# FADE: Towards Fairness-aware Data Generation for Domain Generalization via Classifier-Guided Score-based Diffusion Models

**Yujie Lin**[1,2] , **Dong Li**[3] , **Minglai Shao**[1] , **Guihong Wan**[4] , **Chen Zhao**[3]
[1]School of New Media and Communication, Tianjin University
[2]School of Informatics, Xiamen University
[3]Department of Computer Science, Baylor University
[4]Departments of Biostatistics and Epidemiology, Harvard University
linyujie001213@gmail.com, shaoml@tju.edu.cn, {dong_li1, chen_zhao}@baylor.edu,
guihong_wan@hsph.harvard.edu

## Abstract

Fairness-aware domain generalization (FairDG) has emerged as a critical challenge for deploying trustworthy AI systems, particularly in scenarios involving distribution shifts. Traditional methods for addressing fairness have failed in domain generalization due to their lack of consideration for distribution shifts. Although disentanglement has been used to tackle FairDG, it is limited by its strong assumptions. To overcome these limitations, we propose Fairness-aware Classifier-Guided Score-based Diffusion Models (FADE) as a novel approach to effectively address the FairDG issue. Specifically, we first pre-train a score-based diffusion model (SDM) and two classifiers to equip the model with strong generalization capabilities across different domains. Then, we guide the SDM using these pre-trained classifiers to effectively eliminate sensitive information from the generated data. Finally, the generated fair data is used to train downstream classifiers, ensuring robust performance under new data distributions. Extensive experiments on three real-world datasets demonstrate that FADE not only enhances fairness but also improves accuracy in the presence of distribution shifts. Additionally, FADE outperforms existing methods in achieving the best accuracy-fairness trade-offs.

## 1 Introduction

In recent years, fairness in machine learning has received widespread attention due to its significance in real-world systems such as loan approvals [Purificato *et al.*, 2023] and employee hiring [Alder and Gilbert, 2006]. From a group perspective, algorithmic fairness focuses on the statistical parity among different groups defined by specific sensitive attributes of people [Hardt *et al.*, 2016]. Many efforts have been made to address algorithmic fairness [Mitchell *et al.*, 2021]. A type of method achieves fairness by eliminating bias from the dataset through fair data generation. As a pioneer in this approach, FGAN [Xu *et al.*, 2018] removes sensitive information by ensuring that the discriminator cannot distinguish the sensitive group membership of the generated samples. Due to the powerful generative capabilities, diffusion models have also been employed to ensure the fairness of generated data. FLDGM [Ramachandranpillai *et al.*, 2023] integrate an existing debiasing method with diffusion models to generate unbiased data. Although they have achieved remarkable results, their poor generalization performance has caused them to fail when facing distribution shifts [Shao *et al.*, 2024].

Fairness-aware domain generalization (FairDG), as an emerging topic, has gained increasing attention because it is essential for advancing the deployment of trustworthy AI in the real world [Shao *et al.*, 2024]. Instead of the independent and identically distributed (*i.i.d.*) assumption, this field focuses on fairness under distribution shifts. The most widely studied distribution shift is covariate shift [Shimodaira, 2000], which variations to changes caused by differences in the marginal distributions of instances. One class of methods addresses the FairDG issue by learning fair and domain-invariant representations through feature disentanglement [Oh *et al.*, 2022]. However, these methods often have overly strong assumptions that it can perfectly disentangle sensitive information and domain-specific information from the feature representation. The assumptions limit their applicability and reduce their robustness. Another class of methods enhances the diversity of the training set through data augmentation [Pham *et al.*, 2023], enabling the model to generalize better to unseen target domains and thereby improving overall performance.

In this paper, inspired by data augmentation strategies, we propose a three-stage framework called FADE: Fairness-aware Classifier-Guided Score-based Diffusion Models to generate unbiased and domain-invariant data. Specifically, we pre-train the parameters of the score-based diffusion model (SDM), label classifier, and sensitive classifier using a source dataset across multiple domains in the first stage. This stage equips the three components with generalization capabilities in the target domain. In the second stage, we derive a generator by guiding the SDM with the pre-trained classifiers, which effectively removes bias or sensitive information from the generated data. Simultaneously, during the generation phase, we can obtain fair data without requiring any additional training. Finally, we use the generated data, which has reduced sensitive information, to train a fairness-aware domain-invariant downstream classifier. Our contributions can be summarized as follows:

- We formulate a novel problem: generating unbiased data to train downstream classifiers that are tested on distribution-shifted datasets, while ensuring both accuracy and fairness.
- We have designed a novel fair data generation method called FADE, which ensures the generalization ability of the generative model during the pre-training phase and removes biased information during the generation phase. FADE not only allows for the specification of generated sample categories but also possesses generalization capabilities under new data distribution shifts.
- Experimental results on three real-world datasets demonstrate that FADE achieves the best performance in both fairness and accuracy compared to other baselines when facing the challenge of distribution shifts.

## 2 Related Work

**Fair Data Generation.** Fairness in machine learning aims to ensure equitable performance across different demographic groups, and it can be achieved through three primary approaches: pre-processing, in-processing, and post-processing methods. Pre-processing methods modify the training data to mitigate biases before training the model, using techniques such as data resampling, data transformation, and fair data generation. Fair data generation is similar to pre-processing methods. However, unlike pre-processing methods, fair data generation does not use the original data to train downstream classifiers. Instead, it can generate additional data for training predictive models, which is especially beneficial when the original training data is very limited. FGAN [Xu *et al.*, 2018] is the first method to tackle fair data generation. Also based on GANs [Goodfellow *et al.*, 2020], DECAF [Van Breugel *et al.*, 2021] can achieve various fairness criteria by leveraging causal graphs. FLDGM [Ramachandranpillai *et al.*, 2023] attempts to integrate an existing debiasing method [Liu *et al.*, 2022] with GANs or diffusion models [Ho *et al.*, 2020] to achieve fair data generation.

**Fairness under Distribution Shifts.** Achieving fairness is not devoid of challenges, especially in the presence of distribution shifts. These shifts can pose significant hurdles as models trained on source distributions may not generalize well to target data distributions, potentially exacerbating biases and undermining the intended fairness objectives [Lin *et al.*, 2024; Komanduri *et al.*, 2024; Lin *et al.*, 2023a; Lin *et al.*, 2023b; Zhao and Chen, 2019]. There are two primary approaches to addressing fairness issues across domains: feature disentanglement and data augmentation. Feature disentanglement aims to learn latent representations of data features, enhancing their clarity and mutual independence within the model [Zhao *et al.*, 2024; Lian *et al.*, 2025; Jiang *et al.*, 2024; Li *et al.*, 2024; Zhao *et al.*, 2023]. However, these methods frequently rely on overly strong assumptions, which restrict their applicability and diminish their robustness. Additionally, assessing the quality of disentanglement presents its own challenges. Data augmentation seeks to enhance the diversity of training datasets and improve model generalization performance by systematically applying controlled transformations to the training data [Pham *et al.*, 2023; Zhao *et al.*, 2024].

## 3 Background

**Notations.** Let $\mathcal{X} \subseteq \mathbb{R}^p$ denotes a feature space and $\mathcal{Z} = \{0, 1\}$ is a sensitive space. $\mathcal{Y} = \{0, 1\}$ is defined as an output or a label space. A domain $e \in \mathcal{E}$ is defined as a joint distribution $\mathbb{P}^e_{XZY} := \mathbb{P}(X^e, Z^e, Y^e)$ on $\mathcal{H} = \mathcal{X} \times \mathcal{Z} \times \mathcal{Y}$. A dataset *i.i.d.* sampled from a domain $e$ is represented as $\mathcal{D}^e = \{(\mathbf{x}^e_j, z^e_j, y^e_j)\}_{j=1}^{|\mathcal{D}^e|}$, where $\mathbf{x}, z, y$ are the realizations of random variables $X, Z, Y$ in the corresponding spaces. A function $f$ parameterized by $\boldsymbol{\theta}$ is denoted as $f_{\boldsymbol{\theta}}$. We defer a list of the notations used in this paper in Appendix A.

