

# Fair Sampling in Diffusion Models through Switching Mechanism

Yujin Choi<sup>1\*</sup>, Jinseong Park<sup>1\*</sup>, Hoki Kim<sup>1</sup>, Jaewook Lee<sup>1</sup>, Saerom Park<sup>2†</sup>

<sup>1</sup>Seoul National University

<sup>2</sup>Ulsan National Institute of Science and Technology (UNIST)

uznhigh@snu.ac.kr, jinseong@snu.ac.kr, ghrl9613@snu.ac.kr, jaewook@snu.ac.kr, srompark@unist.ac.kr

## Abstract

Diffusion models have shown their effectiveness in generation tasks by well-approximating the underlying probability distribution. However, diffusion models are known to suffer from an amplified inherent bias from the training data in terms of fairness. While the sampling process of diffusion models can be controlled by conditional guidance, previous works have attempted to find empirical guidance to achieve quantitative fairness. To address this limitation, we propose a fairness-aware sampling method called *attribute switching* mechanism for diffusion models. Without additional training, the proposed sampling can obfuscate sensitive attributes in generated data without relying on classifiers. We mathematically prove and experimentally demonstrate the effectiveness of the proposed method on two key aspects: (i) the generation of fair data and (ii) the preservation of the utility of the generated data.

## Introduction

Generative models have shown promise across diverse domains to generate unseen data examples. Recently, diffusion-based models (Song et al. 2020; Rombach et al. 2022) have shown the effectiveness in various fields, including image (Rombach et al. 2022), video (Blattmann et al. 2023), and audio (Kong et al. 2020) generation. However, diffusion models also suffer from generating data possessing inherited and amplified bias from the training dataset (Friedrich et al. 2023), similar to the problems observed in other using generative models such as autoencoder (Amini et al. 2019) or generative adversarial networks (GAN) (Sattigeri et al. 2019; Xu et al. 2018). Since the diffusion model needs specific methods to generate synthetic data by denoising from stochastic differential equation (SDE) and ordinary differential equation (ODE) (Song, Meng, and Ermon 2020), it requires unique methods for achieving fairness in contrast to previous fair generative models (Kenfack et al. 2021). In response, recent studies have investigated fairness issues in diffusion models (Friedrich et al. 2023; Sinha et al. 2021). These studies have mainly focused on quantitative aspects of achieving fairness within the diffusion model frame-

work. For instance, Friedrich et al. (2023) introduces the fair guidance to generate sample fair data to resolve the imbalanced sampling for certain classes. However, assessing models through a specific classifier falls short of capturing the broader performance spectrum (Xu et al. 2019). In addition, evaluating the fairness of a generative model through classification tasks introduces complexities in distinguishing whether the observed discrimination stems from the generator or the classifier. Instead, in this study, we employ the concept of  $\epsilon$ -fairness, as introduced in (Feldman et al. 2015; Xu et al. 2018). We aim to evaluate data fairness concerning the sensitive attribute itself without a specific label by achieving data fairness in diffusion models.

To achieve the goal, we present a sampling method called *attribute switching* to generate synthetic data under the  $\epsilon$ -fairness framework. Inspired by the finding in (Choi et al. 2022) that diffusion models learn different characteristics at each step, our work aims to generate data that is both independent of sensitive attributes and satisfies utility requirements. Subsequently, we prove the theoretical condition of the transition point, ensuring independence so that achieving fairness. Furthermore, our model does not require additional training on top of standard diffusion models, offering flexibility in its applicability to various pre-trained models. As our method is solely designed for diffusion sampling, we demonstrate the effectiveness of attribute switching using various recent diffusion models. We summarize our main contributions as follows:

- We address the data fairness issue in diffusion models, aiming to ensure sensitive attribute independence without requiring additional classifiers. To the best of our knowledge, this is the first attempt to investigate  $\epsilon$ -fairness within diffusion models.
- We introduce a novel sampling approach, which we call *attribute switching*, that can generate fair synthetic data by matching the data distribution. Additionally, we mathematically prove its effectiveness and present an efficient algorithm to find the optimal transition point to achieve fairness.
- We show the effectiveness of the proposed fair sampling on various datasets including the state-of-art diffusion models without additional training.

\*These authors contributed equally.

†Corresponding author.

