ldots, K$. Indeed, by Assumption 2,

$$p_k(x, z) \lesssim B \exp\left(-C_x\|x\|_2^2/2 - C_z\|z\|_2^2/2\right), \qquad k = 0, \ldots, K. \tag{225}$$

Moreover, the loss envelope (195),

$$\ell(x, z; s_w^o) \lesssim \|x\|_2^2 + M_\ell,$$

holds pointwise and uniformly in z and the mixture weights. Therefore, applying the same decomposition as (219) gives

$$\boxed{\begin{aligned} &\mathbb{E}_{(X,Z)\sim P_k}\left[\left|\ell(X, Z; s_w^o) - \ell^{\mathrm{tr}}(X, Z; s_w^o)\right|\right] \\ &\quad\lesssim \int_{\|x\|_\infty > R} \int_{\mathbb{R}^{d_z}} \left(\|x\|_2^2 + M_\ell\right) \exp\left(-C_x\|x\|_2^2/2 - C_z\|z\|_2^2/2\right) dz\, dx \\ &\qquad + \int_{\|z\|_\infty > R} \int_{\mathbb{R}^{d_x}} \left(\|x\|_2^2 + M_\ell\right) \exp\left(-C_x\|x\|_2^2/2 - C_z\|z\|_2^2/2\right) dx\, dz \\ &\quad\lesssim \exp(-C_{\mathrm{sg}} R^2) R M_\ell, \qquad k = 0, \ldots, K. \end{aligned}}$$

Thus, the same bound holds uniformly over all component distributions P_k.

□

Theorem 7 (Bounds of terms A_1 and A_2.)**.** *Suppose that the conditions of Lemma 4 and Theorem 6 hold. We can bound the term $|A_1| + |A_2|$ in (185) by*

$$|A_1| + |A_2| = \mathcal{O}\left(\exp\left(-C_{\mathrm{sg}} R^2\right) R M_\ell\right). \tag{226}$$

Proof. By the triangle inequality:

$$\left|h_s(x, z) - h_s^{\mathrm{tr}}(x, z)\right| \le \left|\ell(x, z; s) - \ell^{\mathrm{tr}}(x, z; s)\right| + \left|\ell(x, z; s_w^o) - \ell^{\mathrm{tr}}(x, z; s_w^o)\right|. \tag{227}$$

Consequently, by (193) and (226),

$$\mathbb{E}_{(x,z)\sim P_k}\left[\left|h_s(x,z)-h_s^{\mathrm{tr}}(x,z)\right|\right] \lesssim \exp\left(-C_{\mathrm{sg}}R^2\right) RM_\ell. \tag{228}$$

We now apply Lemma 4 and Theorem 6 to bound the two truncation terms A_1 and A_2 in (185).

Recall that

$$A_1 = \mathbb{E}_D\left[H_1^{\mathrm{tr}} - H_1\right]. \tag{229}$$

By the triangle inequality and (228),

$$\begin{aligned}
|A_1| &\le \mathbb{E}_D\left[\left|H_1^{\mathrm{tr}} - H_1\right|\right] \\
&\le \sum_{k=0}^{K}\pi_k \frac{1}{n_k}\sum_{i=1}^{n_k}\mathbb{E}_D\left[\left|h^{\mathrm{tr}}_{\widehat{s}_w}(X_{k,i},Z_{k,i}) - h_{\widehat{s}_w}(X_{k,i},Z_{k,i})\right|\right] \\
&\lesssim \sum_{k=0}^{K}\pi_k \exp\left(-C_{\mathrm{sg}}R^2\right) RM_\ell \\
&= \mathcal{O}\left(\exp\left(-C_{\mathrm{sg}}R^2\right) RM_\ell\right),
\end{aligned} \tag{230}$$

where the last equality follows from $\sum_{k=0}^{K}\pi_k = 1$.

Next we bound $A_2 = \mathbb{E}_D\mathbb{E}_{D'}\left[H_2 - H_2^{\mathrm{tr}}\right]$ by the similar procedure. Again, by (228), we have:

$$\begin{aligned}
|A_2| &\le \mathbb{E}_D\mathbb{E}_{D'}\left[\left|H_2 - H_2^{\mathrm{tr}}\right|\right] \\
&\le \sum_{k=0}^{K}\pi_k \frac{1}{n_k}\sum_{i=1}^{n_k}\mathbb{E}_D\mathbb{E}_{D'}\left[\left|h_{\widehat{s}_w}(X'_{k,i},Z'_{k,i}) - h^{\mathrm{tr}}_{\widehat{s}_w}(X'_{k,i},Z'_{k,i})\right|\right] \\
&\lesssim \sum_{k=0}^{K}\pi_k \exp\left(-C_{\mathrm{sg}}R^2\right) RM_\ell \\
&= \mathcal{O}\left(\exp\left(-C_{\mathrm{sg}}R^2\right) RM_\ell\right).
\end{aligned} \tag{231}$$

Combining (230) and (231), we obtain

$$\boxed{|A_1| + |A_2| = \mathcal{O}\left(\exp\left(-C_{\mathrm{sg}}R^2\right) RM_\ell\right).} \tag{232}$$

□

K.5.4 Bounding C

Recall (189) that

$$C := \mathbb{E}_D\left[H_1\right], \tag{233}$$

where H_1 is defined in (180). By the definition of h_s in (175),

$$\begin{aligned}
H_1 &= \sum_{k=0}^{K}\pi_k \frac{1}{n_k}\sum_{i=1}^{n_k}\left[\ell(x_{k,i},z_{k,i};\widehat{s}_w) - \ell(X_{k,i},Z_{k,i};s_w^o)\right] \\
&= \widehat{\mathcal{L}}_w(\widehat{s}_w) - \widehat{\mathcal{L}}_w(s_w^o),
\end{aligned} \tag{234}$$

where

$$\widehat{\mathcal{L}}_w(s) := \sum_{k=0}^{K}\pi_k \frac{1}{n_k}\sum_{i=1}^{n_k}\ell(x_{k,i},z_{k,i};s) \tag{235}$$

denotes the weighted empirical risk, defined in (10). Since $\widehat{s}_w$ is an empirical risk minimizer over $\mathcal{F}$,

$$\widehat{\mathcal{L}}_w(\widehat{s}_w) \le \widehat{\mathcal{L}}_w(s), \qquad \text{for every } s\in\mathcal{F}. \tag{236}$$

Therefore, for any fixed $s\in\mathcal{F}$,

$$C = \mathbb{E}_D\left[\widehat{\mathcal{L}}_w(\widehat{s}_w) - \widehat{\mathcal{L}}_w(s_w^o)\right] \le \mathbb{E}_D\left[\widehat{\mathcal{L}}_w(s) - \widehat{\mathcal{L}}_w(s_w^o)\right]. \tag{237}$$

Since both s and s_w^o are fixed with respect to the training samples, the empirical risks are unbiased estimators of their corresponding population risks. Because of (17),

$$\mathbb{E}_D\left[\widehat{\mathcal{L}}_w(s)\right] = \sum_{k=0}^{K} \pi_k \mathcal{L}_k(s) = \sum_{k=0}^{K} \pi_k \left[\mathcal{R}_k(s) - C_{\mathrm{DSM},k}\right] = \overline{\mathcal{R}}_w(s) - \sum_{k=0}^{K} \pi_k C_{\mathrm{DSM},k} \tag{238}$$

$$\mathbb{E}_D\left[\widehat{\mathcal{L}}_w(s_w^o)\right] = \sum_{k=0}^{K} \pi_k \mathcal{L}_k(s_w^o) = \sum_{k=0}^{K} \pi_k \left[\mathcal{R}_k(s_w^o) - C_{\mathrm{DSM},k}\right] = \overline{\mathcal{R}}_w(s_w^o) - \sum_{k=0}^{K} \pi_k C_{\mathrm{DSM},k}. \tag{239}$$

Consequently,

$$C \le \overline{\mathcal{R}}_w(s) - \overline{\mathcal{R}}_w(s_w^o) = \overline{\mathcal{E}}_w(s) = \|s - s_w^o\|_{\mathcal{H}_w}^2, \qquad \forall s \in \mathcal{F}. \tag{240}$$

Taking the infimum over $s \in \mathcal{F}$ gives

$$\boxed{C \le \inf_{s\in\mathcal{F}} \overline{\mathcal{E}}_w(s) = \inf_{s\in\mathcal{F}} \|s - s_w^o\|_{\mathcal{H}_w}^2.} \tag{241}$$

Thus, the term C is controlled by the approximation error of the network class $\mathcal{F}$ to the population-optimal score s_w^o under the $\mathcal{H}_w$ norm. Next, we invoke the score approximation result of Fu et al. (2024).

Lemma 5 (Conditional score approximation with unbounded covariates; adapted from Fu et al. (2024), Proposition C.4)**.** *Suppose that $z \in \mathbb{R}^{d_z}$ and the conditional density $p(x|z)$ satisfies*

$$p(x|z) = \exp\left(-C_x\|x\|_2^2/2\right) f(x,z), \tag{242}$$

where

$$f \in \mathcal{H}^\beta\left(\mathbb{R}^{d_x} \times \mathbb{R}^{d_z}, B\right), \qquad \inf_{(x,z)\in\mathbb{R}^{d_x}\times\mathbb{R}^{d_z}} f(x,z) \ge C_f > 0. \tag{243}$$

Suppose further that the marginal density of Z has a sub-Gaussian tail, i.e.,

$$p(z) \lesssim \exp\left(-C_z\|z\|_2^2/2\right) \tag{244}$$

for some constant $C_z > 0$. For $t > 0$, let $p_t(x_t|z)$ denote the conditional density obtained by applying the forward diffusion process to X while keeping Z unchanged, i.e.,

$$p_t(x_t|z) := \int_{\mathbb{R}^{d_x}} q_t(x_t|x)p(x|z), dx, \tag{245}$$

where $q_t(x_t|x)$ is defined in (201). Same with Lemma 3, for sufficiently large N and constants $C_\sigma, C_\alpha > 0$, let

$$t_0 = N^{-C_\sigma}, \qquad T = C_\alpha \log N. \tag{246}$$

Then there exists a ReLU network $s \in \mathcal{F}(M_t, W, \kappa, L, S)$ such that, for every $t \in [t_0, T]$,

$$\mathbb{E}_{Z\sim p(z)}\left[\int_{\mathbb{R}^{d_x}} \|s(x_t,Z,t) - \nabla_{x_t}\log p_t(x_t|Z)\|_2^2\, p_t(x_t|Z)\, dx_t\right] \lesssim \frac{B^2}{\sigma_t^2} N^{-\frac{2\beta}{d_x+d_z}} (\log N)^{\beta+1}. \tag{247}$$

The corresponding network parameters can be chosen such that

$$M_t = \mathcal{O}\left(\frac{\sqrt{\log N}}{\sigma_t}\right), \qquad W = \mathcal{O}\left(N\log^7 N\right), \tag{248}$$

and

$$\kappa = \exp\left(\mathcal{O}(\log^4 N)\right), \qquad L = \mathcal{O}(\log^4 N), \qquad S = \mathcal{O}\left(N\log^9 N\right), \tag{249}$$

where the choice of M_t coincides with (194).

Next, we give the bound of term C defined in (185).