**Fairness-aware Domain Generalization (FairDG).** In the supervised learning setting of FairDG [Shao *et al.*, 2024], the goal is to lean a classifier $f_{\boldsymbol{\theta}} : \mathcal{X} \rightarrow \mathcal{Y}$ across multiple source domains $\mathcal{E}^s = \{e_i\}_{i=1}^S$, where $S = |\mathcal{E}^s|$, such that the learned $f_{\boldsymbol{\theta}}$ can be applied on a distinct shifted target domain $e_t \notin \mathcal{E}^s$, which is unknown and inaccessible during training, achieving both good generalization accuracy and fair decision-making.

To ensure fair predictions, the classifier $f_{\boldsymbol{\theta}}$ is required to guarantee that the outcomes $\hat{Y} = f_{\boldsymbol{\theta}}(X)$ are not biased or discriminatory against any specific protected groups, each characterized by sensitive attributes $Z$, such as race and gender. Specifically, a fairness-aware metric $\omega_{\boldsymbol{\theta}} : \mathcal{Z} \times \mathcal{Y} \rightarrow \mathbb{R}$, such as the difference of demographic parity [Dwork *et al.*, 2012] or the difference of equalized odds [Hardt *et al.*, 2016], is used to control the fair dependence between the sensitive attribute and model outcomes during training. A value of $\omega_{\boldsymbol{\theta}}(Z, \hat{Y})$ closer to 0 indicates greater fairness.

**Problem 1** (FairDG)**.** *Given a source dataset* $\mathcal{D}^{src} = \{\mathcal{D}^{e_i}\}_{i=1}^S$*, where each* $\mathcal{D}^{e_i} = \{(\mathbf{x}^{e_i}_j, z^{e_i}_j, y^{e_i}_j)\}_{j=1}^{|\mathcal{D}^{e_i}|}$ *is i.i.d. sampled from a unique source domain* $e_i \in \mathcal{E}^s$*, a loss function* $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$*, and a fair metric* $\omega_{\boldsymbol{\theta}} : \mathcal{Z} \times \mathcal{Y} \rightarrow \mathbb{R}$*, the goal is to learn a fair classifier* $f_{\boldsymbol{\theta}} : \mathcal{X} \rightarrow \mathcal{Y}$ *that minimized the risk over* $\mathcal{D}^{src}$ *satisfying a fairness constraint:*

$$\begin{aligned} \min_{\boldsymbol{\theta}} \max_{e_i \in \mathcal{E}^s} \quad & \mathbb{E}_{(\mathbf{x}^{e_i}, z^{e_i}, y^{e_i}) \in \mathcal{D}^{src}, \forall e_i \in \mathcal{E}^s} [\ell(f_{\boldsymbol{\theta}}(\mathbf{x}^{e_i}), y^{e_i})] \\ \textit{s.t.} \quad & \omega_{\boldsymbol{\theta}}\left(z^{e_i}, f_{\boldsymbol{\theta}}(\mathbf{x}^{e_i})\right) \leq \epsilon \end{aligned}$$

*where* $\epsilon > 0$ *is an empirical threshold. The learned* $f_{\boldsymbol{\theta}}$ *is required to be generalizable to a target domain* $e_t \notin \mathcal{E}^s$*, which is unknown and inaccessible during training, such that it performs well in terms of predicted accuracy and fairness on* $\mathcal{D}^{tgt} = \{(\mathbf{x}^{e_t}_j, z^{e_t}_j, y^{e_t}_j)\}_{j=1}^{|\mathcal{D}^{tgt}|}$ *sampled from this domain.*

A key challenge in addressing Problem 1 is determining how closely the data distributions in unknown target domains align with those in the observed source domains. Given the various distribution shifts discussed in [Shao *et al.*, 2024], this paper narrows the scope by focusing specifically on the shift between source and target domains that is solely due to covariate shift [Shimodaira, 2000], in which domain variation

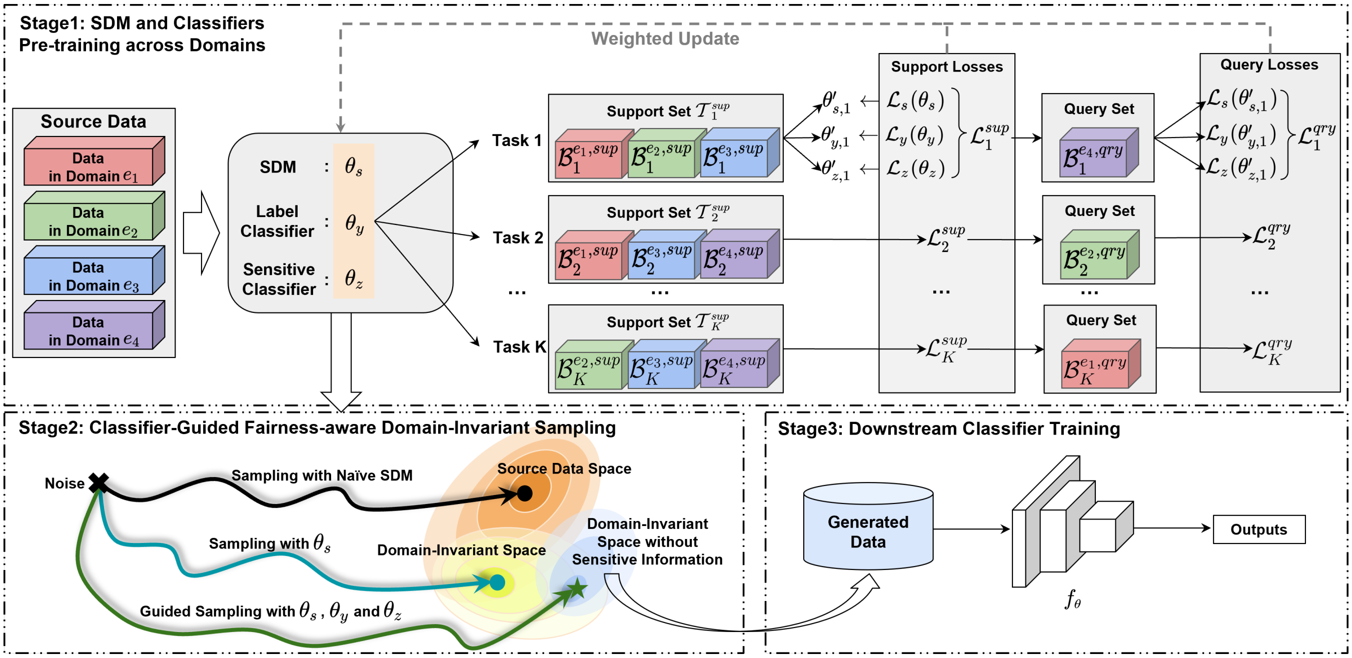



Figure 1: An overview of FADE. Supposing there are 4 domains in the training set $\mathcal{D}^{src}$. For the $k$-th iteration, we randomly select a batch as $\mathcal{T}_k^{sup}$, and sample from the remaining three domains to compose $\mathcal{T}_k^{qry}$. We perform gradient descent using the loss $\mathcal{L}_k^{sup}(\Phi)$ obtained on $\mathcal{T}_k^{sup}$ to obtain a temporary set of parameters $\Phi'$, then use the model with $\Phi'$ to obtain loss $\mathcal{L}_k^{qry}(\Phi_k')$ values on $\mathcal{T}_k^{qry}$. We use the weighted sum of two losses to update $\Phi = \{\boldsymbol{\theta}_s, \boldsymbol{\theta}_y, \boldsymbol{\theta}_z\}$. When using the score network $s_{\boldsymbol{\theta}_s}$ to generate data from noise, we can guide the generation process with $\phi_{\boldsymbol{\theta}_y}$ and $\psi_{\boldsymbol{\theta}_z}$ to obtain unbiased data. Finally, these data are used to train downstream tasks to achieve the final results.

is attributed to disparities in the marginal distributions over input features $\mathbb{P}_X^{e_i} \neq \mathbb{P}_X^{e_t}, \forall e_i \in \mathcal{E}^s$.