Notions	Classification Fairness		Data Fairness	
	Statistical parity	Equalized odds	Statistical fairness	$\epsilon$ -fairness
Equations	$P(\eta(X) S=1) = P(\eta(X) S=0)$	$P(\eta(X)=1 Y,S=1) = P(\eta(X)=1 Y,S=0)$	$P(Y S=1) = P(Y S=0)$	Equation (1) $\text{BER}(f(X), S) > \epsilon$
Independent to Classifier $\eta$	X	X	O	O
Independent to Label $Y$	X	X	X	O

Table 1: Various notions for fairness. Among the fair notions,  $\epsilon$ -fairness does not rely on classifier  $\eta$  and label  $Y$ .

## Preliminary and Related Work

In this section, we introduce key notations, definitions, and related works to explain fairness issues in generative models.

**Notations** We consider the sensitive attributes  $S \in \mathcal{S}$  and input data  $X \in \mathcal{X}$ , where  $(X, S) \sim p_D$  represents the underlying distribution of the training data. For synthetic data, we introduce a hat notation  $\hat{X}$ . In addition, when defining fairness notions for the classification tasks, we consider the label  $Y \in \mathcal{Y}$  or a classifier  $\eta : \mathcal{X} \rightarrow \mathcal{Y}$ , where data distribution becomes  $(X, Y, S) \sim p_D$ . For diffusion models,  $p_t(X_t)$  denotes the distribution of  $X_t$  for the diffusion process  $\{X_t\}_{t=0}^T$  and  $X_0 \sim p_D(X)$ .

## Fairness Notions

With regards to group fairness, prior research has introduced a range of fairness concepts (Feldman et al. 2015; Verma and Rubin 2018). Among these, Table 1 summarizes the predominant definitions (Verma and Rubin 2018), which categorizes the notions into classification fairness (prediction-dependent) and data fairness (prediction-independent) categories. Among them, classification fairness requires the consideration of downstream tasks with the corresponding classifiers. For instance, the *statistical parity* focuses on ensuring that predicted labels are independent of the sensitive attribute (Dwork et al. 2012), and *equalized odds* fairness aims that the classifier  $\eta$  maintains the equal values of ROC for label  $Y$  (Hardt, Price, and Srebro 2016).

Therefore, to circumvent the necessity for an additional classifier, we focus on the concept of data fairness. However, although *statistical fairness* does not directly depend on the classifier  $\eta$ , it requires taking into account a specific downstream task based on the relationship between the sensitive attribute  $S$  and the label  $Y$ . Intuitively, it is more natural to consider the input data  $X$  alongside the sensitive attribute  $S$  to assess generative models. Thus, we focus on  *$\epsilon$ -fairness* (Feldman et al. 2015; Xu et al. 2018) defined as follows:

**Definition 1. ( $\epsilon$ -fairness)** For any function  $f : \mathcal{X} \rightarrow \mathcal{S}$ , a dataset  $D = (X, S)$  satisfies  $\epsilon$ -fairness if

$$\text{BER}(f(X), S) > \epsilon,$$

where the balanced error rate (BER) is defined as

$$\text{BER}(f(X), S) = \frac{P(f(X)=0|S=1) + P(f(X)=1|S=0)}{2}. \quad (1)$$

The  $\epsilon$ -fairness evaluates fairness concerning the sensitive attribute  $S$  itself. Xu et al. (2018) pointed out that  $\epsilon$ -fairness

can be measured by the classifiers trained on synthetic data. With a high classification error rate on real data, sensitive attribute  $S$  is not predictable by synthetic data  $\hat{X}$  and the inherent disparate impact in the real data can be removed in synthetic data. In this context, the  $\epsilon$ -fairness implies that the synthetic data  $\hat{X}$  from the fair generator can assure the independence of  $S$ , denoted as  $\hat{X} \perp\!\!\!\perp S$ .