Theorem 8 (Approximation error of the weighted score)**.** *Suppose that the conditions of Lemma 1 and 5 hold. Same with Lemma 3, let*

$$t_0 = N^{-C_\sigma}, \qquad T = C_\alpha \log N,$$

where C_σ and $C_\alpha > 0$ are constants. Then, for sufficiently large N, there exists $s_N \in \mathcal{F}(M_t, W, \kappa, L, S)$ such that

$$\|s_N - s_w^o\|_{\mathcal{H}_w}^2 \lesssim B^2 N^{-\frac{2\beta}{d_x+d_z}} (\log N)^{\beta+1}. \tag{250}$$

Consequently,

$$\inf_{s\in\mathcal{F}} \|s - s_w^o\|_{\mathcal{H}_w}^2 \lesssim B^2 N^{-\frac{2\beta}{d_x+d_z}} (\log N)^{\beta+1}. \tag{251}$$

Therefore, the approximation term C defined in (185) satisfies

$$C \lesssim B^2 N^{-\frac{2\beta}{d_x+d_z}} (\log N)^{\beta+1}. \tag{252}$$

Since B is the radius of the Hölder ball and it is a constant, we have

$$C = \mathcal{O}\left(N^{-\frac{2\beta}{d_x+d_z}} (\log N)^{\beta+1}\right). \tag{253}$$

Proof. As established in Lemma 1, the mixture conditional density $p_{w,t}(x_t|z)$ is defined in (108); the fact that $p_{w,0}(x|z)$ satisfies the smoothness and tail conditions required by Lemma 5 is established in (104)–(107); and

$$s_w^o(x_t, z, t) = \nabla_{x_t} \log p_{w,t}(x_t \mid z)$$

is proved in (109). Therefore, Lemma 5 can be applied directly to the weighted conditional density $p_{w,0}(x|z)$. For sufficiently large N, there exists

$$s_N \in \mathcal{F}(M_t, W, \kappa, L, S)$$

such that, for every $t \in [t_0, T]$,

$$\mathbb{E}_{Z\sim p_w(z)}\left[\int_{\mathbb{R}^{d_x}} \|s_N(x_t, Z, t) - s_w^o(x_t, Z, t)\|_2^2 \, p_{w,t}(x_t|Z)\, dx_t\right] \lesssim \frac{B^2}{\sigma_t^2} N^{-\frac{2\beta}{d_x+d_z}} (\log N)^{\beta+1}. \tag{254}$$

Equivalently, since

$$p_{w,t}(x_t, z) = p_{w,t}(x_t|z) p_w(z),$$

we have

$$\mathbb{E}_{(X_t,Z)\sim p_{w,t}}\left[\|s_N(X_t, Z, t) - s_w^o(X_t, Z, t)\|_2^2\right] \lesssim \frac{B^2}{\sigma_t^2} N^{-\frac{2\beta}{d_x+d_z}} (\log N)^{\beta+1}. \tag{255}$$

By the definition of the $\mathcal{H}_w$ norm in (91),

$$\begin{aligned}\|s_N - s_w^o\|_{\mathcal{H}_w}^2 &= \frac{1}{T-t_0}\int_{t_0}^T \mathbb{E}_{(X_t,Z)\sim p_{w,t}}\left[\|s_N(X_t, Z, t) - s_w^o(X_t, Z, t)\|_2^2\right] dt \\ &\lesssim B^2 N^{-\frac{2\beta}{d_x+d_z}} (\log N)^{\beta+1} \frac{1}{T-t_0}\int_{t_0}^T \frac{1}{\sigma_t^2}\, dt. \end{aligned} \tag{256}$$

To evaluate the remaining time integral, define

$$I_\sigma := \int_{t_0}^T \frac{1}{\sigma_t^2}\, dt \asymp M_\ell. \tag{257}$$

Under the choices

$$t_0 = N^{-C_\sigma}, \qquad T = C_\alpha \log N, \tag{258}$$

where $C_\sigma, C_\alpha > 0$ are constants, we next determine the order of I_σ.

For the Ornstein–Uhlenbeck forward process defined in Section 2.2,

$$\sigma_t^2 = 1 - e^{-t}.$$

Hence,

$$\begin{aligned} I_\sigma &= \int_{t_0}^{T} \frac{1}{1-e^{-t}}\, dt \\ &= \left[\log(e^t-1)\right]_{t_0}^{T} \\ &= \log\left(\frac{e^T-1}{e^{t_0}-1}\right). \end{aligned} \tag{259}$$

Since $T = C_\alpha \log N$, we have

$$\begin{aligned} \log(e^T-1) &= \log\left(N^{C_\alpha}-1\right) \\ &= C_\alpha \log N + \log\left(1-N^{-C_\alpha}\right) \\ &= C_\alpha \log N + o(1). \end{aligned} \tag{260}$$

On the other hand, since $t_0 = N^{-C_\sigma} \to 0$, the expansion

$$e^{t_0}-1 = t_0\left(1+\mathcal{O}(t_0)\right)$$

gives

$$\begin{aligned} \log(e^{t_0}-1) &= \log t_0 + \log\left(\frac{e^{t_0}-1}{t_0}\right) \\ &= -C_\sigma \log N + o(1). \end{aligned} \tag{261}$$

Combining (259)–(261), we obtain

$$I_\sigma = (C_\alpha + C_\sigma)\log N + o(\log N) = \mathcal{O}(\log N) \asymp M_\ell. \tag{262}$$

Meanwhile,

$$\begin{aligned} T - t_0 &= C_\alpha \log N - N^{-C_\sigma} \\ &= C_\alpha \log N\,(1+o(1)) = \mathcal{O}(\log N). \end{aligned} \tag{263}$$

Therefore,

$$\begin{aligned} \frac{I_\sigma}{T-t_0} &= \frac{(C_\alpha+C_\sigma)\log N + o(\log N)}{C_\alpha \log N + o(\log N)} \\ &= \frac{C_\alpha + C_\sigma}{C_\alpha} + o(1) \\ &= 1 + \frac{C_\sigma}{C_\alpha} + o(1) = \mathcal{O}(1). \end{aligned} \tag{264}$$

Substituting (264) into (256) yields

$$\|s_N - s_w^o\|_{\mathcal{H}_w}^2 \lesssim B^2 N^{-\frac{2\beta}{d_x+d_z}} (\log N)^{\beta+1}. \tag{265}$$

Since $s_N \in \mathcal{F}$, it follows that

$$\boxed{\inf_{s\in\mathcal{F}} \|s - s_w^o\|_{\mathcal{H}_w}^2 \le \|s_N - s_w^o\|_{\mathcal{H}_w}^2 \lesssim B^2 N^{-\frac{2\beta}{d_x+d_z}} (\log N)^{\beta+1}.} \tag{266}$$

Combining this with (241), we obtain

$$\boxed{C \lesssim B^2 N^{-\frac{2\beta}{d_x+d_z}} (\log N)^{\beta+1}.} \tag{267}$$

By Lemma 1, B_w is uniformly bounded over the mixture weights and is independent of the network complexity parameter N. Therefore,

$$\boxed{C = \mathcal{O}\left(N^{-\frac{2\beta}{d_x+d_z}} (\log N)^{\beta+1}\right).} \tag{268}$$

□

K.5.5 BOUNDING B

Population and empirical operators. For convenience, we introduce the population and empirical operators used in the subsequent analysis. For any measurable function $f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_z} \to \mathbb{R}$, define the population operator associated with the k-th domain by

$$\mathbb{P}_k f := \mathbb{E}_{(X,Z)\sim P_k}\left[f(X,Z)\right], \qquad k = 0, \ldots, K. \tag{269}$$

The corresponding empirical operator is defined as

$$\mathbb{P}_{k,n_k} f := \frac{1}{n_k}\sum_{i=1}^{n_k} f(x_{k,i}, z_{k,i}). \tag{270}$$

Recall that the weighted density defined in (86) is

$$p_{w,0}(x,z) = \sum_{k=0}^{K} \pi_k\, p_{k,0}(x,z), \qquad \pi_k \geq 0, \qquad \sum_{k=0}^{K} \pi_k = 1.$$

Accordingly, define the weighted population operator by

$$\begin{aligned}\mathbb{P}_w f &:= \sum_{k=0}^{K} \pi_k \mathbb{P}_k f \\ &= \sum_{k=0}^{K} \pi_k \mathbb{E}_{(X,Z)\sim P_k}\left[f(X,Z)\right] \\ &= \mathbb{E}_{(X,Z)\sim P_w}\left[f(X,Z)\right].\end{aligned} \tag{271}$$

Similarly, define the weighted empirical operator by

$$\begin{aligned}\mathbb{P}_{w,\mathbf{n}} f &:= \sum_{k=0}^{K} \pi_k \mathbb{P}_{k,n_k} f \\ &= \sum_{k=0}^{K} \frac{\pi_k}{n_k} \sum_{i=1}^{n_k} f(x_{k,i}, z_{k,i}).\end{aligned} \tag{272}$$

Hence, the weight assigned to each individual observation from the k-th domain is

$$\frac{\pi_k}{n_k} = \begin{cases} \dfrac{1}{n_0 + W_N}, & k = 0, \\ \dfrac{w_k}{n_0 + W_N}, & k = 1, \ldots, K. \end{cases} \tag{273}$$

Therefore, the weighted empirical operator can equivalently be written as

$$\mathbb{P}_{w,\mathbf{n}} f = \frac{\displaystyle\sum_{i=1}^{n_0} f(x_{0,i}, z_{0,i}) + \sum_{k=1}^{K} w_k \sum_{i=1}^{n_k} f(x_{k,i}, z_{k,i})}{n_0 + W_N}. \tag{274}$$

Define the ghost-sample empirical operator for the k-th domain by

$$\mathbb{P}'_{k,n_k} f := \frac{1}{n_k}\sum_{i=1}^{n_k} f(x'_{k,i}, z'_{k,i}), \tag{275}$$

and define the corresponding weighted ghost-sample empirical operator as

$$\begin{aligned}\mathbb{P}'_{w,\mathbf{n}} f &:= \sum_{k=0}^{K} \pi_k \mathbb{P}'_{k,n_k} f \\ &= \sum_{k=0}^{K} \frac{\pi_k}{n_k} \sum_{i=1}^{n_k} f(x'_{k,i}, z'_{k,i}).\end{aligned} \tag{276}$$

Recall that

$$h_s(x,z) := \ell(x,z;s) - \ell(x,z;s_w^o), \tag{277}$$

and

$$h_s^{\mathrm{tr}}(x,z) := \ell^{\mathrm{tr}}(x,z;s) - \ell^{\mathrm{tr}}(x,z;s_w^o). \tag{278}$$

Then the weighted population excess risk in (173) can be written as

$$\overline{\mathcal{E}}_w(s) = \mathbb{P}_w h_s. \tag{279}$$

Moreover, the quantities introduced in the decomposition of the expected excess risk (180)-(183) can be expressed compactly as

$$H_1 = \mathbb{P}_{w,\mathbf{n}} h_{\widehat{s}_w}, \qquad H_1^{\mathrm{tr}} = \mathbb{P}_{w,\mathbf{n}} h_{\widehat{s}_w}^{\mathrm{tr}}, \tag{280}$$

$$H_2 = \mathbb{P}'_{w,\mathbf{n}} h_{\widehat{s}_w}, \qquad H_2^{\mathrm{tr}} = \mathbb{P}'_{w,\mathbf{n}} h_{\widehat{s}_w}^{\mathrm{tr}}. \tag{281}$$