Inspired by existing domain generalization efforts [Zhou *et al.*, 2022], data augmentation strategies, such as diffusion models [Ho *et al.*, 2020], applied to source domains enhance the diversity of the training data, thereby improving the ability of $f_{\boldsymbol{\theta}}$ to generalize to unseen target domains and ensuring domain invariance.

**Score-based Diffusion Models (SDMs).** Denoising diffusion probabilistic models (DDPMs) [Ho *et al.*, 2020] have demonstrated remarkable performance in generating high-fidelity data. The forward process in DDPMs gradually adds Gaussian noise to the data over a series of timesteps, eventually transforming the data into pure noise. Let $\mathbf{x}_0$ be the original data, and $\mathbf{x}_{i>0}, i = 1, \cdots N$ be the noisy data after $i$ steps. The forward process is defined as:

$$\mathbf{x}_i = \sqrt{1-\beta_i}\mathbf{x}_{i-1} + \sqrt{\beta_i}\mathbf{u}_{i-1}, \quad i = 1, \cdots, N \tag{1}$$

where noise $\mathbf{u}_{i-1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\beta_i$ is a hyperparameter to control the intensity of noise addition and $N$ is the total number of noise addition times.

Score-based diffusion models [Song *et al.*, 2020] are an alternative implementation of DDPMs. In the limit of $N \to \infty$, we define an auxiliary set of noise scales $\{\beta_i = N\beta_i\}_{i=1}^N$. Let $\beta\left(\frac{i}{N}\right) = \beta_i$, $\{\beta_i\}_{i=1}^N$ becomes a function $\beta(t)$ indexed by $t$. Given $t \in [0, T]$, $\mathbf{x}_t$ represents $\mathbf{x}$ at moment $t$. The forward diffusion in Eq.(1) can be defined by a stochastic differential equation (SDE) [Ho *et al.*, 2020]:

$$\mathrm{d}\mathbf{x}_t = -\frac{1}{2}\beta(t)\mathbf{x}_t \mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}\mathbf{w}, \tag{2}$$

where the function $\beta(t)$ is determined by the discrete hyper parameter $\beta_t$ and $\mathbf{w}$ is the standard Wiener process [Zhang *et al.*, 2018]. Defining $h(\mathbf{x}_t) = -\frac{1}{2}\beta(t)$ and $g(t) = \sqrt{\beta(t)}$. When generating samples, the corresponding reverse diffusion process can be described by the following system of SDEs:

$$\mathrm{d}\mathbf{x}_t = \left[h(\mathbf{x}_t) - g^2(t)\nabla_{\mathbf{x}_t}\log p_t(\mathbf{x}_t)\right]\mathrm{d}\bar{t} + g(t)\mathrm{d}\overline{\mathbf{w}}, \tag{3}$$

where $\mathbf{w}$ is the reverse-time standard Wiener processes, and $\mathrm{d}\bar{t}$ is an infinitesimal negative time step. The score networks $s_{\boldsymbol{\theta}_s}$ is trained to approximate the partial score functions $\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t)$ and used to generate the sample features by evolving backward in time. Specially, the score loss function [Song *et al.*, 2020] $\mathcal{L}_s : \Theta_s \times \mathcal{H} \to \mathbb{R}$ can be formulate:

$$\begin{aligned}\mathcal{L}_s(\boldsymbol{\theta}_s, \mathcal{D}) =& \mathbb{E}_t\{\lambda(t)\mathbb{E}_{\mathbf{x}_t(0)}\mathbb{E}_{\mathbf{x}_t(t)|\mathbf{x}_t(0)}[\|s_{\boldsymbol{\theta}_s}(\mathbf{x}_t(t), t) \\ &- \nabla_{\mathbf{x}_t(t)}\log p_{0t}(\mathbf{x}_t(t)|\mathbf{x}_t(0))\|_2^2]\},\end{aligned} \tag{4}$$

where $\lambda : [0, T] \to \mathbb{R}_{>0}$ is a positive weighting function, $t$ is uniformly sampled over $[0, T]$, $\mathbf{x}_0 \sim p(\mathbf{x})$ and $\mathbf{x}_t \sim p(\mathbf{x}_t|\mathbf{x}_0)$. For DDPM, we can typically choose

$$\lambda \propto 1/\mathbb{E}\left[\|\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t|\mathbf{x}_0\|_2^2\right]. \tag{5}$$

**Problem Setting.** To address the problem of FairDG, given data $\mathcal{D}^{src} = \{\mathcal{D}^{e_i}\}_{i=1}^S$, where each $\mathcal{D}^{e_i}$ is sampled from a unique source domain $e_i \in \mathcal{E}^s$, the key of this paper is to seek a generator $G : \mathcal{X} \times \mathcal{Z} \times \mathcal{Y} \to \mathcal{X} \times \mathcal{Z} \times \mathcal{Y}$, initialized by a pre-trained SDM. A dataset generated by $G$, denoted $\mathcal{D}^{gen} = \{\mathbf{x}_j', z_j', y_j'\}_{j=1}^{|\mathcal{D}^{gen}|}$, can then be used to learn a downstream classifier $f_{\boldsymbol{\theta}}$ by minimizing the risk and satisfying algorithmic fairness, $\omega_{\boldsymbol{\theta}}(Z', f_{\boldsymbol{\theta}}(X')) \leq \epsilon$, over $\mathcal{D}^{gen}$.

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(\mathbf{x}^{e_i}, z^{e_i}, y^{e_i}) \in \mathcal{D}^{src}, \forall e_i \in \mathcal{E}^s}[\ell(f_{\boldsymbol{\theta}}(\mathbf{x}'), y')], \tag{6}$$

where the sample $(\mathbf{x}', z', y') \in \mathcal{D}^{gen}$ is generated by $G$ using $(\mathbf{x}^{e_i}, z^{e_i}, y^{e_i})$. During inference, the learned $f_{\boldsymbol{\theta}}$ is expected to perform well in terms of accuracy and model fairness on $\mathcal{D}^{tgt}$, which is sampled from a target domain $e_t \notin \mathcal{E}^s$ and is unknown and inaccessible during training.

# 4 Methodology

We propose a novel framework, FADE, to address Eq.(6) as shown in Figure 1. This framework has three stages. In the first stage, the parameters of the SDM, label classifier, and sensitive classifier are pre-trained using the source dataset across various domains. In the second stage, the generator $G$ is derived by guiding the SDM using classifiers with the learned parameters. Finally, in the last stage, the data generated from $G$ with mitigated sensitive information are used to learn a fairness-aware domain invariant downstream classifier $f_{\boldsymbol{\theta}}$.

## 4.1 SDM and Classifiers Pre-training across Domains

As described in Figure 1, given the source data $\mathcal{D}^{src}$ consisting of $S$ training domains, we consider $K$ tasks. Each task $\mathcal{T}_k$ is associated with a support and a query set. The support set $\mathcal{T}_k^{sup} = \cup_{e_i \in \mathcal{T}_k^{sup}} \mathcal{B}_k^{e_i, sup}$ contains $S-1$ data batches, where each batch $\mathcal{B}_k^{e_i,sup}$ is sampled from the corresponding $\mathcal{D}^{e_i} \in \mathcal{D}^{src}$ and the query set $\mathcal{T}_k^{qry} = \cup_{e_i \in \mathcal{T}_k^{qry}} \mathcal{B}_k^{e_i,qry}$ is sampled from the remaining distinct data subset in $\mathcal{D}^{src}$.