## Diffusion Models

Diffusion models (Ho, Jain, and Abbeel 2020; Dhariwal and Nichol 2021) become the de-facto standard method for generative models by achieving both sample qualities and diversity. To approximate the data distribution, diffusion models forward sample data into noise distribution from time steps  $t = 0$  to  $t = T$ , where  $T$  is the number of steps in the denoising process. Then, the model learns the reverse diffusion process by matching score functions via SDE and ODE (Song, Meng, and Ermon 2020; Song et al. 2020). The forward diffusion process is described with an Itô SDE:

$$dX_t = f(X_t, t)dt + g(t)dW_t, \quad (2)$$

for drift coefficient  $f : \mathcal{X} \times [0, T] \rightarrow \mathcal{X}$ , diffusion coefficient  $g : [0, T] \rightarrow \mathbb{R}$ , and Brownian motion  $W$ . For the forward diffusion, corresponding reverse SDE (Anderson 1982) conditioned on  $S$  (Ho and Salimans 2021) can be written as follows:

$$d\bar{X}_t = (f(\bar{X}_t, t) - g^2(t)\nabla_x \log p_t(\bar{X}_t|S))dt - g(t)d\bar{W}_t. \quad (3)$$

Song et al. (2020) showed that this reverse SDE solution is equivalent to the following reverse ODE:

$$d\bar{X}_t = (f(\bar{X}_t, t) - \frac{1}{2}g^2(t)\nabla_x \log p_t(\bar{X}_t|S))dt. \quad (4)$$

To learn the score function via neural networks  $\theta$ , diffusion models minimize the following loss:

$$\mathcal{L}(\theta) = \mathbb{E}_{X_t, t} [\|\psi_\theta(X_t, t, S) - \nabla_{X_t} \log p(X_t|S)\|_2^2], \quad (5)$$

where  $\psi_\theta(X_t, t, S)$  is a score function to learn.

**Recent Diffusion Sampling** A distinct advantage of the diffusion model lies in its capacity to generate data without needing extra conditioning architectures (Ho and Salimans 2021). Thus, previous studies have investigated exploring flexible guidance sampling, such as leveraging text embedding (Rombach et al. 2022; Ramesh et al. 2022) or designing new sampling methods using pre-trained models (Brack et al. 2023). Different sampling methods can vary outcomes in terms of both sampling quality (Humayun, Balestriero, and Baraniuk 2022) and speed (Lu et al. 2022).

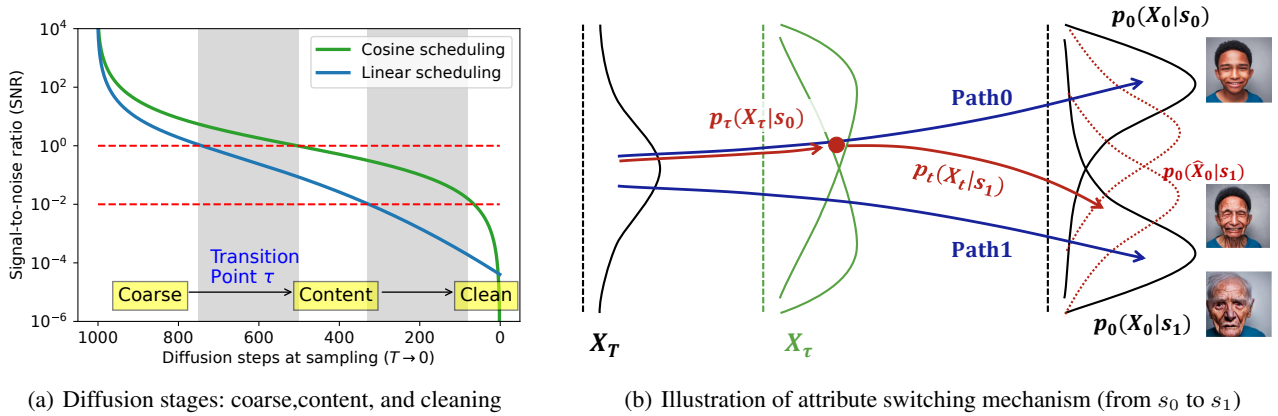


Figure 1: Motivation concerning diffusion learning stages and illustration of the proposed method. In transition point  $\tau$ , the proposed method switches the condition of sensitive attributes from  $s_0$  to  $s_1$  to generate synthetic data satisfying  $\epsilon$ -fairness.