Conditioning on the training data D, we have

$$\begin{aligned}
\mathbb{E}_{D'}\left[\mathbb{P}'_{w,\mathbf{n}} h_{\widehat{s}_w} \,\middle|\, D\right] &= \sum_{k=0}^{K} \pi_k \mathbb{P}_k h_{\widehat{s}_w} \\
&= \mathbb{P}_w h_{\widehat{s}_w} \\
&= \overline{\mathcal{E}}_w(\widehat{s}_w).
\end{aligned} \tag{282}$$

$$\begin{aligned}
\mathbb{E}_D\left[\overline{\mathcal{E}}_w(\widehat{s}_w)\right] &= \mathbb{E}_{D,D'}\left[\mathbb{P}'_{w,\mathbf{n}} h_{\widehat{s}_w}\right] \\
&= \underbrace{\mathbb{E}_{D,D'}\left[\mathbb{P}'_{w,\mathbf{n}}\left(h_{\widehat{s}_w} - h_{\widehat{s}_w}^{\mathrm{tr}}\right)\right]}_{A_2} \\
&\quad + \underbrace{\mathbb{E}_D\left[\mathbb{E}_{D'}\left[\mathbb{P}'_{w,\mathbf{n}} h_{\widehat{s}_w}^{\mathrm{tr}} \,\middle|\, D\right] - \mathbb{P}_{w,\mathbf{n}} h_{\widehat{s}_w}^{\mathrm{tr}}\right]}_{B} \\
&\quad + \underbrace{\mathbb{E}_D\left[\mathbb{P}_{w,\mathbf{n}}\left(h_{\widehat{s}_w}^{\mathrm{tr}} - h_{\widehat{s}_w}\right)\right]}_{A_1} \\
&\quad + \underbrace{\mathbb{E}_D\left[\mathbb{P}_{w,\mathbf{n}} h_{\widehat{s}_w}\right]}_{C}.
\end{aligned} \tag{283}$$

We aim to bound $B = \mathbb{E}_{D,D'}\left[\left(\mathbb{P}'_{w,\mathbf{n}} - \mathbb{P}_{w,\mathbf{n}}\right) h_{\widehat{s}_w}^{\mathrm{tr}}\right]$.

Covering number of the loss function class. For a function class $\mathcal{G}$ equipped with a norm $\|\cdot\|$, let

$$\mathcal{N}\left(\delta, \mathcal{G}, \|\cdot\|\right) \tag{284}$$

denote its δ-covering number, defined as

$$\mathcal{N}\left(\delta, \mathcal{G}, \|\cdot\|\right) := \min\left\{ M \in \mathbb{N} : \exists\, g_1, \ldots, g_M \in \mathcal{G} \text{ such that } \sup_{g\in\mathcal{G}} \min_{1\le j\le M} \|g - g_j\| \le \delta \right\}. \tag{285}$$

In (218), the joint truncation region of x and z is defined as

$$\mathcal{A}_R = \left\{(x,z) : \|x\|_\infty \le R,\ \|z\|_\infty \le R\right\}. \tag{286}$$

We choose the truncation radius as

$$R = \mathcal{O}(\sqrt{\log N}), \tag{287}$$

where N is same with (190) and Theorem 8. We also choose $t_0 = N^{-C_\sigma}$ and $T = C_\alpha \log N$, same with Theorem 8. By (194) and (262), we have

$$M_\ell = \int_{t_0}^{T} m_t^2\, dt \asymp \int_{t_0}^{T} \frac{1}{\sigma_t^2}\, dt \asymp \log N. \tag{288}$$

Hence,

$$R^2 \lesssim M_\ell. \tag{289}$$

We define the loss function class restricted to $\mathcal{A}_R$ as

$$\mathcal{S}(R) := \left\{ \ell^{\mathrm{tr}}(\cdot, \cdot; s) : s \in \mathcal{F} \right\}. \tag{290}$$

Since

$$\ell^{\mathrm{tr}}(x, z; s) = \ell(x, z; s), \qquad (x, z) \in \mathcal{A}_R,$$

the covering number of the truncated loss class on $\mathcal{A}_R$ is identical to that of the original loss class restricted to $\mathcal{A}_R$. For any function $g : \mathcal{A}_R \to \mathbb{R}$, define

$$\|g\|_{L^\infty(\mathcal{A}_R)} := \sup_{(x,z)\in\mathcal{A}_R} |g(x, z)|. \tag{291}$$

The next Lemma gives the δ-covering number of of $\mathcal{S}(R)$ with respect to $\|\cdot\|_{L^\infty(\mathcal{A}_R)}$.

Lemma 6 (Covering number of the loss function class; Fu et al. (2024), Lemma D.8 and Eqs. (D.29)–(D.30))**.** *Let $\delta > 0$ and $R > 0$. For the ReLU network class $\mathcal{F}(M_t, W, \kappa, L, S)$ defined in (83), the δ-covering number of $\mathcal{S}(R)$ with respect to $\|\cdot\|_{L^\infty(\mathcal{A}_R)}$ satisfies*

$$\mathcal{N}\left(\delta, \mathcal{S}(R), \|\cdot\|_{L^\infty(\mathcal{A}_R)}\right) \lesssim \left(\frac{2L^2\big(W \max\{R, T\} + 2\big)\kappa^L W^{L+1} \log N}{\delta} \right)^{2S}. \tag{292}$$

Here, N denotes the network complexity parameter controlling the architecture of the ReLU network, whereas S denotes the number of nonzero network parameters.

In particular, under the network configuration associated with Lemma 5, namely,

$$W = \mathcal{O}\left(N \log^7 N\right), \qquad S = \mathcal{O}\left(N \log^9 N\right), \tag{293}$$

together with

$$L = \mathcal{O}(\log^4 N), \qquad \kappa = \exp\left(\mathcal{O}(\log^4 N)\right), \tag{294}$$

the logarithm of the covering number satisfies

$$\begin{aligned} &\log \mathcal{N}\left(\delta, \mathcal{S}(R), \|\cdot\|_{L^\infty(\mathcal{A}_R)}\right) \\ &\quad \lesssim N \log^9 N \left(\mathrm{Poly}(\log\log N) + \mathrm{Poly}(\log\log N) \log N \log R + \log^8 N + \log \frac{1}{\delta} \right) \\ &\quad \lesssim N \log^9 N \left(\log^8 N + \log^2 N \log R + \log \frac{1}{\delta} \right). \end{aligned} \tag{295}$$

Finite-cover reduction of B. Next, we discretize the truncated loss class using a finite δ-cover, thereby reducing the bound on B to a finite empirical process plus a covering approximation error. Recall that

$$B = \mathbb{E}_{D,D'}\left[\left(\mathbb{P}'_{w,\mathbf{n}} - \mathbb{P}_{w,\mathbf{n}} \right) h^{\mathrm{tr}}_{\widehat{s}_w} \right]. \tag{296}$$

Let

$$\mathcal{J} := \{\ell_1, \ldots, \ell_{\mathcal{N}}\} \subset \mathcal{S}(R) \tag{297}$$

be a δ-cover of $\mathcal{S}(R)$ with respect to the $L^\infty(\mathcal{A}_R)$ norm, where

$$\mathcal{N} := \mathcal{N}\left(\delta, \mathcal{S}(R), \|\cdot\|_{L^\infty(\mathcal{A}_R)}\right). \tag{298}$$

Since $\widehat{s}_w$ depends on the training sample D, there exists a random index

$$J = J(D) \in \{1, \ldots, \mathcal{N}\} \tag{299}$$

such that

$$\left\| \ell^{\mathrm{tr}}(\cdot, \cdot; \widehat{s}_w) - \ell_J \right\|_{L^\infty(\mathcal{A}_R)} \le \delta. \tag{300}$$

For each $r = 1, \ldots, \mathcal{N}$, define

$$h_r(x, z) := \ell_r(x, z) - \ell^{\mathrm{tr}}(x, z; s^o_w). \tag{301}$$

Recall that

$$h^{\mathrm{tr}}_{\widehat{s}_w}(x, z) = \ell^{\mathrm{tr}}(x, z; \widehat{s}_w) - \ell^{\mathrm{tr}}(x, z; s^o_w). \tag{302}$$

Therefore,

$$h^{\mathrm{tr}}_{\widehat{s}_w}(x,z) - h_J(x,z) = \ell^{\mathrm{tr}}(x,z;\widehat{s}_w) - \ell_J(x,z). \tag{303}$$

By (300) and the definition of the truncated loss function class defined in (290), both ℓ^{tr} and ℓ_J vanish outside $\mathcal{A}_R$. Hence, we have

$$\left\|h^{\mathrm{tr}}_{\widehat{s}_w} - h_J\right\|_{L^\infty(\mathcal{A}_R)} = \left\|h^{\mathrm{tr}}_{\widehat{s}_w} - h_J\right\|_\infty \le \delta. \tag{304}$$

Using the linearity of the empirical operators, we decompose

$$\left(\mathbb{P}'_{w,\mathbf{n}} - \mathbb{P}_{w,\mathbf{n}}\right) h^{\mathrm{tr}}_{\widehat{s}_w} = \left(\mathbb{P}'_{w,\mathbf{n}} - \mathbb{P}_{w,\mathbf{n}}\right) h_J + \left(\mathbb{P}'_{w,\mathbf{n}} - \mathbb{P}_{w,\mathbf{n}}\right)\left(h^{\mathrm{tr}}_{\widehat{s}_w} - h_J\right). \tag{305}$$

For the second term, since both weighted empirical operators are positive linear operators and

$$\mathbb{P}'_{w,\mathbf{n}} 1 = \mathbb{P}_{w,\mathbf{n}} 1 = 1, \tag{306}$$

we have

$$\begin{aligned}\left|\left(\mathbb{P}'_{w,\mathbf{n}} - \mathbb{P}_{w,\mathbf{n}}\right)\left(h^{\mathrm{tr}}_{\widehat{s}_w} - h_J\right)\right| &\le \mathbb{P}'_{w,\mathbf{n}}\left|h^{\mathrm{tr}}_{\widehat{s}_w} - h_J\right| + \mathbb{P}_{w,\mathbf{n}}\left|h^{\mathrm{tr}}_{\widehat{s}_w} - h_J\right| \\ &\le \left(\mathbb{P}'_{w,\mathbf{n}} 1 + \mathbb{P}_{w,\mathbf{n}} 1\right)\left\|h^{\mathrm{tr}}_{\widehat{s}_w} - h_J\right\|_\infty \\ &\le 2\delta. \end{aligned} \tag{307}$$

Consequently,

$$|B| \le \mathbb{E}_{D,D'}\left[\left|\left(\mathbb{P}'_{w,\mathbf{n}} - \mathbb{P}_{w,\mathbf{n}}\right) h_J\right|\right] + 2\delta. \tag{308}$$

Self-normalized control of the finite empirical process. We next introduce a self-normalized empirical process. For an arbitrary constant $a > 0$ and each $r = 1, \ldots, \mathcal{N}$, define

$$U_r := \max\left\{a,\ \sqrt{(\mathbb{P}_w h_r)_+}\right\}, \tag{309}$$

where

$$x_+ := \max\{x, 0\}.$$

Furthermore, define

$$D_w := \max_{1 \le r \le \mathcal{N}} \frac{\left|\left(\mathbb{P}'_{w,\mathbf{n}} - \mathbb{P}_{w,\mathbf{n}}\right) h_r\right|}{U_r}. \tag{310}$$

Since $J \in \{1, \ldots, \mathcal{N}\}$, it follows pointwise that

$$\begin{aligned}\left|\left(\mathbb{P}'_{w,\mathbf{n}} - \mathbb{P}_{w,\mathbf{n}}\right) h_J\right| &= U_J \frac{\left|\left(\mathbb{P}'_{w,\mathbf{n}} - \mathbb{P}_{w,\mathbf{n}}\right) h_J\right|}{U_J} \\ &\le U_J D_w. \end{aligned} \tag{311}$$