Inspired by [Finn *et al.*, 2017], this stage aims to learn a good parameter $\Phi$ over the $K$ tasks that can be generalized on all source domains. The pre-training model contains three components: a SDM $s_{\boldsymbol{\theta}_s}$ introduced in the previous section, a label classifier $\phi_{\boldsymbol{\theta}_y} : \mathcal{X} \to \mathcal{Y}$, and a sensitive classifier $\psi_{\boldsymbol{\theta}_z} : \mathcal{X} \to \mathcal{Z}$. To optimize $\Phi = \{\boldsymbol{\theta}_s, \boldsymbol{\theta}_y, \boldsymbol{\theta}_z\}$ across domains, the objective function is given in Eq.(7).

$$\min_{\Phi} \frac{1}{K} \sum_{k=1}^{K} \Big( \delta \mathcal{L}_k^{sup}(\Phi) + (1-\delta) \mathcal{L}_k^{qry}(\Phi'_k) \Big) \tag{7}$$

where

$$\begin{aligned} &\Phi = \{\boldsymbol{\theta}_s, \boldsymbol{\theta}_y, \boldsymbol{\theta}_z\} \quad \text{and} \quad \Phi'_k = \{\boldsymbol{\theta}'_{s,k}, \boldsymbol{\theta}'_{y,k}, \boldsymbol{\theta}'_{z,k}\}; \\ &\boldsymbol{\theta}'_{s,k} = \boldsymbol{\theta}_s - \xi_s \nabla_{\boldsymbol{\theta}_s} s_{\boldsymbol{\theta}_s}(\mathcal{T}_k^{sup}); \\ &\boldsymbol{\theta}'_{y,k} = \boldsymbol{\theta}_y - \xi_y \nabla_{\boldsymbol{\theta}_y} \phi_{\boldsymbol{\theta}_y}(\mathcal{T}_k^{sup}); \\ &\boldsymbol{\theta}'_{z,k} = \boldsymbol{\theta}_z - \xi_z \nabla_{\boldsymbol{\theta}_z} \psi_{\boldsymbol{\theta}_z}(\mathcal{T}_k^{sup}); \\ &\mathcal{L}_k^{sup}(\Phi) = \mathcal{L}_s(\boldsymbol{\theta}_s, \mathcal{T}_k^{sup}) + \mathcal{L}_y(\boldsymbol{\theta}_y, \mathcal{T}_k^{sup}) + \mathcal{L}_z(\boldsymbol{\theta}_z, \mathcal{T}_k^{sup}); \\ &\mathcal{L}_k^{qry}(\Phi'_k) = \mathcal{L}_s(\boldsymbol{\theta}'_{s,k}, \mathcal{T}_k^{qry}) + \mathcal{L}_y(\boldsymbol{\theta}'_{y,k}, \mathcal{T}_k^{qry}) + \mathcal{L}_z(\boldsymbol{\theta}'_{z,k}, \mathcal{T}_k^{qry}). \end{aligned} \tag{8}$$

In Eq.(7), $\xi_s, \xi_y, \xi_z > 0$ are learning rates and $\delta \in [0,1]$ is an empirical weight balancing support and query losses. $\mathcal{L}_s$ is the score loss function defined in Eq.(4). $\mathcal{L}_y : \Theta_y \times \mathcal{H} \to \mathbb{R}$ and $\mathcal{L}_z : \Theta_z \times \mathcal{H} \to \mathbb{R}$ are cross-entropy loss functions for predicting class labels and sensitive attributes, respectively. Notice that $\boldsymbol{\theta}'_{s,k}, \boldsymbol{\theta}'_{y,k}, \boldsymbol{\theta}'_{z,k}$ are updated using one or few gradient steps for rapid optimization.

After sufficient iterations, $\Phi$ becomes optimized and invariant to domains. This occurs because each task can be viewed as a domain generalization problem within a sub-condition where the support sets act as "source data" and the query sets as "target data".

Furthermore, in the second stage of FADE, the SDM is guided to the generator $G$ by the label and sensitive classifiers using samples in $\mathcal{D}^{src}$, while keeping their parameters $\boldsymbol{\theta}_s, \boldsymbol{\theta}_y, \boldsymbol{\theta}_z$ unchanged.

## 4.2 Guiding the SDM to $G$ with Classifiers

**Debiasing with Fair Control.** To mitigate sensitive information from samples during the reverse diffusion process, we propose a novel sampling strategy. Assuming a strength signal $a_{fair} \in [0,1]$, we should sample from the conditional distribution $p(\mathbf{x}_t | a_{fair} = \lambda_z)$, hyperparameter $\lambda_z$ represents the strengh of fair control. Consequently, we need to solve the conditional reverse-time SDE:

$$\mathrm{d}\mathbf{x}_t = \left[ h(\mathbf{x}_t) - g^2(t) \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t | a_{fair}) \right] \mathrm{d}\bar{t} + g(t) \mathrm{d}\bar{\mathbf{w}}. \tag{9}$$

Since $p(\mathbf{x}_t | a_{fair} = \lambda_z) \propto p(\mathbf{x}_t) p(a_{fair} = \lambda_z | \mathbf{x}_t)$, we can derive the gradient relationship (proved in Appendix B) as follows:

$$\begin{aligned} \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t | a_{fair} = \lambda_z) = &\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t) \\ &+ \nabla_{\mathbf{x}_t} \log p(a_{fair} = \lambda_z | \mathbf{x}_t). \end{aligned} \tag{10}$$

Therefore, the problem we need to address is transformed into modeling $p(a_{fair} = \lambda_z | \mathbf{x}_t)$. For a sample that does not contain sensitive information, it will be uncertain to classify it definitively into any sensitive category. In other words, for debiased samples, training a classifier to predict their sensitive subgroups will result in a predicted distribution that approaches a uniform distribution. At this point, the entropy of the sensitive attribute prediction distribution $p_{\boldsymbol{\theta}_z}(z|\mathbf{x}_t)$ will reach its maximum. Based on this property, we define $p(a_{fair} = \lambda_z | \mathbf{x}_t)$ as:

$$p(a_{fair} = \lambda_z | \mathbf{x}_t) = \frac{\lambda_z H(p_{\boldsymbol{\theta}_z}(z|\mathbf{x}_t))}{C_t}, \tag{11}$$

where $H(\cdot)$ denotes the entropy function and $C_t$ is a normalization constant. In practice, $p_{\boldsymbol{\theta}_z}(z|\mathbf{x}_t)$ is approximated using a pre-trained sensitive classifier $\psi_{\boldsymbol{\theta}_z}$. Adding the gradient of the logarithm of Eq.(11) in the reverse diffusion process corresponds to maximizing the entropy of $p_{\boldsymbol{\theta}_z}(z|\mathbf{x}_t)$ at each time step $t$. This ensures that the samples drawn at each step contain minimal sensitive information, making it difficult for the classifier $\psi_{\boldsymbol{\theta}_z}$ to determine their sensitive category. Substituting Eq.(10) and Eq.(11) into Eq.(9) yields the final reverse SDE:

$$\begin{aligned} \mathrm{d}\mathbf{x}_t = &\Big[ h(\mathbf{x}_t) - g^2(t) (\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t) \\ &+ \lambda_z \nabla_{\mathbf{x}} \log H(p_{\boldsymbol{\theta}_z}(z|\mathbf{x}_t))) \Big] \mathrm{d}\bar{t} + g(t) \mathrm{d}\bar{\mathbf{w}}, \end{aligned} \tag{12}$$

where $\lambda_z$ is a hyperparameter that controls the guidance strength of the sensitive classifier $\psi_{\boldsymbol{\theta}_z}$. So far, we are able to generate samples that are free from sensitive information.