### Fair Generative Model

When considering downstream fairness, a common approach involves incorporating fairness loss into the training objective. FairGAN(+) (Xu et al. 2018, 2019) and Fairness GAN (Sattigeri et al. 2019) focus on generating fair datasets composed of input and label pairs, enabling downstream classifiers to achieve fairness concerning sensitive attributes. Alternatively, DB-VAE (Amini et al. 2019) employs a de-biased classifier and adaptive resampling of rarer data during training to mitigate the potential biases related to downstream tasks rather than the generative task itself. Other lines of studies concentrate on the fairness notion related to imbalanced distributions of sensitive attributes. Choi et al. (2020) introduced fair generative modeling via weak supervision, utilizing a small, unbiased reference dataset to address imbalanced distribution biases.

For diffusion models, Sinha et al. (2021) addresses bias in datasets generated by unconditional diffusion through a few-shot conditional generation approach. Their goal is to achieve a balanced number of class labels. Similarly, Friedrich et al. (2023) achieves fairness by fair guidance, which aims to make a balanced number of class labels.

### Methodology

#### Problem Statement

Our objective is to achieve  $\epsilon$ -fairness within the generated data by leveraging the independence relationship  $\hat{X} \perp\!\!\!\perp S$  (Xu et al. 2018). Moreover, we aim to obtain  $p_G$  that ensures not only the distributional independence of sensitive attributes  $p_G(\hat{X}|S = 0) = p_G(\hat{X}|S = 1)$ , but also maintains utility with respect to the inherent data manifold. While the fairness comes from the independence of generated samples concerning the sensitive groups, the utility of  $p_G$  implies that the generated samples from  $p_G$  preserve the underlying manifold of original data distribution.

In this section, we demonstrate how our method can achieve these fair characteristics of  $p_G$  by introducing a new

mechanism to control the sensitive attributes during the sampling. For simplicity, we will omit the subscript  $G$  when referring to the distribution associated with the diffusion process and use the notation  $p(X|s) := p(X|S = s)$  when referring to the conditional distribution to attribute  $s$ . Note that we present the proposed method in a binary context; however, its applicability is broader, covering multi-nominal sensitive attributes or specified text-based conditions.

#### Sampling with Attribute Switching Mechanism

To make the two data distributions similar, it is well known that controlling high-level features (e.g., semantics, outlines, and image color tones) are important (Long et al. 2015; Ilyas et al. 2019). Recent findings indicate that diffusion models learn distinct attributes at each sampling step (Choi et al. 2022). Within the context of signal-to-noise ratio (SNR) for diffusion (Kingma et al. 2021), the sampling process involves three steps: (i) learning **coarse** features, (ii) generating rich **content**, and (iii) **cleaning** up residual noise, as depicted in Figure 1(a). Specifically, the coarse stage generates high-level features, while the content and clean stages handle fine detail generation (Kwon, Jeong, and Uh 2023).

Thus, through the transfer of high-level features during the sampling process, we might mitigate the inherent distributional disparity between the two sensitive attributes. This approach ensures that a sensitive attribute encompasses coarse features similar to those of the other attribute. To achieve this goal, we propose a general sampling framework to handle sensitive attributes, which we call *attribute switching*. The sampling process of the proposed method is graphically depicted in Figure 1(b). Simply, at the transition point  $\tau$ , we alter the condition of the diffusion model, transitioning it from the **initial sensitive attribute**  $s_0$  to the **switched sensitive attribute**  $s_1$ . By leveraging high-level features learned from the distribution of  $s_0$ , our method can generate images from  $p(\hat{X}_0|s_1)$ , which exhibits an independent relationship to the sensitive attribute. The detailed sampling procedure is explained in Algorithm 1.

**Algorithm 1: Fair Sampling with Attribute Switching**


---

**Input:** Pre-trained model  $\theta$ , subgroup of time steps for sampling  $\{0, k, \dots, T - k, T\}$ , transition point  $\tau$ , and sensitive attributes  $s_0, s_1$

**Output:** Generated fair images  $\hat{X}_0$

**for** a random batch **do**

**Initialize:**  $\hat{X}_T \sim N(0, I)$

**for**  $t = T - k, T - 2k, \dots, \tau - k$  **do**

Sampling under the condition of  $S = s_0$

$d\hat{X}_t = (f(\hat{X}_t, t) - \frac{1}{2}g^2(t)\psi_\theta(\hat{X}_t, t, s_0))dt$

**end**

**for**  $t = \tau, \dots, 0$  **do**

Switching: sampling under condition of  $S = s_1$

$d\hat{X}_t = (f(\hat{X}_t, t) - \frac{1}{2}g^2(t)\psi_\theta(\hat{X}_t, t, s_1))dt$