Substituting (311) into (308) gives

$$|B| \le \mathbb{E}_{D,D'}\left[U_J D_w\right] + 2\delta. \tag{312}$$

Finally, applying Young's inequality $uv \le (u^2 + v^2)/2$, we obtain

$$|B| \le \frac{1}{2}\mathbb{E}_{D,D'}\left[U_J^2\right] + \frac{1}{2}\mathbb{E}_{D,D'}\left[D_w^2\right] + 2\delta. \tag{313}$$

Bounding $\mathbb{E}[U_J^2]$. We first bound $\mathbb{E}_{D,D'}[U_J^2]$. By the definition of U_r in (309), we have,

$$U_J^2 = \max\left\{a^2,\ (\mathbb{P}_w h_J)_+\right\} \le a^2 + (\mathbb{P}_w h_J)_+. \tag{314}$$

By (304) and the positivity of $\mathbb{P}_w$,

$$\left|\mathbb{P}_w h_J - \mathbb{P}_w h^{\mathrm{tr}}_{\widehat{s}_w}\right| \le \mathbb{P}_w\left|h_J - h^{\mathrm{tr}}_{\widehat{s}_w}\right| \le \left\|h_J - h^{\mathrm{tr}}_{\widehat{s}_w}\right\|_\infty \le \delta. \tag{315}$$

Consequently,

$$(\mathbb{P}_w h_J)_+ \le \left(\mathbb{P}_w h^{\mathrm{tr}}_{\widehat{s}_w}\right)_+ + \delta. \tag{316}$$

Combining (314) and (316) gives

$$\mathbb{E}_{D,D'}\left[U_J^2\right] \le a^2 + \mathbb{E}_D\left[\left(\mathbb{P}_w h^{\mathrm{tr}}_{\widehat{s}_w}\right)_+\right] + \delta$$
$$= a^2 + \mathbb{E}_D\left[\left(\overline{\mathcal{E}}^{\mathrm{tr}}_w(\widehat{s}_w)\right)_+\right] + \delta, \tag{317}$$

where

$$\overline{\mathcal{E}}^{\mathrm{tr}}_w(s) := \mathbb{P}_w h^{\mathrm{tr}}_s \tag{318}$$

denotes the truncated population excess risk.

Moreover, by the truncation bound in Lemma 4 and Theorem 6,

$$\left|\overline{\mathcal{E}}^{\mathrm{tr}}_w(\widehat{s}_w) - \overline{\mathcal{E}}_w(\widehat{s}_w)\right| \le \mathbb{P}_w\left|\ell^{\mathrm{tr}}(\cdot,\cdot;\widehat{s}_w) - \ell(\cdot,\cdot;\widehat{s}_w)\right| + \mathbb{P}_w\left|\ell^{\mathrm{tr}}(\cdot,\cdot;s^o_w) - \ell(\cdot,\cdot;s^o_w)\right|$$
$$\lesssim \exp(-C_{\mathrm{sg}}R^2)RM_\ell. \tag{319}$$

Since $\overline{\mathcal{E}}_w(\widehat{s}_w) \ge 0$, it follows that

$$\mathbb{E}_{D,D'}\left[U_J^2\right] \lesssim a^2 + \mathbb{E}_D\left[\overline{\mathcal{E}}_w(\widehat{s}_w)\right] + \delta + \exp(-C_{\mathrm{sg}}R^2)RM_\ell. \tag{320}$$

Bounding $\mathbb{E}[D_w^2]$. We next control the second moment of the normalized empirical process D_w. For each fixed covering index $r \in \{1,\dots,\mathcal{N}\}$, define

$$Y_r := \frac{\left(\mathbb{P}'_{w,\mathbf{n}} - \mathbb{P}_{w,\mathbf{n}}\right)h_r}{U_r}. \tag{321}$$

Let

$$\lambda_k := \frac{\pi_k}{n_k}, \qquad \rho_w := \max_{0\le k\le K}\lambda_k. \tag{322}$$

Then

$$Y_r = \sum_{k=0}^{K}\sum_{i=1}^{n_k} X^{(r)}_{k,i}, \tag{323}$$

where

$$X^{(r)}_{k,i} := \frac{\lambda_k}{U_r}\left[h_r(X'_{k,i},Z'_{k,i}) - h_r(X_{k,i},Z_{k,i})\right]. \tag{324}$$

Since $(X'_{k,i},Z'_{k,i})$ and $(X_{k,i},Z_{k,i})$ are independent and identically distributed according to P_k,

$$\mathbb{E}\left[X^{(r)}_{k,i}\right] = 0. \tag{325}$$

Moreover, the random variables $\{X^{(r)}_{k,i}\}_{k,i}$ are mutually independent for every fixed r. By the uniform loss bound in (191), the mixture-score loss envelope in (195), and $R^2 \lesssim M_\ell$ in (289), there exists a universal constant $C_M > 0$ such that

$$\|h_r\|_\infty \le C_M M_\ell, \qquad r = 1,\dots,\mathcal{N}. \tag{326}$$

Since $U_r \ge a$, it follows that

$$\left|X^{(r)}_{k,i}\right| \le \frac{2C_M M_\ell \lambda_k}{a} \le \frac{2C_M M_\ell \rho_w}{a}. \tag{327}$$

For the total second moment, using $(u-v)^2 \le 2u^2 + 2v^2$, we obtain

$$\sum_{k=0}^{K}\sum_{i=1}^{n_k}\mathbb{E}\left[\left(X^{(r)}_{k,i}\right)^2\right] \le 4\sum_{k=0}^{K}\frac{n_k\lambda_k^2}{U_r^2}\mathbb{P}_k h_r^2$$
$$= 4\sum_{k=0}^{K}\frac{\lambda_k\pi_k}{U_r^2}\mathbb{P}_k h_r^2$$
$$\le \frac{4\rho_w}{U_r^2}\sum_{k=0}^{K}\pi_k\mathbb{P}_k h_r^2$$
$$= \frac{4\rho_w}{U_r^2}\mathbb{P}_w h_r^2. \tag{328}$$

A Bernstein-type bound for the truncated excess loss. We next establish a Bernstein-type second-moment bound for the functions $\{h_r\}_{r=1}^{\mathcal{N}}$.

Recall that each covering element $\ell_r \in \mathcal{S}(R)$ corresponds to some $s_r \in \mathcal{F}$ and (301), we have

$$h_r(x,z) = \ell^{\mathrm{tr}}(x,z;s_r) - \ell^{\mathrm{tr}}(x,z;s_w^o). \tag{329}$$

For notational simplicity, for a fixed (x,z) define

$$A_r(x,z) := \frac{1}{T-t_0}\int_{t_0}^{T} \mathbb{E}_{X_t\sim\mathcal{N}(a_t x,\sigma_t^2 I_{d_x})}\left[\|s_r(X_t,z,t) - s_w^o(X_t,z,t)\|_2^2\right] dt, \tag{330}$$

$$B_r(x,z) := \frac{1}{T-t_0}\int_{t_0}^{T} \mathbb{E}_{X_t\sim\mathcal{N}(a_t x,\sigma_t^2 I_{d_x})}\Bigg[\Big\langle s_r(X_t,z,t) - s_w^o(X_t,z,t), s_w^o(X_t,z,t) + \frac{X_t - a_t x}{\sigma_t^2}\Big\rangle\Bigg] dt. \tag{331}$$

By expanding the difference of the two squared losses, we obtain

$$h_r(x,z) = \mathbf{1}_{\mathcal{A}_R}(x,z)\left[A_r(x,z) + 2B_r(x,z)\right]. \tag{332}$$

By the Cauchy–Schwarz inequality,

$$\begin{aligned}|B_r(x,z)|^2 &\le A_r(x,z)\cdot\frac{1}{T-t_0}\int_{t_0}^{T}\mathbb{E}_{X_t\sim\mathcal{N}(a_t x,\sigma_t^2 I_{d_x})}\left[\left\|s_w^o(X_t,z,t) + \frac{X_t-a_t x}{\sigma_t^2}\right\|_2^2\right] dt\\ &= A_r(x,z)\cdot\ell(x,z;s_w^o).\end{aligned} \tag{333}$$

Moreover, using $\|u-v\|_2^2 \le 2\|u-y\|_2^2 + 2\|v-y\|_2^2$, we have

$$A_r(x,z) \le 2\ell(x,z;s_r) + 2\ell(x,z;s_w^o). \tag{334}$$

By the uniform loss bounds in (191), (195), and (289), there exists a constant $C_\ell > 0$ such that, uniformly over r and $(x,z)\in\mathcal{A}_R$,

$$\ell(x,z;s_r) \le C_\ell M_\ell, \qquad \ell(x,z;s_w^o) \le R^2 + M_\ell \le C_\ell M_\ell. \tag{335}$$

Consequently,

$$A_r(x,z) \le 4C_\ell M_\ell, \qquad |B_r(x,z)|^2 \le C_\ell M_\ell A_r(x,z). \tag{336}$$

Applying $(a+b)^2 \le 2a^2 + 2b^2$ to (332), we obtain

$$\begin{aligned}h_r^2(x,z) &\le \mathbf{1}_{\mathcal{A}_R}(x,z)\left[2A_r^2(x,z) + 8B_r^2(x,z)\right]\\ &\lesssim M_\ell \mathbf{1}_{\mathcal{A}_R}(x,z)A_r(x,z).\end{aligned} \tag{337}$$

Taking expectation with respect to P_w gives

$$\mathbb{P}_w h_r^2 \lesssim M_\ell \mathbb{P}_w\left[\mathbf{1}_{\mathcal{A}_R}A_r\right] \le M_\ell \mathbb{P}_w A_r. \tag{338}$$

We now identify $\mathbb{P}_w A_r$ with the population excess risk. By the denoising-score identity (Vincent, 2011) and (271),

$$\mathbb{E}_{X_0\sim p_w(\cdot|X_t=x_t,Z=z)}\left[\frac{x_t - a_t X_0}{\sigma_t^2}\right] = -s_w^o(x_t,z,t). \tag{339}$$

$$\begin{aligned}\mathbb{P}_w B_r &= \frac{1}{T-t_0}\int_{t_0}^{T}\mathbb{E}_{(X_0,Z)\sim P_w}\mathbb{E}_{X_t\sim\mathcal{N}(a_t X_0,\sigma_t^2 I_{d_x})}\Bigg[\Big\langle s_r(X_t,Z,t) - s_w^o(X_t,Z,t), s_w^o(X_t,Z,t) + \frac{X_t - a_t X_0}{\sigma_t^2}\Big\rangle\Bigg] dt\\ &= \frac{1}{T-t_0}\int_{t_0}^{T}\mathbb{E}_{(X_t,Z)\sim p_{w,t}}\Bigg[\Big\langle s_r(X_t,Z,t) - s_w^o(X_t,Z,t), s_w^o(X_t,Z,t) + \mathbb{E}_{X_0\sim p_w(\cdot|X_t,Z)}\left[\frac{X_t - a_t X_0}{\sigma_t^2}\right]\Big\rangle\Bigg] dt\\ &= 0.\end{aligned} \tag{340}$$

Hence, for the untruncated excess loss

$$h_r^{\mathrm{full}}(x,z) := \ell(x,z;s_r) - \ell(x,z;s_w^o), \tag{341}$$

we have

$$\mathbb{P}_w h_r^{\mathrm{full}} = \mathbb{P}_w A_r = \|s_r - s_w^o\|_{\mathcal{H}_w}^2 \geq 0. \tag{342}$$