**Label Generation with Classifier Guidance.** To address our problem, we need to generate both the data and their labels simultaneously. Previous methods solve this by concatenating the labels with the features and then inputting them into the generative model for joint training [Xu *et al.*, 2018; Van Breugel *et al.*, 2021; Ramachandranpillai *et al.*, 2023]. However, a fair control process can affect these approaches by influencing the generated labels, leading to undesired labels. To tackle this challenge, we propose to employ a label generation process with classifier guidance [Dhariwal and Nichol, 2021]. We first specify a label $y$ and then use the SDM to model the conditional distribution $p(\mathbf{x}_t|y)$. We approach this by sampling from the conditional distribution $p(\mathbf{x}_t|y)$ where $y$ represents the label condition, by solving the conditional reverse-time SDE:

$$\mathrm{d}\mathbf{x}_t = \left[h(\mathbf{x}_t) - g^2(t)\nabla_{\mathbf{x}_t}\log p(\mathbf{x}|a_{fair}, y)\right]\mathrm{d}\bar{t} + g(t)\mathrm{d}\overline{\mathbf{w}}. \quad (13)$$

Since $\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t|a_{fair}, y) = \nabla_{\mathbf{x}_t}\log p(a_{fair}|\mathbf{x}_t) + \nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t) + \nabla_{\mathbf{x}_t}\log p(y|\mathbf{x}_t, a_{fair})$ (proved in Appendix B), we need a pre-trained classifier to simulate $p(y|\mathbf{x}_t, a_{fair})$. For any given sample, regardless of whether we apply fair control to it, we need to specify the generation of its label $y$ ( $y$ is independent of $a_{fair}$). Therefore, $p_{\boldsymbol{\theta}_y}(y|\mathbf{x}_t) = p_{\boldsymbol{\theta}_y}(y|\mathbf{x}_t, a_{fair})$ can be approximated using a pre-trained sensitive classifier $\phi_{\boldsymbol{\theta}_y}$. we can rewrite Eq.(13) as:

$$\begin{aligned}\mathrm{d}\mathbf{x}_t = \Big[&h(\mathbf{x}_t) - g^2(t)(\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t) + \lambda_z\nabla_{\mathbf{x}_t}\log p_{\boldsymbol{\theta}_z}(a_{fair}|\mathbf{x}_t)\\ &+ \lambda_y\nabla_{\mathbf{x}_t}\log p_{\boldsymbol{\theta}_y}(y|\mathbf{x}_t))\Big]\mathrm{d}\bar{t} + g(t)\mathrm{d}\overline{\mathbf{w}}, \quad (14)\end{aligned}$$

where $\lambda_y$ is a hyperparameter that controls the guidance strength of the label classifier $\phi_{\boldsymbol{\theta}_y}$. We can specify labels for generating samples, rather than generating labels and samples simultaneously at random. Using this method, we can specify the same number of samples to be generated for different classes, thereby creating a balanced dataset to avoid insufficient feature information for minority classes in downstream classifiers. At this point, we can generate a dataset that does not contain sensitive information with labels.

### 4.3 Learning Downstream Classifier Using the Generated Data by $G$

Let $G$ be the score network $\boldsymbol{\theta}_s$ guided by $\boldsymbol{\theta}_y$ and $\boldsymbol{\theta}_z$. Given a target class $y$, iteratively input Gaussian noise into the generator $G$ to output a subset of data with label $y$ by Eq.(14). Perform the same operation for other classes as well, and then merge all these subsets into the final dataset $\mathcal{D}^{gen}$. Our downstream training objective is to train a classifier $f_\theta$ that satisfies:

$$\min_{\boldsymbol{\theta}}\mathbb{E}_{(\mathbf{x}', z', y')\in\mathcal{D}^{gen}}[\ell(f_{\boldsymbol{\theta}}(\mathbf{x}'), y')] \quad (15)$$

where $\ell$ represents a classification loss function such as cross-entropy.

### 4.4 Algorithmic Summary of FADE

The overall process of FADE is illustrated in Algorithm 1. Lines 1 to 8 represent the pre-training phase of the model parameter $\Phi = \{\boldsymbol{\theta}_s, \boldsymbol{\theta}_y, \boldsymbol{\theta}_z\}$, while lines 9 and 10 respectively represent the data generation phase and the downstream classifier training. During the pre-training phase of the model, by constructing different $\mathcal{T}_k^{sup}$ and $\mathcal{T}_k^{qry}$, the model gains the ability to generate data from non-training domains. In the data generation phase, we do not require any additional training. We can simply use the pre-trained classifiers, $\phi_{\boldsymbol{\theta}_y}$ and $\psi_{\boldsymbol{\theta}_z}$, as a guide to generate data that is free from biased information. Finally, during the downstream classifier training, only the simplest classification loss is required.

**Algorithm 1** FADE

**Input**: Labeled source datasets $\mathcal{D}^{src}$ with $S$ domains; score network $s_{\boldsymbol{\theta}_s}$; label classifier $\phi_{\boldsymbol{\theta}_y}$; sensitive classifier $\psi_{\boldsymbol{\theta}_z}$; downstream classifier $f_{\boldsymbol{\theta}}$; hyperparameters $\alpha_s, \alpha_y, \alpha_z, \beta_s, \beta_y, \beta_z, \gamma_s, \gamma_y, \gamma_z$.
**Initialize:** $\boldsymbol{\theta}_s, \boldsymbol{\theta}_y, \boldsymbol{\theta}_z$, and $\boldsymbol{\theta}$

1: **repeat**
2: Sample $K$ tasks $\{\mathcal{T}_k\}_{k=1}^K$ from $\mathcal{D}^{src}$
3: **for** each task $\mathcal{T}_k = \{\mathcal{T}_k^{sup}, \mathcal{T}_k^{qry}\}$ **do**
4: Evaluate $\mathcal{L}_k^{sup}$ using $\boldsymbol{\theta}_s, \boldsymbol{\theta}_y, \boldsymbol{\theta}_z$ on $\mathcal{T}_k^{sup}$
5: Update $\boldsymbol{\theta}'_{s,k}, \boldsymbol{\theta}'_{y,k}, \boldsymbol{\theta}'_{z,k} \leftarrow \boldsymbol{\theta}_s, \boldsymbol{\theta}_y, \boldsymbol{\theta}_z$ in Eq.(8)
6: Evaluate $\mathcal{L}_k^{qry}$ using $\boldsymbol{\theta}'_{s,k}, \boldsymbol{\theta}'_{y,k}, \boldsymbol{\theta}'_{z,k}$ on $\mathcal{T}_k^{qry}$
7: **end for**
8: Update $\boldsymbol{\theta}_s, \boldsymbol{\theta}_y, \boldsymbol{\theta}_z$ using $\{\mathcal{L}_k^{sup}, \mathcal{L}_k^{qry}\}_{k=1}^K$ in Eq.(7)
9: **until** convergence
10: Derive $G$ from $s_{\boldsymbol{\theta}_s}$ guided by $\phi_{\boldsymbol{\theta}_y}$ and $\psi_{\boldsymbol{\theta}_z}$ and generate $\mathcal{D}^{gen}$ from $G$ in Eq.(14)
11: Iteratively optimize $f_{\boldsymbol{\theta}}$ using $\mathcal{D}^{gen}$

## 5 Experiments

**Evaluation Metrics**. We measure the domain generalization performance using *Accuracy* and evaluate the algorithmic fairness using four evaluation metrics as follows:

- Difference of Demographic Parity ($\Delta_{DP}$) [Dwork *et al.*, 2012] requires that the acceptance rate provided by the algorithm should be the same across all sensitive subgroups, which is formalized as

$$\Delta_{DP} = \left|\mathbb{P}(\hat{Y}=1|Z=0) - \mathbb{P}(\hat{Y}=1|Z=1)\right|$$

where $\hat{Y}$ is the predicted class label.