**end**

**end**

---

**Preserving the distribution of the ODE solution** We first argue that attribute switching still leads to generate the synthetic data samples on the same data manifold because  $p(X_\tau|s_0)$  and  $p(X_\tau|s_1)$  have the smooth noised distribution (Choi et al. 2022). In a more formal expression, the attribute switching is represented as:

$$S_t \sim p(S_t = s) \\ = p_D(S = s_0) \cdot \delta(s - s_0) \cdot \mathbb{1}(t > \tau) \\ + p_D(S = s_1) \cdot \delta(s - s_1) \cdot \mathbb{1}(t \leq \tau), \quad (6)$$

for  $s_0, s_1 \in \{0, 1\}$ ,  $s_0 \neq s_1$ , and  $\mathbb{1}$  denotes an indicator function. The vanilla sampling methods, such as DDPM or DDIM, perform sampling under  $p(S_t = s) = p_D(S = s)$ . For mathematical proof, we now demonstrate that the solution to the original SDE maintains an identical distribution to the SDE after attribute switching as follows:

**Theorem 1.** *Assuming a pre-trained model is trained with Equation (2), and the subsequent reverse ODE represents an equivalent probability flow ODE corresponding to the pre-trained model Equation (4). Then, the solution of following reverse ODE has the same distribution with the pre-trained ODE,*

$$d\hat{X}_t = (f(\hat{X}_t, t) - \frac{1}{2}g^2(t)\nabla_x \log p_t(\hat{X}_t|S_t))dt, \quad (7)$$

where  $S_t$  follows Equation (6).

*Proof. (Sketch of the proof)* Using the Fokker-Plank equation, we can show the solution of the given ODE has the same probabilistic as the original ODE. Refer to the Appendix for the details.  $\square$

Therefore, we do not need extra training and no adjustments are requisite for existing training methodologies. In the following section, we will elucidate how switching can obtain fairness and suggest a straightforward algorithm to find proper  $\tau$ .

**Algorithm 2: Fair Transition Point Searching**


---

**Input:** Pre-trained model  $\theta$ , sensitive attribute  $s_0, s_1$ , subgroup of time steps for sampling  $\{0, k, \dots, T - k, T\}$

**Output:** Transition point  $\tau$

**Initialize:**  $X_T \sim N(0, I)$

**for**  $t = T - k, T - 2k, \dots, 0$  **do**

/\* Single sampling for  $\tau$  search \*/

Calculate score  $\psi_\theta(X_t, t, s_0)$ ,  $\psi_\theta(X_t, t, s_1)$  using pre-trained model  $\theta$

Obtain diffusion coefficient  $g(t)$  in Equation (2)

Store difference between  $s_0$  and  $s_1$  using  $\theta$

$D_t = g^2(t)\{\psi_\theta(X_t, t, s_0) - \psi_\theta(X_t, t, s_1)\}$

**end**

Calculate the transition point  $\tau$

$$\tau = \arg \min_{\tau} \left\| \sum_{i \leq \tau} D_i - \sum_{i > \tau} D_i \right\|$$


---

**Fair Sampling with Attribute Switching**

As vanilla diffusion sampling fails to satisfy  $\epsilon$ -fairness, two data distributions conditioned on the sensitive attribute become distinct as denoising progresses from time  $T$  to 0. This is illustrated in Figure 1(b), where ‘Path0’ and ‘Path1’ diverge from each other. In contrast, our method facilitates attaining the distribution matching for  $p(\hat{X}_0|s_0)$  and  $p(\hat{X}_0|s_1)$  by switching the condition of the sensitive attribute of Equation (4) from  $s_0$  to  $s_1$  at the transition point  $\tau$ . For a detailed explanation, we now present a theoretical analysis aimed at identifying an optimal transition point  $\tau$  that guarantees fairness. We also provide theoretical support for the practicality of utilizing a pre-trained diffusion model.