Combining (338) and (342), we obtain

$$\mathbb{P}_w h_r^2 \lesssim M_\ell \mathbb{P}_w h_r^{\mathrm{full}}. \tag{343}$$

By Lemma 4 and Theorem 6,

$$\begin{aligned}\left|\mathbb{P}_w h_r^{\mathrm{full}} - \mathbb{P}_w h_r\right| &\leq \sum_{k=0}^{K} \pi_k \mathbb{E}_{P_k}\left[\left|h_r^{\mathrm{full}} - h_r\right|\right] \\ &\lesssim \exp(-C_{\mathrm{sg}} R^2) R M_\ell.\end{aligned} \tag{344}$$

Define

$$\tau_R := C_{\mathrm{tr}} \exp(-C_{\mathrm{sg}} R^2) R M_\ell \tag{345}$$

for a sufficiently large constant $C_{\mathrm{tr}} > 0$. Since $\mathbb{P}_w h_r^{\mathrm{full}} \geq 0$, it follows that

$$\mathbb{P}_w h_r^{\mathrm{full}} \leq (\mathbb{P}_w h_r)_+ + \tau_R. \tag{346}$$

Combining (343) and (346), we arrive at

$$\boxed{\mathbb{P}_w h_r^2 \lesssim M_\ell \left[(\mathbb{P}_w h_r)_+ + \tau_R\right].} \tag{347}$$

If $a > 0$ is chosen such that

$$a^2 \geq \tau_R, \tag{348}$$

then, by the definition of U_r in (309),

$$(\mathbb{P}_w h_r)_+ + \tau_R \leq 2 \max\left\{a^2, (\mathbb{P}_w h_r)_+\right\} = 2U_r^2. \tag{349}$$

Therefore,

$$\boxed{\mathbb{P}_w h_r^2 \lesssim M_\ell U_r^2, \qquad r = 1, \ldots, \mathcal{N}.} \tag{350}$$

Equivalently,

$$\boxed{\frac{\mathbb{P}_w h_r^2}{U_r^2} \lesssim M_\ell.} \tag{351}$$

Bounding the normalized empirical process D_w. In (321), we defined

$$Y_r = \frac{\left(\mathbb{P}'_{w,\mathbf{n}} - \mathbb{P}_{w,\mathbf{n}}\right) h_r}{U_r} = \sum_{k=0}^{K} \sum_{i=1}^{n_k} X_{k,i}^{(r)}, \tag{352}$$

where $X_{k,i}^{(r)}$ is defined in (324). For every fixed r, the random variables $\{X_{k,i}^{(r)}\}_{k,i}$ are mutually independent and centered:

$$\mathbb{E}\left[X_{k,i}^{(r)}\right] = 0. \tag{353}$$

Let $C_1 = 2C_M > 0$ and by (327), we have

$$\left|X_{k,i}^{(r)}\right| \leq \frac{C_1 M_\ell \rho_w}{a}, \tag{354}$$

By (328) and the Bernstein-type bound (350), provided that $a^2 \geq \tau_R$ we have

$$\begin{aligned}\sum_{k=0}^{K} \sum_{i=1}^{n_k} \mathbb{E}\left[\left(X_{k,i}^{(r)}\right)^2\right] &\leq \frac{4\rho_w}{U_r^2} \mathbb{P}_w h_r^2 \\ &\leq C_2 M_\ell \rho_w\end{aligned} \tag{355}$$

for some constant $C_2 > 0$. Therefore, Bernstein's inequality (Boucheron et al., 2013) gives, for every $t > 0$,

$$\mathbb{P}\left(|Y_r| \geq t\right) \leq 2\exp\left\{-\frac{t^2/2}{C_2 M_\ell \rho_w + \frac{C_1 M_\ell \rho_w}{3a} t}\right\} \leq 2\exp\left[-c\min\left\{\frac{t^2}{M_\ell \rho_w}, \frac{at}{M_\ell \rho_w}\right\}\right] \tag{356}$$

for some universal constant $c > 0$. Recall the definition of D_w in (310), Y_r in (321), and (356),

$$\mathbb{P}\left(D_w \geq t\right) \leq \sum_{r=1}^{\mathcal{N}} \mathbb{P}\left(|Y_r| \geq t\right) \leq 2\mathcal{N}\exp\left[-c\min\left\{\frac{t^2}{M_\ell \rho_w}, \frac{at}{M_\ell \rho_w}\right\}\right]. \tag{357}$$

Let

$$L_{\mathcal{N}} := \log(2\mathcal{N}). \tag{358}$$

Set

$$A_w := M_\ell \rho_w, \qquad B_w^{\mathrm{emp}} := \frac{M_\ell \rho_w}{a}. \tag{359}$$

Then (357) can be written as

$$\mathbb{P}\left(D_w \geq t\right) \leq \exp\left[L_{\mathcal{N}} - c\min\left\{\frac{t^2}{A_w}, \frac{t}{B_w^{\mathrm{emp}}}\right\}\right]. \tag{360}$$

A standard integration of the sub-exponential tail bound yields

$$\mathbb{E}_{D,D'}\left[D_w^2\right] \lesssim A_w L_{\mathcal{N}} + \left(B_w^{\mathrm{emp}}\right)^2 L_{\mathcal{N}}^2. \tag{361}$$

Therefore,

$$\mathbb{E}_{D,D'}\left[D_w^2\right] \lesssim M_\ell \rho_w L_{\mathcal{N}} + \frac{M_\ell^2 \rho_w^2}{a^2} L_{\mathcal{N}}^2. \tag{362}$$

Indeed, using

$$\mathbb{E}[D_w^2] = 2\int_0^\infty t\,\mathbb{P}(D_w \geq t)\,dt,$$

and splitting the integral at a constant multiple of

$$\sqrt{A_w L_{\mathcal{N}}} + B_w^{\mathrm{emp}} L_{\mathcal{N}},$$

gives (361).

We choose $a > 0$ such that

$$a^2 \asymp M_\ell \rho_w L_{\mathcal{N}} + \tau_R, \tag{363}$$

where

$$\tau_R = C_{\mathrm{tr}} \exp(-C_{\mathrm{sg}} R^2) R M_\ell. \tag{364}$$

In particular,

$$a^2 \gtrsim M_\ell \rho_w L_{\mathcal{N}}, \qquad a^2 \geq \tau_R. \tag{365}$$

Under the choice (363),

$$\frac{M_\ell^2 \rho_w^2}{a^2} L_{\mathcal{N}}^2 \lesssim M_\ell \rho_w L_{\mathcal{N}}. \tag{366}$$

Consequently,

$$\boxed{\mathbb{E}_{D,D'}\left[D_w^2\right] \lesssim M_\ell \rho_w L_{\mathcal{N}}.} \tag{367}$$

Recall from (320) that

$$\mathbb{E}_D\left[U_J^2\right] \le a^2 + \mathbb{E}_D\left[\overline{\mathcal{E}}_w(\widehat{s}_w)\right] + \delta + \tau_R. \tag{368}$$

Combining (313), (367), and (368), we obtain

$$|B| \le \frac{1}{2}\mathbb{E}_D\left[\overline{\mathcal{E}}_w(\widehat{s}_w)\right] + \frac{1}{2}a^2 + C_D M_\ell \rho_w L_{\mathcal{N}} + \frac{1}{2}\tau_R + \frac{5}{2}\delta, \tag{369}$$

for some constant $C_D > 0$. By (363),

$$a^2 \lesssim M_\ell \rho_w L_{\mathcal{N}} + \tau_R.$$

Therefore, absorbing the term $a^2/2$ and the numerical constants into a constant $C_B > 0$, we obtain

$$\boxed{|B| \le \frac{1}{2}\mathbb{E}_D\left[\overline{\mathcal{E}}_w(\widehat{s}_w)\right] + C_B\left[M_\ell \rho_w L_{\mathcal{N}} + \tau_R + \delta\right].} \tag{370}$$

Effective sample size and maximal observation weight. To characterize the variance of the weighted empirical risk, recall (274) that

$$\mathbb{P}_{w,\mathbf{n}} f = \sum_{k=0}^{K} \pi_k \mathbb{P}_{k,n_k} f = \sum_{k=0}^{K} \frac{\pi_k}{n_k} \sum_{i=1}^{n_k} f(X_{k,i}, Z_{k,i}). \tag{371}$$

Since the samples are independent across domains, for any measurable function f with finite second moments,

$$\mathrm{Var}\left(\mathbb{P}_{w,\mathbf{n}} f\right) = \sum_{k=0}^{K} \frac{\pi_k^2}{n_k} \mathrm{Var}_{P_k}\left(f(X,Z)\right). \tag{372}$$

Hence, the quantity

$$\nu_w := \sum_{k=0}^{K} \frac{\pi_k^2}{n_k} \tag{373}$$

plays the role of the variance factor of the weighted empirical risk. Motivated by the usual variance factor $1/n$ in standard ERM, we define the effective sample size by

$$N_{\mathrm{eff}} := \nu_w^{-1} = \left(\sum_{k=0}^{K} \frac{\pi_k^2}{n_k}\right)^{-1}. \tag{374}$$

Recall (33) that

$$\pi_0 = \frac{n_0}{n_0 + W_N}, \qquad \pi_k = \frac{w_k n_k}{n_0 + W_N}, \quad k = 1, \dots, K, \tag{375}$$

where $W_N = \sum_{k=1}^{K} w_k n_k$. Therefore,

$$\begin{aligned} \frac{1}{N_{\mathrm{eff}}} &= \frac{\pi_0^2}{n_0} + \sum_{k=1}^{K} \frac{\pi_k^2}{n_k} \\ &= \frac{n_0 + \sum_{k=1}^{K} w_k^2 n_k}{(n_0 + W_N)^2}, \end{aligned} \tag{376}$$

or equivalently,

$$\boxed{N_{\mathrm{eff}} = \frac{(n_0 + W_N)^2}{n_0 + \sum_{k=1}^{K} w_k^2 n_k}.} \tag{377}$$

Recall the definition of λ_k and ρ_w in (322), and by the definition of π_k, we have

$$\rho_w = \frac{\max\{1, w_1, \dots, w_K\}}{n_0 + W_N}. \tag{378}$$

Moreover,

$$\frac{1}{N_{\text{eff}}} = \sum_{k=0}^{K} \frac{\pi_k^2}{n_k} = \sum_{k=0}^{K} n_k \lambda_k^2 \leq \rho_w \sum_{k=0}^{K} n_k \lambda_k = \rho_w \sum_{k=0}^{K} \pi_k = \rho_w. \tag{379}$$

Thus,

$$\frac{1}{N_{\text{eff}}} \leq \rho_w. \tag{380}$$

Using the assumption in (26), we obtain

$$\max\{1, w_1, \ldots, w_K\} \leq C_{\text{reg}} \frac{n_0 + \sum_{k=1}^{K} w_k^2 n_k}{n_0 + W_N}, \tag{381}$$

where $C_{\text{reg}} \geq 1$. Using (377) and (378), condition (381) is equivalent to

$$\rho_w \leq \frac{C_{\text{reg}}}{N_{\text{eff}}}. \tag{382}$$

Together with (379), this yields

$$\boxed{\rho_w \asymp \frac{1}{N_{\text{eff}}}.} \tag{383}$$

Consequently, under (381),

$$M_\ell \rho_w \log(2\mathcal{N}) \lesssim \frac{M_\ell \log(2\mathcal{N})}{N_{\text{eff}}}. \tag{384}$$