- Difference of Equalized Odds ($\Delta_{EO}$) [Hardt *et al.*, 2016] examines fairness in terms of equal true positive and false positive rates, which is formalized as

$$\Delta_{EO} = \frac{1}{2}\sum_y\Big|\mathbb{P}(\hat{Y}=1|Z=0, Y=y) - \mathbb{P}(\hat{Y}=1|Z=1, Y=y)\Big|$$

- Difference of Equalized Opportunity ($\Delta_{EOp}$) [Hardt *et al.*, 2016] evaluates the fairness of true positive rates specifically, which is formalized as

$$\Delta_{EOp} = \left|\mathbb{P}(\hat{Y}=1|Z=0, Y=1) - \mathbb{P}(\hat{Y}=1|Z=1, Y=1)\right|$$

- Difference of dataset distance ($D_{fair}$) measures the fairness of a dataset by evaluating the difference between the distances from the mean point of the generated dataset to the mean points of two sensitive groups in the original dataset. We formulally define it as:

$$D_{fair} = \Big|\,\|\mu(\mathcal{D}^{gen}) - \mu(\mathcal{D}^{src}_{z=0})\|_2 - \|\mu(\mathcal{D}^{gen}) - \mu(\mathcal{D}^{src}_{z=1})\|_2\,\Big|$$

where $\mu(\cdot)$ denotes the operation of taking the mean point of a dataset, $\mathcal{D}^{src}_{z=0}$ and $\mathcal{D}^{src}_{z=1}$ represent different sensitive subsets within $\mathcal{D}^{src}$, respectively.

The smaller the values of $\Delta_{DP}$, $\Delta_{EO}$ $\Delta_{EOp}$ and $D_{fair}$, the fairer the algorithm.

**Compared Methods**. We compare FADE with three classic generative models VAE [Kingma and Welling, 2013], Accuracy ↑ / $\Delta_{DP}$ ↓ / $\Delta_{EO}$ ↓ / $\Delta_{EOp}$ ↓

| | | White | | | | Black | | | | A-I-E | | | | A-P-I | | | | Others | | | | Avg | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Adult | VAE | 79.12 | 0.23 | 0.25 | 0.37 | 89.05 | 0.12 | 0.21 | 0.36 | 85.03 | 0.16 | 0.21 | 0.30 | 76.79 | 0.44 | 0.43 | 0.55 | 86.37 | 0.12 | 0.20 | 0.34 | 83.27 | 0.21 | 0.26 | 0.38 |
| | GAN | 72.27 | 0.26 | 0.27 | 0.42 | 72.07 | 0.09 | 0.33 | **0.20** | 76.51 | 0.16 | 0.22 | 0.35 | 77.57 | 0.23 | 0.27 | 0.22 | 83.70 | 0.11 | 0.23 | 0.25 | 76.42 | 0.17 | 0.26 | 0.29 |
| | DDPM | 79.51 | 0.27 | 0.28 | 0.42 | **89.47** | 0.09 | 0.18 | 0.31 | 86.74 | 0.11 | 0.17 | 0.26 | **78.37** | 0.28 | 0.33 | 0.53 | **87.34** | 0.11 | 0.20 | 0.34 | 84.29 | 0.17 | 0.23 | 0.37 |
| | FGAN | 74.06 | **0.14** | 0.17 | **0.22** | 83.44 | 0.06 | 0.12 | 0.22 | 84.89 | 0.06 | **0.11** | 0.20 | 71.66 | 0.14 | 0.19 | 0.27 | 79.91 | **0.04** | 0.06 | 0.08 | 78.99 | **0.09** | 0.13 | 0.20 |
| | DECAF | 73.74 | 0.16 | 0.22 | 0.26 | 87.38 | 0.08 | 0.23 | 0.28 | 84.72 | 0.06 | 0.26 | 0.31 | 71.99 | 0.15 | 0.24 | 0.30 | 84.26 | 0.05 | 0.12 | 0.15 | 80.42 | 0.10 | 0.21 | 0.26 |
| | FDisCo | 73.89 | 0.31 | 0.33 | 0.42 | 85.05 | 0.05 | 0.19 | 0.33 | 71.49 | 0.11 | 0.23 | 0.32 | 72.43 | **0.05** | **0.12** | **0.20** | 83.15 | 0.06 | 0.07 | **0.06** | 77.20 | 0.12 | 0.19 | 0.27 |
| | FLDGM | 61.22 | 0.17 | 0.20 | 0.26 | 58.09 | 0.09 | 0.17 | 0.27 | 64.29 | 0.23 | 0.27 | 0.31 | 61.94 | 0.19 | 0.26 | 0.37 | 62.13 | 0.25 | 0.27 | 0.31 | 61.53 | 0.19 | 0.24 | 0.30 |
| | FADE (Ours) | **81.62** | 0.15 | **0.15** | 0.23 | 89.19 | **0.04** | **0.11** | 0.21 | **87.99** | **0.05** | 0.12 | **0.19** | 77.04 | 0.17 | 0.18 | 0.27 | 86.92 | **0.04** | **0.04** | **0.06** | **84.55** | **0.09** | **0.12** | **0.19** |

| | | Basic-4-years | | | | High-school | | | | Basic-9-years | | | | University-degree | | | | Professional-course | | | | Avg | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Bank | VAE | 82.78 | 0.12 | 0.15 | 0.23 | **83.34** | 0.09 | 0.10 | 0.16 | 85.42 | 0.07 | 0.14 | 0.24 | **79.03** | 0.08 | 0.11 | 0.17 | **83.86** | 0.05 | 0.09 | 0.14 | **82.88** | 0.08 | 0.12 | 0.19 |
| | GAN | 79.09 | 0.12 | 0.14 | 0.26 | 72.84 | 0.05 | 0.08 | 0.13 | 80.38 | 0.05 | 0.13 | 0.26 | 75.18 | 0.06 | 0.09 | 0.12 | 81.62 | 0.08 | 0.11 | 0.22 | 78.48 | 0.08 | 0.12 | 0.22 |
| | DDPM | **83.86** | 0.11 | 0.16 | 0.25 | 82.71 | 0.04 | 0.10 | 0.16 | **85.58** | 0.05 | 0.13 | 0.22 | 78.13 | 0.07 | 0.11 | 0.17 | 82.85 | 0.04 | 0.10 | 0.19 | 82.65 | 0.06 | 0.12 | 0.20 |
| | FGAN | 80.91 | 0.07 | 0.11 | 0.18 | 78.75 | 0.03 | 0.06 | 0.11 | 82.64 | 0.05 | 0.10 | 0.17 | 73.33 | 0.02 | 0.04 | **0.07** | 75.38 | 0.04 | **0.04** | 0.12 | 78.20 | 0.05 | 0.07 | 0.13 |
| | DECAF | 80.17 | **0.06** | 0.10 | 0.19 | 77.60 | **0.02** | 0.05 | 0.12 | 83.24 | 0.05 | 0.09 | 0.15 | 72.93 | 0.02 | 0.05 | 0.13 | 71.70 | 0.12 | 0.21 | 0.13 | 77.13 | 0.06 | 0.10 | 0.14 |
| | FDisCo | 81.73 | 0.08 | 0.14 | 0.24 | 80.53 | 0.06 | 0.10 | 0.17 | 82.87 | 0.04 | 0.08 | 0.16 | 76.37 | 0.06 | 0.10 | 0.15 | 81.66 | 0.07 | 0.11 | 0.17 | 80.63 | 0.06 | 0.11 | 0.18 |
| | FLDGM | 75.68 | 0.17 | 0.19 | 0.21 | 74.69 | 0.07 | 0.11 | 0.15 | 78.45 | 0.08 | 0.13 | 0.19 | 70.26 | 0.08 | 0.12 | 0.15 | 73.63 | 0.07 | 0.11 | 0.15 | 74.54 | 0.10 | 0.13 | 0.17 |
| | FADE (Ours) | 82.52 | **0.06** | **0.09** | **0.14** | 80.73 | **0.02** | **0.04** | **0.07** | 85.31 | **0.03** | **0.07** | **0.13** | 75.91 | **0.01** | **0.03** | **0.07** | 79.67 | **0.03** | 0.06 | **0.11** | 80.84 | **0.03** | **0.06** | **0.11** |

Table 1: Performance on Adult and Bank datasets (**bold** is the best, underline is the second best).