**Fair Condition of Transition Point  $\tau$**  Our primary goal is to achieve a distribution of generated samples wherein  $p(\hat{X}|S = s_0) = p(\hat{X}|S = s_1)$ . Thus, we suggest the condition of transition point  $\tau$  to achieve the goal as follows:

**Theorem 2. (Fair condition of transition point  $\tau$ )** *Let  $\tau$  be a transition point satisfying the following condition:*

$$\int_0^\tau D(t)dt = \int_\tau^T D(t)dt. \quad (8)$$

where

$$D(t) = g^2(t)(\nabla_x \log p_t(\bar{X}_t|s_0) - \nabla_x \log p_t(\bar{X}_t|s_1)), \quad (9)$$

*Then, the generated distribution from the following reverse ODE becomes independent of the sensitive attribute, where  $S_t$  is from Equation (6).*

*Proof. (Sketch of the proof)* By using Equation (4), we can get the equation about the generated image from reverse ODE. Moreover, by Equation (6) and (7), we can get the equation about the switched image. Refer to the Appendix for the details.  $\square$

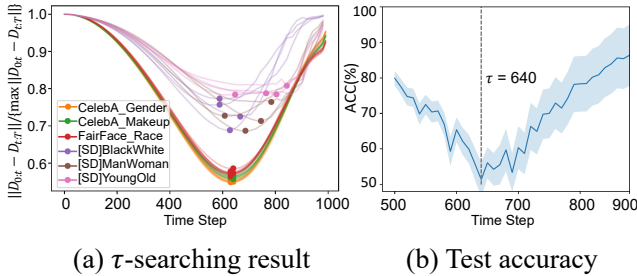


Figure 2:  $\tau$ -search algorithm and its efficiency. (a)  $\tau$ -searching results with diverse diffusion models and datasets and (b) test accuracy of true data evaluated with models trained using synthetic data with varying  $\tau$  values.

The above theorem states that the attribute switching guarantees fairness of the generated data distribution, by ensuring the independence of the sensitive attribute when  $\tau$  satisfies Equation (8).

**Empirical Validation of Transition Points  $\tau$**  To guarantee fairness in diffusion sampling according to Theorem 2, our goal is to find  $\tau$  that minimizes the following objective:

$$\left\| \sum_{i \leq \tau} D_i - \sum_{i > \tau} D_i \right\|. \quad (10)$$

Thus, we first introduce an efficient algorithm to determine an appropriate  $\tau$  in Algorithm 2. Within a mini-batch setting, we store the estimated value of the difference between two sensitive attributes in Equation (9) and estimate the score in Equation (5). The optimal  $\tau$  that minimizes Equation (10) can be found by performing a one-time sampling with cumulative sum calculation for a given mini-batch. To verify whether the selection of  $\tau$  from a single mini-batch can be applied to the entire sampling process, we illustrated the maximum scaled values of Equation (10), varying the model and dataset in Figure 2(a). The top three correspond to experiments conducted with a batch size of 256, while the remaining three at the bottom represent experiments with a batch size of 5. For a large batch size, individual computations of differences for each mini-batch are unnecessary; a single batch calculation suffices for determining  $\tau$  for a given dataset. More detailed experiments with diverse batch sizes are illustrated in the Appendix. Note that the function  $g(t)$  in Equation (9) might be differently defined with the scheduler and sampler (Ho, Jain, and Abbeel 2020; Karras et al. 2022).

To verify the optimal transition point  $\tau$  from Algorithm 2, we plot the accuracy of a model trained with the synthetic data and test with the original FairFace (Karkkainen and Joo 2021) “race” (black-white) dataset. In Figure 2(b), we observe that the theoretically derived value of  $\tau = 640$ , which ensures fairness, closely aligns with the  $\tau$  value that exhibits fairness in practice. Interestingly, this value is highly aligned to their empirical boundaries of coarse-content phases ( $\approx \text{SNR} = 10^0$ ) (Choi et al. 2022) in Figure 1(a). These empirical results highly support transferring high-level features of  $s_0$  to content phases of  $s_1$  is effective for fair sampling.

Classifier	Methods	$S = 0$	$S = 1$	gap	BER
	Real	9.51	4.89	4.62	7.20
Syn (Tr)	Vanilla	8.80	8.20	0.60	8.50
	Mixing	39.59	43.93	4.34	41.76
Orig (Te)	Editing	9.58	18.03	8.46	14.42
	Ours	54.63	54.92	0.29	<b>54.78</b>
Orig (Tr)	Vanilla	19.59	10.68	8.91	15.14
	Mixing	62.64	20.55	42.09	41.60
Syn (Te)	Editing	31.56	10.92	20.64	21.24
	Ours	62.59	38.86	23.73	<b>50.73</b>

Table 2: Error rates for classifier (%), trained (Tr) with synthetic data (Syn) generated from each sampling method and tested (Te) with original data (Orig), and vice versa.