K.5.6 Proof of Theorem 1

Combining the bounds. Combining (185), (226), (268), and (370), we obtain

$$\begin{aligned}
\mathbb{E}_D\left[\overline{\mathcal{E}}_w(\widehat{s}_w)\right] &\leq |A_1| + |A_2| + |B| + C \\
&\leq \frac{1}{2}\mathbb{E}_D\left[\overline{\mathcal{E}}_w(\widehat{s}_w)\right] + C_B M_\ell \rho_w L_{\mathcal{N}} + C_B \tau_R + C_B \delta \\
&\quad + C_A \exp\left(-C_{\text{sg}} R^2\right) R M_\ell + C_{\text{app}} N^{-\frac{2\beta}{d_x + d_z}} (\log N)^{\beta+1},
\end{aligned} \tag{385}$$

where C_A, C_B, and C_{app} are positive constants independent of the sample sizes and N. Therefore, moving the first term on the right-hand side to the left-hand side yields

$$\begin{aligned}
\mathbb{E}_D\left[\overline{\mathcal{E}}_w(\widehat{s}_w)\right] &\lesssim N^{-\frac{2\beta}{d_x + d_z}} (\log N)^{\beta+1} + M_\ell \rho_w L_{\mathcal{N}} + \delta \\
&\quad + \tau_R + \exp\left(-C_{\text{sg}} R^2\right) R M_\ell.
\end{aligned} \tag{386}$$

Consequently, since $\tau_R \lesssim \exp(-C_{\text{sg}} R^2) R M_\ell$, we arrive at

$$\boxed{\mathbb{E}_D\left[\overline{\mathcal{E}}_w(\widehat{s}_w)\right] \lesssim N^{-\frac{2\beta}{d_x + d_z}} (\log N)^{\beta+1} + M_\ell \rho_w L_{\mathcal{N}} + \delta + \exp\left(-C_{\text{sg}} R^2\right) R M_\ell.} \tag{387}$$

Choice of tuning parameters and convergence rate. Recall that in (24), the effective sample size is defined as

$$N_{\text{eff}} = \frac{\left(n_0 + \sum_{k=1}^{K} w_k n_k\right)^2}{n_0 + \sum_{k=1}^{K} w_k^2 n_k}. \tag{388}$$

Moreover, recall that

$$\rho_w = \frac{\max\{1, w_1, \ldots, w_K\}}{n_0 + \sum_{k=1}^{K} w_k n_k}. \tag{389}$$

Under Assumption (26),

$$\begin{aligned}
\rho_w &\lesssim \frac{n_0 + \sum_{k=1}^{K} w_k^2 n_k}{\left(n_0 + \sum_{k=1}^{K} w_k n_k\right)^2} \\
&= \frac{1}{N_{\text{eff}}}.
\end{aligned} \tag{390}$$

Following the parameter choices in Fu et al. (2024), proof of Theorem 4.1, we take

$$\delta = N^{-\frac{2\beta}{d_x+d_z}}, \qquad R^2 = C_R \log N, \tag{391}$$

where $C_R > 0$ is chosen sufficiently large. By (358) and the covering-number bound in Lemma 6,

$$\begin{aligned} L_{\mathcal{N}} &= \log(2\mathcal{N}) \\ &\lesssim N \log^9 N \left(\log^8 N + \log^2 N \log R + \log \frac{1}{\delta} \right) \\ &\lesssim N \log^{17} N. \end{aligned} \tag{392}$$

Moreover, in (192)

$$M_\ell \lesssim \log \frac{1}{t_0}. \tag{393}$$

By choosing C_R sufficiently large, the truncation remainder $\exp(-C_{\mathrm{sg}} R^2) R M_\ell$ is of smaller order than the approximation term and can therefore be absorbed into the leading terms. The choice of δ in (391) is also dominated by the approximation term. Therefore, using (390), we obtain

$$\mathbb{E}_D \left[\overline{\mathcal{E}}_w(\widehat{s}_w) \right] \lesssim N^{-\frac{2\beta}{d_x+d_z}} (\log N)^{\beta+1} + \frac{N}{N_{\mathrm{eff}}} \log \frac{1}{t_0} \log^{17} N. \tag{394}$$

Balancing the polynomial parts of the approximation and estimation terms in (394), namely,

$$N^{-\frac{2\beta}{d_x+d_z}} \qquad \text{and} \qquad \frac{N}{N_{\mathrm{eff}}},$$

we choose

$$N \asymp N_{\mathrm{eff}}^{\frac{d_x+d_z}{d_x+d_z+2\beta}}. \tag{395}$$

Under this choice,

$$N^{-\frac{2\beta}{d_x+d_z}} \asymp \frac{N}{N_{\mathrm{eff}}} \asymp N_{\mathrm{eff}}^{-\frac{2\beta}{d_x+d_z+2\beta}}, \tag{396}$$

and

$$\log N \asymp \log N_{\mathrm{eff}}. \tag{397}$$

Since $t_0 = N^{-C_\sigma}$,

$$\log \frac{1}{t_0} = C_\sigma \log N. \tag{398}$$

Therefore,

$$N^{-\frac{2\beta}{d_x+d_z}} (\log N)^{\beta+1} \lesssim \log \frac{1}{t_0} N^{-\frac{2\beta}{d_x+d_z}} (\log N)^{\beta}. \tag{399}$$

Consequently, substituting (395) into (394) yields

$$\boxed{\mathbb{E}_D \left[\overline{\mathcal{E}}_w(\widehat{s}_w) \right] \lesssim \log \frac{1}{t_0} N_{\mathrm{eff}}^{-\frac{2\beta}{d_x+d_z+2\beta}} (\log N_{\mathrm{eff}})^{\max\{17,\beta\}}.} \tag{400}$$

Target-domain risk bound. Combining (100), (164), (172), (400), and , we obtain

$$\begin{aligned} \mathbb{E}_{D_0,\ldots,D_K} \left[\mathcal{R}_0(\widehat{s}_w) \right] \lesssim{}& \log \frac{1}{t_0} N_{\mathrm{eff}}^{-\frac{2\beta}{d_x+d_z+2\beta}} (\log N_{\mathrm{eff}})^{\max\{17,\beta\}} \\ &+ \frac{W_N^2}{(n_0+W_N)^2} \left[\overline{w}^\top G \overline{w} + \mathrm{diag}(G)^\top \overline{w} \right]. \end{aligned} \tag{401}$$

□

K.6 THE TV ERROR BOUND OF DA-DIFF

K.6.1 PROOF OF THEOREM 2

Lemma 7 (Target-domain total variation bound; Fu et al. (2024), Appendix D.3)**.** *Let $\widehat{P}_w(\cdot|z)$ denote the conditional distribution generated by the backward diffusion process using the estimated score network $\widehat{s}_w$, and let $P_0(\cdot|z)$ denote the target conditional distribution. Then*

$$\begin{aligned}&\mathbb{E}_{D_0,\dots,D_K}\mathbb{E}_{Z\sim p_0}\left[d_{\mathrm{TV}}\left(P_0(\cdot|Z),\widehat{P}_w(\cdot|z)\right)\right]\\&\lesssim \sqrt{t_0}\left(\log\frac{1}{t_0}\right)^{\frac{d_x+1}{2}}+\exp(-T)+\sqrt{T}\sqrt{\mathbb{E}_{D_0,\dots,D_K}\left[\mathcal{R}_0(\widehat{s}_w)\right]}.\end{aligned}\tag{402}$$

Consequently, by (401),

$$\begin{aligned}&\mathbb{E}_{D_0,\dots,D_K}\mathbb{E}_{Z\sim p_0}\left[d_{\mathrm{TV}}\left(P_0(\cdot|Z),\widehat{P}_w(\cdot|z)\right)\right]\\&\lesssim \sqrt{t_0}\left(\log\frac{1}{t_0}\right)^{\frac{d_x+1}{2}}+\exp(-T)\\&\quad+\sqrt{T}\Bigg[\log\frac{1}{t_0}\,N_{\mathrm{eff}}^{-\frac{2\beta}{d_x+d_z+2\beta}}\,(\log N_{\mathrm{eff}})^{\max\{17,\beta\}}\\&\qquad+\frac{W_N^2}{(n_0+W_N)^2}\left[\overline{w}^\top G\overline{w}+\mathrm{diag}(G)^\top\overline{w}\right]\Bigg]^{1/2}.\end{aligned}\tag{403}$$

Corollary 1 (Target-domain total variation convergence rate)**.** *Suppose the conditions of Lemma 7 and Theorem 1 hold. Let*

$$T=\frac{2\beta}{d_x+d_z+2\beta}\log N_{\mathrm{eff}},\qquad t_0=N_{\mathrm{eff}}^{-\frac{4\beta}{d_x+d_z+2\beta}-1}.\tag{404}$$

Then

$$\log\frac{1}{t_0}\asymp\log N_{\mathrm{eff}},\qquad T\asymp\log N_{\mathrm{eff}}.\tag{405}$$

Moreover,

$$\exp(-T)=N_{\mathrm{eff}}^{-\frac{2\beta}{d_x+d_z+2\beta}},\tag{406}$$

$$\sqrt{t_0}\left(\log\frac{1}{t_0}\right)^{\frac{d_x+1}{2}}=o\left(N_{\mathrm{eff}}^{-\frac{\beta}{d_x+d_z+2\beta}}\right).\tag{407}$$

Combining Lemma 7 with (401), we therefore obtain

$$\begin{aligned}&\mathbb{E}_{D_0,\dots,D_K}\mathbb{E}_{Z\sim p_0}\left[d_{\mathrm{TV}}\left(P_0(\cdot|Z),\widehat{P}_w(\cdot|Z)\right)\right]\\&\lesssim N_{\mathrm{eff}}^{-\frac{\beta}{d_x+d_z+2\beta}}\,(\log N_{\mathrm{eff}})^{\max\left\{\frac{19}{2},\frac{\beta+2}{2}\right\}}\\&\qquad+\sqrt{\log N_{\mathrm{eff}}}\left[\frac{W_N^2}{(n_0+W_N)^2}\left[\overline{w}^\top G\overline{w}+\mathrm{diag}(G)^\top\overline{w}\right]\right]^{1/2}\\&\lesssim N_{\mathrm{eff}}^{-\frac{\beta}{d_x+d_z+2\beta}}\,(\log N_{\mathrm{eff}})^{\max\left\{\frac{19}{2},\frac{\beta+2}{2}\right\}}\\&\qquad+\sqrt{\log N_{\mathrm{eff}}}\cdot\frac{W_N}{n_0+W_N}\sqrt{\overline{w}^\top G\overline{w}+\mathrm{diag}(G)^\top\overline{w}}.\end{aligned}\tag{408}$$

Suppressing polylogarithmic factors in N_{eff} and absorbing the lower-order remainder term, this can be written as

$$\begin{aligned}&\mathbb{E}_{D_0,\dots,D_K}\mathbb{E}_{Z_0\sim P_0}\left[d_{\mathrm{TV}}\left(P_0(\cdot|Z_0),\widehat{P}_w(\cdot|Z_0)\right)\right]\\&=\widetilde{\mathcal{O}}\left(N_{\mathrm{eff}}^{-\frac{\beta}{d_x+d_z+2\beta}}+\frac{W_N}{n_0+W_N}\sqrt{\overline{w}^\top G\overline{w}+\mathrm{diag}(G)^\top\overline{w}}\right),\end{aligned}\tag{409}$$

which proves Theorem 2.