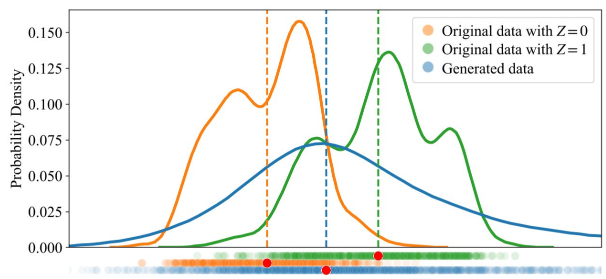
GAN [Goodfellow *et al.*, 2020], and DDPM [Ho *et al.*, 2020]. Additionally, four models for fair data generation methods have also been considered: FGAN [Xu *et al.*, 2018], DECAF [Van Breugel *et al.*, 2021], FDisCo [Liu *et al.*, 2022] and FLDGM [Ramachandranpillai *et al.*, 2023].

**Datasets and settings**. We use three datasets: Adult [Kohavi and others, 1996], Bank [Moro *et al.*, 2014] and NYSF [Goel *et al.*, 2016]. Please refer to Appendix C for all datasets, network architecture, hardware environment, training and testing strategies.

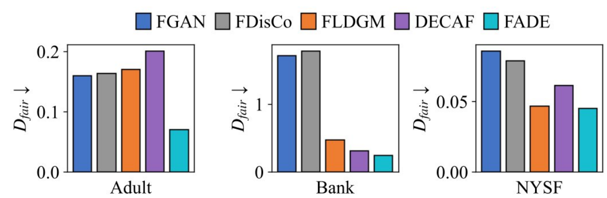
## 6 Results

**Overall Performance.** The overall performance of FADE and its competing methods on Adult and Bank datasets is presented in Table 1. Due to space limitations, we have omitted the full results with standard deviations and those for NYSF dataset, including them in Appendix D.1. Focus on the average of each metric across all domains, FADE achieves the best performance in both DG and fairness on Adult datasets simultaneously. Although the accuracy of FADE on the Bank dataset is lower compared to some other methods, it achieves the most equitable results with minimal performance degradation, and its accuracy is also higher than those methods that focus on fairness. Both VAE and DDPM achieve decent classification accuracy, but due to their lack of fairness consideration, they cannot guarantee the algorithmic fairness. Although FairGAN and DECAF outperform FADE in fairness performance in some domains, their DG performance is not competitive. Overall, FADE ensures fairness while maintaining strong classification capabilities, which means that it has successfully generated domain-invariant fair data.

**Analysis of Generated Dataset.** We further analyze the dataset generated by FADE from a data perspective. We use linear discriminant analysis [Balakrishnama and Ganapathiraju, 1998] (LDA) to reduce the dimensionality of both the original and generated datasets to one dimension and plot their probability densities. Figure 2(a) shows the visualization result on the White domain of Adult dataset. The original data for different sensitive groups have significantly different probability density distributions and means. The mean of the data generated by FADE falls between the means of the two sensitive group data, indicating that it produces a more fair data distribution. Additionally, the distribution of the generated data covers a wider range compared to the original data, suggesting that it generates generalized data beyond




(a) LDA visualization on Adult dataset.




(b) $D_{fair}$ performance on three datasets.

Figure 2: (a) LDA visualization of the original dataset and the dataset generated by FADE. The orange, green, and blue curves represent the probability densities of the original data for two sensitive groups and the generated data, respectively. The three dashed lines indicate the means. The markers below the x-axis show the one-dimensional data, with red points representing the mean points. (b) Average $D_{fair}$ results for all fairness-aware methods across three datasets.

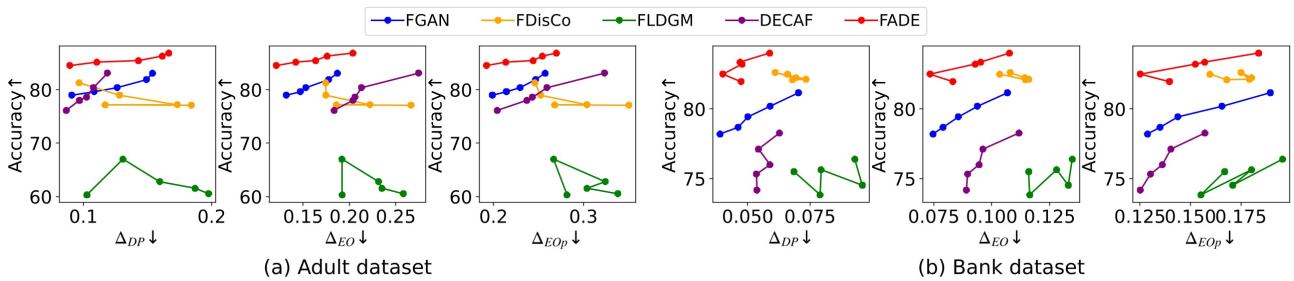



Figure 3: Accuracy-fairness trade-offs of FADE by various $\lambda_z \in \{0.1, 1, 10, 50, 100\}$ on (a) Adult and (b) Bank datasets over different baselines. The upper left indicates a better trade-off performance. Results are averaged across all domains.

| | | Accuracy↑ | $\Delta_{DP}\downarrow$ | $\Delta_{EO}\downarrow$ | $\Delta_{EOp}\downarrow$ |
|---|---|---|---|---|---|
| Adult | FADE | 84.55 | 0.089 | 0.121 | 0.192 |
| | *w/o* FG | 85.17(+0.62) | 0.177 (+0.088) | 0.204 (+0.083) | 0.304 (+0.112) |
| | *w/o* LG | 83.18 (-1.37) | 0.106 (+0.017) | 0.138 (+0.017) | 0.219 (+0.027) |
| | *w/o* Query | 81.55 (-3.00) | 0.231 (+0.142) | 0.217 (+0.096) | 0.269 (+0.077) |
| Bank | FADE | 80.84 | 0.030 | 0.062 | 0.108 |
| | *w/o* FG | 82.45(+1.61) | 0.052 (+0.022) | 0.104 (+0.042) | 0.175 (+0.067) |
| | *w/o* LG | 80.43 (-0.41) | 0.090 (+0.060) | 0.126 (+0.064) | 0.178 (+0.070) |
| | *w/o* Query | 80.10 (-0.74) | 0.032 (+0.002) | 0.069 (+0.007) | 0.111 (+0.003) |

Table 2: Ablation study on Adult and Bank datasets. Results are averaged across all domains.

the original distribution.

Furthermore, we compare the average of $D_{fair}$ across all domains between the generated datasets and the original datasets for all fairness-aware methods across three datasets, as shown in Figure 2(b). FADE achieves the best $D_{fair}$ results across three datasets, indicating that FADE generates data with the smallest distance to the two original sensitive subgroups. Compared to other fairness-aware baselines, FADE produced the fairest data, further validating that the fairness guidance component enables FADE to generate unbiased data effectively. Meanwhile, we also observe that other methods do not perform consistently across different datasets, indicating that their debiasing performance is significantly influenced by the distribution of the original data. Particularly for DECAF, its performance is highly dependent on the causal structure of the original data. In contrast, FADE's consistent results demonstrate its strong robustness.

**Accuracy-Fairness Trade-offs.** To evaluate the accuracy-fairness trade-offs of FADE, we vary $\lambda_z \in \{0.1, 1, 10, 50, 100\}$ to obtain different classification and fairness results. A larger $\lambda_z$ indicates a greater emphasis on algorithmic fairness by FADE. We compare the results with other fairness-aware methods using three fairness metrics across Adult and Bank datasets, as shown in Figure 3. FADE's results across various $\lambda_z$ are positioned in the upper-left corner compared to all other methods. This suggests that FADE not only delivers superior fairness performance but also maintains competitive classification performance, achieving the best accuracy-fairness trade-offs. Although FGAN and DECAF are comparable to FADE in terms of algorithmic fairness, their lower classification accuracy results in a poorer accuracy-fairness trade-off.

**Ablation Study.** To evaluate the effect of components in the design of FADE, we conduct ablation studies as shown in Table 2. The results on the NYPD dataset are provided in Appendix D.2. We propose 3 variants of FADE: **(i) Without fair guidance (*w/o* FG).** We set $\lambda_z = 0$ to ensure that there is no fair control during the sampling process. Despite achieving a marginal advantage in prediction accuracy on the adult and bank dataset, a sharp increase in all three fairness metrics resulted in unfair classification outcomes (Table 2). **(ii) Without label guidance (*w/o* LG).** We set $\lambda_y$ as 0 firstly. To generate the labels for the dataset, we concatenate the labels with the features and jointly train the generator using this combined input. However, by doing so, the labels generated are influenced by fair control, leading to an increase in incorrect labels and, consequently, a decline in accuracy performance. Furthermore, our empirical observations indicate that the fairness of both datasets has also decreased to varying degrees. **(iii) Without Query Set (*w/o* Query).** Setting $\delta = 1$ in Eq.(7), the diffusion model's sampling process may retain training domain information, which can impact the accuracy in the target domain, particularly evident in the Adult dataset. Even if the accuracy in the Bank dataset does not significantly decrease, the shift in the target distribution can hinder the sensitive classifier from accurately discriminating sensitive attributes, thereby impeding the effective application of fair control and resulting in a decline in fairness.

# 7 Conclusion

In this paper, we introduce FADE, a three-stage framework designed to generate unbiased and domain-invariant data for fairness-aware domain generalization. By leveraging a combination of pre-trained score-based diffusion models (SDM) and classifier guidance, FADE effectively eliminates bias during data generation without the need for additional training. Our approach not only enhances the generalization capabilities of the generated data but also ensures fairness when applied to downstream tasks, particularly in scenarios involving distribution shifts.

# Contribution Statement

Yujie Lin and Dong Li contributed equally to this work and are considered co-first authors. Minglai Shao is the corresponding author.

## References

[Alder and Gilbert, 2006] G Stoney Alder and Joseph Gilbert. Achieving ethics and fairness in hiring: Going beyond the law. *Journal of Business Ethics*, 68:449–464, 2006.

[Balakrishnama and Ganapathiraju, 1998] Suresh Balakrishnama and Aravind Ganapathiraju. Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing*, 18(1998):1–8, 1998.

[Dhariwal and Nichol, 2021] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[Dwork *et al.*, 2012] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[Finn *et al.*, 2017] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[Goel *et al.*, 2016] Sharad Goel, Justin M Rao, and Ravi Shroff. Precinct or prejudice? understanding racial disparities in new york city's stop-and-frisk policy. 2016.

[Goodfellow *et al.*, 2020] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[Gulrajani and Lopez-Paz, 2020] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

[Hardt *et al.*, 2016] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Jiang *et al.*, 2024] Kai Jiang, Chen Zhao, Haoliang Wang, and Feng Chen. Feed: Fairness-enhanced meta-learning for domain generalization. In *2024 IEEE International Conference on Big Data (BigData)*, pages 949–958. IEEE, 2024.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[Kohavi and others, 1996] Ron Kohavi et al. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pages 202–207, 1996.

[Komanduri *et al.*, 2024] Aneesh Komanduri, Chen Zhao, Feng Chen, and Xintao Wu. Causal diffusion autoencoders: Toward counterfactual generation via diffusion probabilistic models. In *ECAI 2024*, pages 2516–2523. IOS Press, 2024.

[Li *et al.*, 2024] Dong Li, Chen Zhao, Minglai Shao, and Wenjun Wang. Learning fair invariant representations under covariate and correlation shifts simultaneously. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1174–1183, 2024.

[Lian *et al.*, 2025] Heng Lian, Chen Zhao, Zhong Chen, Xingquan Zhu, My T Thai, and Yi He. Metric-agnostic continual learning for sustainable group fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18648–18657, 2025.

[Lin *et al.*, 2023a] Yujie Lin, Chen Zhao, Minglai Shao, Baoluo Meng, Xujiang Zhao, and Haifeng Chen. Towards counterfactual fairness-aware domain generalization in changing environments. *arXiv preprint arXiv:2309.13005*, 2023.

[Lin *et al.*, 2023b] Yujie Lin, Chen Zhao, Minglai Shao, Xujiang Zhao, and Haifeng Chen. Adaptation speed analysis for fairness-aware causal models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 1421–1430, 2023.

[Lin *et al.*, 2024] Yujie Lin, Dong Li, Chen Zhao, Xintao Wu, Qin Tian, and Minglai Shao. Supervised algorithmic fairness in distribution shifts: A survey. *arXiv preprint arXiv:2402.01327*, 2024.

[Liu *et al.*, 2022] Ji Liu, Zenan Li, Yuan Yao, Feng Xu, Xiaoxing Ma, Miao Xu, and Hanghang Tong. Fair representation learning: An alternative to mutual information. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1088–1097, 2022.

[Mitchell *et al.*, 2021] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual review of statistics and its application*, 8(1):141–163, 2021.

[Moro *et al.*, 2014] Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

[Oh *et al.*, 2022] Changdae Oh, Heeji Won, Junhyuk So, Taero Kim, Yewon Kim, Hosik Choi, and Kyungwoo Song. Learning fair representation via distributional contrastive disentanglement. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1295–1305, 2022.

[Pham *et al.*, 2023] Thai-Hoang Pham, Xueru Zhang, and Ping Zhang. Fairness and accuracy under domain generalization. *arXiv preprint arXiv:2301.13323*, 2023.

[Purificato *et al.*, 2023] Erasmo Purificato, Flavio Lorenzo, Francesca Fallucchi, and Ernesto William De Luca. The use of responsible artificial intelligence techniques in the context of loan approval processes. *International Journal of Human–Computer Interaction*, 39(7):1543–1562, 2023.

[Ramachandranpillai *et al.*, 2023] Resmi Ramachandranpillai, Md Fahim Sikder, and Fredrik Heintz. Fair latent deep generative models (fldgms) for syntax-agnostic and fair synthetic data generation. In *ECAI 2023*, pages 1938–1945. IOS Press, 2023.

[Shao *et al.*, 2024] Minglai Shao, Dong Li, Chen Zhao, Xintao Wu, Yujie Lin, and Qin Tian. Supervised algorithmic fairness in distribution shifts: A survey. *arXiv preprint arXiv:2402.01327*, 2024.

[Shimodaira, 2000] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

[Song *et al.*, 2020] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[Van Breugel *et al.*, 2021] Boris Van Breugel, Trent Kyono, Jeroen Berrevoets, and Mihaela Van der Schaar. Decaf: Generating fair synthetic data using causally-aware generative networks. *Advances in Neural Information Processing Systems*, 34:22221–22233, 2021.

[Xu *et al.*, 2018] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE international conference on big data (big data)*, pages 570–575. IEEE, 2018.

[Zhang *et al.*, 2018] Zhengxin Zhang, Xiaosheng Si, Changhua Hu, and Yaguo Lei. Degradation data analysis and remaining useful life estimation: A review on wiener-process-based methods. *European Journal of Operational Research*, 271(3):775–796, 2018.

[Zhao and Chen, 2019] Chen Zhao and Feng Chen. Rank-based multi-task learning for fair regression. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 916–925. IEEE, 2019.

[Zhao *et al.*, 2023] Chen Zhao, Feng Mi, Xintao Wu, Kai Jiang, Latifur Khan, Christan Grant, and Feng Chen. Towards fair disentangled online learning for changing environments. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023.

[Zhao *et al.*, 2024] Chen Zhao, Kai Jiang, Xintao Wu, Haoliang Wang, Latifur Khan, Christan Grant, and Feng Chen. Algorithmic fairness generalization under covariate and dependence shifts simultaneously. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4419–4430, 2024.

[Zhou *et al.*, 2022] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, 2022.