□

K.7 Type I error control of DA-CIT

Lemma 8 (Type I error control of CRT; Theorem 4 of Berrett et al. (2020))**.** *Suppose that the null hypothesis H_0 holds. Let $D_0^{\mathrm{CRT}} := \{(x_{0,i}, y_{0,i}, z_{0,i})\}_{i=1}^{n_0}$ denote the target samples used to conduct the CRT, and let $D_0^{\mathrm{Tr}} := \{(x_{0,i}, y_{0,i}, z_{0,i})\}_{i=n_0+1}^{2n_0}$ denote the target samples used to train the conditional diffusion model. Since the observations are independent, D_0^{CRT} and D_0^{Tr} are independent. Suppose that $\widehat{P}_0(\cdot|z)$ is estimated using D_0^{Tr} and, possibly, the source datasets $D_1, \ldots, D_K$, and is therefore independent of D_0^{CRT}. Then, for any $\alpha \in (0,1)$, the CRT p-value satisfies*

$$\mathbb{P}(p \le \alpha) \le \alpha + \mathcal{O}\left(n_0\, \mathbb{E}_{D_0^{\mathrm{Tr}}, D_1, \ldots, D_K} \mathbb{E}_{Z_0}\left[d_{\mathrm{TV}}\left(P_0(\cdot|Z_0), \widehat{P}_0(\cdot|Z_0)\right)\right]\right). \tag{410}$$

Because D_0^{CRT} and D_0^{Tr} are independent, and D_0^{CRT} is also independent of $\widehat{s}_w$, we have $\mathbb{E}_{D_0^{\mathrm{Tr}}, D_1, \ldots, D_K} = \mathbb{E}_{D_0, D_1, \ldots, D_K}$. By Lemma 8, the excess Type I error of DA-CIT is controlled by TV error bound of $\widehat{P}_w(\cdot|Z_0)$ which is estimated by DA-Diff. Substituting the TV convergence rate in (29) into (410) therefore yields the Type I error bound of DA-CIT in (31).

□

K.8 Estimation of source discrepancy matrix G

In this section, we show that the source discrepancy matrix G can be accurately estimated as mentioned in Remark 2. To this end, we assume that an independent holdout validation sample

$$D_0^{\mathrm{val}} := \left\{(x_{0,\ell}^{\mathrm{val}}, z_{0,\ell}^{\mathrm{val}})\right\}_{\ell=1}^{n_{\mathrm{val}}} \overset{\text{i.i.d.}}{\sim} P_0 \tag{411}$$

is available, which is independent of all samples used to train $\widehat{s}_w, \widehat{s}_1, \ldots, \widehat{s}_K$ and is used only for estimating G. Recall the definition $\Delta_{k,t} = s_k^* - s_0^*$ in (147) and the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ in (90). Then,

$$G_{ij} = \langle \Delta_{i,t}, \Delta_{j,t} \rangle_{\mathcal{H}_0}. \tag{412}$$

Define

$$\widehat{\Delta}_{k,t} := \widehat{s}_k - \widehat{s}_w, \qquad k = 1, \ldots, K. \tag{413}$$

We estimate G_{ij} using the holdout validation sample as

$$\widehat{G}_{ij} := \frac{1}{n_{\mathrm{val}}} \sum_{\ell=1}^{n_{\mathrm{val}}} \frac{1}{T - t_0} \int_{t_0}^{T} \mathbb{E}_{X_t \sim \mathcal{N}(a_t x_{0,\ell}^{\mathrm{val}}, \sigma_t^2 I_{d_x})} \Big[\big\langle \widehat{s}_w(X_t, z_{0,\ell}^{\mathrm{val}}, t) - \widehat{s}_i(X_t, z_{0,\ell}^{\mathrm{val}}, t), \\ \widehat{s}_w(X_t, z_{0,\ell}^{\mathrm{val}}, t) - \widehat{s}_j(X_t, z_{0,\ell}^{\mathrm{val}}, t) \big\rangle\Big] dt. \tag{414}$$

We first establish the convergence rates of the score estimators appearing in (414). To achieve that, we need an extra overlap assumption on individual densities.

Assumption 3 (Individual source overlap)**.** *For the source-specific estimators, we additionally assume that there exists a constant $C_{\mathrm{ov}} > 0$, independent of the sample sizes, such that*

$$p_0(z) \le C_{\mathrm{ov}} p_k(z), \qquad k = 1, \ldots, K, \quad p_0\text{-almost everywhere}. \tag{415}$$

Together with Assumption 2, similar to Appendix K.5.1, we can derive

$$\|g\|_{\mathcal{H}_0}^2 \lesssim \|g\|_{\mathcal{H}_k}^2, \qquad k = 1, \ldots, K. \tag{416}$$

Lemma 9 (Score estimation errors for estimating G)**.** *Suppose that Assumptions 1 and 2 hold, together with (415), and that the conditions of Theorem 1 are satisfied. For $k = 1, \ldots, K$, let $\widehat{s}_k$ be the conditional score estimator trained using D_k.*

Define

$$r_w := \mathbb{E}_{D_0, \ldots, D_K}\left[\|\widehat{s}_w - s_0^*\|_{\mathcal{H}_0}^2\right], \qquad r_k := \mathbb{E}_{D_k}\left[\|\widehat{s}_k - s_k^*\|_{\mathcal{H}_0}^2\right], \tag{417}$$

and

$$\varepsilon_G^2 := r_w + \max_{1\le k\le K} r_k. \tag{418}$$

Let

$$n_{\min} := \min_{1\le k\le K} n_k.$$

Then,

$$\varepsilon_G^2 \lesssim \widetilde{\mathcal{O}}\left(N_{\text{eff}}^{-\frac{2\beta}{d_x+d_z+2\beta}}\right) + \frac{W_N^2}{(n_0+W_N)^2} n_0^{-2\gamma} + \log\frac{1}{t_0}\, n_{\min}^{-\frac{2\beta}{d_x+d_z+2\beta}} (\log n_{\min})^{\max\{17,\beta\}}. \tag{419}$$

Under the feasible regime (32), the first two terms on the right-hand side of (419) converge to zero. Moreover, since we choose $t_0 = N^{-C_\sigma}$ and $N \asymp N_{\text{eff}}^{(d_x+d_z)/(d_x+d_z+2\beta)}$ in (258) and (395), the last term converges to zero under the additional growth condition

$$N_{\text{eff}} \ll \exp\left(n_{\min}^{c_s}\right), \qquad 0 < c_s < 2\Gamma, \tag{420}$$

where $\Gamma = \beta/(d_x+d_z+2\beta)$ is defined in Section 3.2 and c_s is a constant. Therefore, $\varepsilon_G = o(1)$ under (32) and (420) when $n_0, n_{\min} \to \infty$.

Proof. First, Theorem 1 gives

$$r_w \lesssim \widetilde{\mathcal{O}}\left(N_{\text{eff}}^{-\frac{2\beta}{d_x+d_z+2\beta}}\right) + \frac{W_N^2}{(n_0+W_N)^2}\left[\overline{w}^\top G \overline{w} + \operatorname{diag}(G)^\top \overline{w}\right]. \tag{421}$$

By Assumption 1,

$$G_{kk} = \|\Delta_{k,t}\|_{\mathcal{H}_0}^2 = \mathcal{O}(n_0^{-2\gamma}), \qquad k = 1,\dots,K. \tag{422}$$

Since G is a Gram matrix, the Cauchy–Schwarz inequality gives

$$|G_{ij}| \le \sqrt{G_{ii}G_{jj}} = \mathcal{O}(n_0^{-2\gamma}).$$

Moreover, since $\overline{w}$ belongs to the probability simplex,

$$\overline{w}^\top G \overline{w} \le \sum_{i=1}^K \sum_{j=1}^K \overline{w}_i \overline{w}_j |G_{ij}| = \mathcal{O}(n_0^{-2\gamma}), \tag{423}$$

$$\operatorname{diag}(G)^\top \overline{w} = \sum_{k=1}^K \overline{w}_k G_{kk} = \mathcal{O}(n_0^{-2\gamma}). \tag{424}$$

Substituting these bounds into (421) yields

$$r_w \lesssim \widetilde{\mathcal{O}}\left(N_{\text{eff}}^{-\frac{2\beta}{d_x+d_z+2\beta}}\right) + \frac{W_N^2}{(n_0+W_N)^2} n_0^{-2\gamma}. \tag{425}$$

We next consider the source-domain score estimators. Lemma D.7 of Fu et al. (2024) shows that, under Assumption 2, the conditional score estimator trained using n_k observations satisfies

$$\mathbb{E}_{D_k}\left[\|\widehat{s}_k - s_k^*\|_{\mathcal{H}_k}^2\right] \lesssim \log\frac{1}{t_0}\, n_k^{-\frac{2\beta}{d_x+d_z+2\beta}} (\log n_k)^{\max\{17,\beta\}}. \tag{426}$$

By (416),

$$\begin{aligned} r_k &= \mathbb{E}_{D_k}\left[\|\widehat{s}_k - s_k^*\|_{\mathcal{H}_0}^2\right] \\ &\lesssim \mathbb{E}_{D_k}\left[\|\widehat{s}_k - s_k^*\|_{\mathcal{H}_k}^2\right] \\ &\lesssim \log\frac{1}{t_0}\, n_k^{-\frac{2\beta}{d_x+d_z+2\beta}} (\log n_k)^{\max\{17,\beta\}}. \end{aligned} \tag{427}$$

Taking the maximum over $k = 1,\dots,K$ and combining (425) and (427) gives (419).

□

We next separate the score estimation error from the empirical approximation error. Define the population plug-in counterpart of (414) as

$$\widetilde{G}_{ij} := \left\langle \widehat{\Delta}_{i,t}, \widehat{\Delta}_{j,t} \right\rangle_{\mathcal{H}_0}, \qquad i,j = 1,\dots,K. \tag{428}$$

Lemma 10 (Error of the population plug-in estimator)**.** *Under the conditions of Lemma 9, uniformly over $i,j = 1,\dots,K$,*

$$\mathbb{E}_{D_0,\dots,D_K}\left[\left|\widetilde{G}_{ij} - G_{ij}\right|\right] \lesssim n_0^{-2\gamma} + \varepsilon_G^2. \tag{429}$$

Under the feasible growth regime (32) and (420), the right-hand side converges to zero.

Proof. Let

$$e_0 := \widehat{s}_w - s_0^*, \qquad e_k := \widehat{s}_k - s_k^*, \qquad u_k := e_k - e_0.$$

By the definitions of $\Delta_{k,t}$ and $\widehat{\Delta}_{k,t}$ in (147) and (413),

$$\begin{aligned} \widehat{\Delta}_{k,t} &= \widehat{s}_k - \widehat{s}_w \\ &= (s_k^* + e_k) - (s_0^* + e_0) \\ &= \Delta_{k,t} + u_k. \end{aligned} \tag{430}$$

Therefore,

$$\begin{aligned} \widetilde{G}_{ij} - G_{ij} &= \langle \Delta_{i,t} + u_i, \Delta_{j,t} + u_j \rangle_{\mathcal{H}_0} - \langle \Delta_{i,t}, \Delta_{j,t} \rangle_{\mathcal{H}_0} \\ &= \langle u_i, \Delta_{j,t} \rangle_{\mathcal{H}_0} + \langle \Delta_{i,t}, u_j \rangle_{\mathcal{H}_0} + \langle u_i, u_j \rangle_{\mathcal{H}_0}. \end{aligned} \tag{431}$$

By Assumption 1,

$$\|\Delta_{k,t}\|_{\mathcal{H}_0} = \sqrt{G_{kk}} = \mathcal{O}(n_0^{-\gamma}). \tag{432}$$

Moreover,

$$\begin{aligned} &\mathbb{E}_{D_0,\dots,D_K}\left[\|u_k\|_{\mathcal{H}_0}^2\right] \\ &\quad = \mathbb{E}_{D_0,\dots,D_K}\left[\|e_k - e_0\|_{\mathcal{H}_0}^2\right] \\ &\quad \le 2\mathbb{E}_{D_0,\dots,D_K}\left[\|e_0\|_{\mathcal{H}_0}^2\right] + 2\mathbb{E}_{D_0,\dots,D_K}\left[\|e_k\|_{\mathcal{H}_0}^2\right] \\ &\quad = 2r_w + 2r_k \lesssim \varepsilon_G^2. \end{aligned} \tag{433}$$

Applying the Cauchy–Schwarz inequality to the first term in (431) gives

$$\begin{aligned} &\mathbb{E}_{D_0,\dots,D_K}\left[\left|\langle u_i, \Delta_{j,t} \rangle_{\mathcal{H}_0}\right|\right] \\ &\quad \le \|\Delta_{j,t}\|_{\mathcal{H}_0} \left(\mathbb{E}_{D_0,\dots,D_K}\left[\|u_i\|_{\mathcal{H}_0}^2\right]\right)^{1/2} \\ &\quad \lesssim n_0^{-\gamma}\varepsilon_G \\ &\quad \lesssim n_0^{-2\gamma} + \varepsilon_G^2 \end{aligned} \tag{434}$$

Similarly,

$$\mathbb{E}_{D_0,\dots,D_K}\left[\left|\langle \Delta_{i,t}, u_j \rangle_{\mathcal{H}_0}\right|\right] \lesssim n_0^{-\gamma}\varepsilon_G \lesssim n_0^{-2\gamma} + \varepsilon_G^2. \tag{435}$$

For the last term, another application of the Cauchy–Schwarz inequality yields

$$\begin{aligned} &\mathbb{E}_{D_0,\dots,D_K}\left[\left|\langle u_i, u_j \rangle_{\mathcal{H}_0}\right|\right] \\ &\quad \le \left(\mathbb{E}_{D_0,\dots,D_K}\left[\|u_i\|_{\mathcal{H}_0}^2\right]\right)^{1/2} \left(\mathbb{E}_{D_0,\dots,D_K}\left[\|u_j\|_{\mathcal{H}_0}^2\right]\right)^{1/2} \\ &\quad \lesssim \varepsilon_G^2. \end{aligned} \tag{436}$$

Combining (434)–(436) proves (429). Under the feasible growth regime (32), $\varepsilon_G \to 0$, and hence the right-hand side of (429) converges to zero.

□

It remains to control the empirical approximation error $\widehat{G}_{ij} - \widetilde{G}_{ij}$. For fixed trained score networks, define

$$H_{ij}(x,z) := \frac{1}{T-t_0}\int_{t_0}^{T} \mathbb{E}_{X_t\sim\mathcal{N}(a_t x,\sigma_t^2 I_{d_x})}\Big[\big\langle \widehat{\Delta}_{i,t}(X_t,z), \widehat{\Delta}_{j,t}(X_t,z)\big\rangle\Big]dt. \tag{437}$$

Recall (413) that $\widehat{\Delta}_{k,t} = \widehat{s}_k - \widehat{s}_w$. Then,

$$\widehat{G}_{ij} = \frac{1}{n_{\mathrm{val}}}\sum_{\ell=1}^{n_{\mathrm{val}}} H_{ij}\left(x_{0,\ell}^{\mathrm{val}}, z_{0,\ell}^{\mathrm{val}}\right), \qquad \widetilde{G}_{ij} = \mathbb{E}_{(X,Z)\sim P_0}\left[H_{ij}(X,Z)\right]. \tag{438}$$

Lemma 11 (Empirical approximation error of $\widehat{G}$)**.** *Suppose that the holdout sample D_0^{val} is independent of $D_0,\dots,D_K$, and that $\widehat{s}_w, \widehat{s}_1,\dots,\widehat{s}_K$ belong to the network class (83). Define*

$$M_G := \frac{4d_x}{T-t_0}\int_{t_0}^{T} M_t^2\, dt. \tag{439}$$

Then, uniformly over $i,j = 1,\dots,K$,

$$\mathbb{E}_{D_0,\dots,D_K,D_0^{\mathrm{val}}}\left[\left|\widehat{G}_{ij} - \widetilde{G}_{ij}\right|\right] \lesssim \frac{M_G}{\sqrt{n_{\mathrm{val}}}}. \tag{440}$$

Under the network configuration used in Lemma 3 Proposition C.4 of (Fu et al., 2024), and by the choose of T in Theorem 8, we have

$$M_G = \mathcal{O}(\log N). \tag{441}$$

Consequently,

$$\mathbb{E}_{D_0,\dots,D_K,D_0^{\mathrm{val}}}\left[\left|\widehat{G}_{ij} - \widetilde{G}_{ij}\right|\right] \lesssim \frac{\log N}{\sqrt{n_{\mathrm{val}}}}. \tag{442}$$

Proof. Condition on the training datasets $D_0,\dots,D_K$. Then the trained score networks, and hence H_{ij}, are fixed. Since the holdout observations are i.i.d. from P_0 and independent of $D_0,\dots,D_K$, we have

$$\mathbb{E}_{D_0^{\mathrm{val}}}\left[\widehat{G}_{ij}\,\middle|\, D_0,\dots,D_K\right] = \widetilde{G}_{ij}. \tag{443}$$

By the definition of the network class (83),

$$\sup_{x,z}\|\widehat{s}_k(x,z,t)\|_\infty \le M_t, \qquad k=1,\dots,K,$$

and the same bound holds for $\widehat{s}_w$. Therefore,

$$\|\widehat{\Delta}_{k,t}(x,z)\|_2 = \|\widehat{s}_k(x,z,t) - \widehat{s}_w(x,z,t)\|_2 \le 2\sqrt{d_x}M_t. \tag{444}$$

It follows from the Cauchy–Schwarz inequality that

$$\begin{aligned}|H_{ij}(x,z)| &\le \frac{1}{T-t_0}\int_{t_0}^{T}\mathbb{E}\left[\|\widehat{\Delta}_{i,t}(X_t,z)\|_2\|\widehat{\Delta}_{j,t}(X_t,z)\|_2\right]dt\\ &\le \frac{4d_x}{T-t_0}\int_{t_0}^{T} M_t^2\, dt = M_G.\end{aligned} \tag{445}$$

Thus, conditionally on $D_0,\dots,D_K$,

$$\mathrm{Var}_{(X,Z)\sim P_0}\left[H_{ij}(X,Z)\,|\,D_0,\dots,D_K\right] \le M_G^2. \tag{446}$$

Since $\widehat{G}_{ij}$ is the empirical average of n_{val} independent copies of $H_{ij}(X,Z)$,

$$\begin{aligned}\mathbb{E}_{D_0^{\mathrm{val}}}\left[(\widehat{G}_{ij} - \widetilde{G}_{ij})^2\,\middle|\, D_0,\dots,D_K\right] &= \mathrm{Var}\left(\widehat{G}_{ij}\,\middle|\, D_0,\dots,D_K\right)\\ &\le \frac{M_G^2}{n_{\mathrm{val}}}.\end{aligned} \tag{447}$$

Hence, by the Cauchy–Schwarz inequality,

$$\mathbb{E}_{D_0^{\mathrm{val}}}\left[\left|\widehat{G}_{ij}-\widetilde{G}_{ij}\right| \,\middle|\, D_0,\dots,D_K\right] \le \left\{\mathbb{E}_{D_0^{\mathrm{val}}}\left[(\widehat{G}_{ij}-\widetilde{G}_{ij})^2 \,\middle|\, D_0,\dots,D_K\right]\right\}^{1/2} \le \frac{M_G}{\sqrt{n_{\mathrm{val}}}}. \tag{448}$$

Taking expectation over $D_0,\dots,D_K$ proves (440). Finally, by (194),

$$M_\ell = \int_{t_0}^{T} \frac{M_t^2}{\log N}\,dt = \mathcal{O}(\log N).$$

Therefore,

$$M_G = \frac{4d_x}{T-t_0}\int_{t_0}^{T} M_t^2\,dt = \frac{4d_x \log N}{T-t_0} M_\ell. \tag{449}$$

Since $T - t_0 \asymp \log N$ shown in (263), we obtain

$$M_G = \mathcal{O}(\log N), \tag{450}$$

which completes the proof. □

Combining the population plug-in error with the empirical approximation error yields the following estimation bound for G.

Theorem 9 (Estimation error of the source discrepancy matrix)**.** *Under the conditions of Lemmas 9, 10, and 11, uniformly over $i, j = 1,\dots,K$,*

$$\mathbb{E}_{D_0,\dots,D_K,D_0^{\mathrm{val}}}\left[\left|\widehat{G}_{ij}-G_{ij}\right|\right] \lesssim n_0^{-2\gamma} + \varepsilon_G^2 + \frac{\log N}{\sqrt{n_{\mathrm{val}}}}. \tag{451}$$

Furthermore,under the feasible regime (32) and (420), provided that

$$\frac{\log N}{\sqrt{n_{\mathrm{val}}}} \asymp \frac{\log N_{\mathrm{eff}}}{\sqrt{n_{\mathrm{val}}}} \to 0, \tag{452}$$

we have

$$\mathbb{E}_{D_0,\dots,D_K,D_0^{\mathrm{val}}}\left[\left|\widehat{G}_{ij}-G_{ij}\right|\right] \to 0. \tag{453}$$

Proof. By the triangle inequality,

$$\mathbb{E}_{D_0,\dots,D_K,D_0^{\mathrm{val}}}\left[\left|\widehat{G}_{ij}-G_{ij}\right|\right] \le \mathbb{E}_{D_0,\dots,D_K,D_0^{\mathrm{val}}}\left[\left|\widehat{G}_{ij}-\widetilde{G}_{ij}\right|\right] + \mathbb{E}_{D_0,\dots,D_K}\left[\left|\widetilde{G}_{ij}-G_{ij}\right|\right]. \tag{454}$$

Applying Lemmas 11 and 10 gives

$$\mathbb{E}_{D_0,\dots,D_K,D_0^{\mathrm{val}}}\left[\left|\widehat{G}_{ij}-G_{ij}\right|\right] \lesssim n_0^{-2\gamma} + \varepsilon_G^2 + \frac{\log N}{\sqrt{n_{\mathrm{val}}}}, \tag{455}$$

which proves (451).

By Lemma 9, $\varepsilon_G^2 \to 0$ under the feasible regime (32) and (420), together with (452). Since $n_0^{-2\gamma} \to 0$, we further have

$$n_0^{-2\gamma} + \varepsilon_G^2 \to 0. \quad \text{and} \quad \frac{\log N}{\sqrt{n_{\mathrm{val}}}} \asymp \frac{\log N}{\sqrt{n_{\mathrm{val}}}} \to 0$$

Therefore, every term on the right-hand side of (451) converges to zero, which proves (453). We can further take $N \asymp N_{\mathrm{eff}}^{\frac{d_x+d_z}{d_x+d_z+2\beta}}$ as specified in (395). Finally, we have

$$\mathbb{E}_{D_0,\dots,D_K,D_0^{\mathrm{val}}}\left[\left|\widehat{G}_{ij}-G_{ij}\right|\right] \lesssim n_0^{-2\gamma} + \varepsilon_G^2 + \frac{\log N_{\mathrm{eff}}}{\sqrt{n_{\mathrm{val}}}} \to 0. \tag{456}$$

□