For complexity analysis, let  $N_B = \lceil N/\text{batch\_size} \rceil$  be the number of batch samplings for  $N$  data samples. The attribute switching sampling in Algorithm 1 requires  $O(N_B \times T)$  for  $T$  denoising steps, the same as vanilla sampling since switching only requires `if-else` without additional space complexity. For  $\tau$  searching in Algorithm 2, we only need one-time mini-batch sampling of  $O(T)$ , which is marginal to fair sampling of  $O(N_B \times T)$ . Only for Algorithm 2, we require twice the memory space for sampling both sensitive attributes and comparing their step-wise differences.

## Experiments

### Experimental Setups

The success of diffusion models is attributed to their ability to provide effective guidance in a manifold, performing well even under linear interpolation (Ramesh et al. 2022). As our approach represents the first attempt to investigate fairness by matching the distribution given sensitive attributes in diffusion models, the most straightforward method for comparison is mixing conditional embeddings (Song, Meng, and Ermon 2020). Specifically, we linearly combine the sensitive attribute embeddings for sampling, utilizing the embedding of  $s_1$  with a probability  $p > 0.5$ , and that of  $s_0$  with a probability of  $1 - p$ . As another baseline, we employ DiffEdit (Couairon et al. 2022), a state-of-the-art image diffusion-based editing method. We generate images using the original sampling method of diffusion models and edit them by subtracting corresponding text for  $s_0$  and adding corresponding text for  $s_1$ .

Following Xu et al. (2018), we employ a classifier trained on synthetic data and evaluate it on the original dataset to measure fairness. Specifically, we train a ResNet18 (He et al. 2016) model as a classifier to predict the sensitive attribute  $S$ , achieving a training accuracy of 100% in most cases. For datasets, we employ FairFace, CelebA (Liu et al. 2018) to quantify data fairness and utility, and CIFAR-10 (Krizhevsky, Hinton et al. 2009) for dimension reduction experiments. Regarding evaluation metrics, we assess classification performance using accuracy, data  $\epsilon$ -fairness with BER as defined in Equation (1), and data utility with the Frechet Inception Distance (FID) score. All experiments are conducted using PyTorch-based diffusion libraries (von Platen et al. 2022). Additional details regarding the experimental setup can be found in the Appendix and the code is

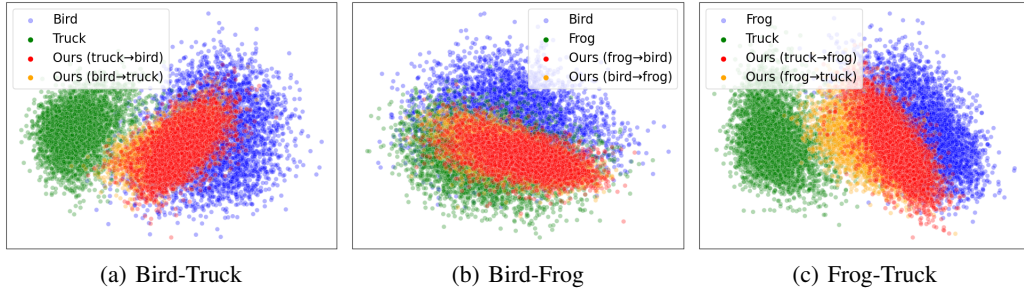


Figure 3: Dimension reduction (PCA,  $n = 2$ ) results of the data distribution for both the true data and the generated data of the proposed methods for Bird-Truck, Bird-Frog, and Frog-Truck pairs within the CIFAR-10 datasets.

	Methods	FairFace Race (12k)				CelebA Male (12k)				CelebA Heavy-Makeup (12k)			
		$S = 0$	$S = 1$	total	gap	$S = 0$	$S = 1$	total	gap	$S = 0$	$S = 1$	total	gap
Total	Real (Test)	10.12*	8.77*	6.35*	1.34*	21.66	21.25	2.50	0.41	12.92	18.88	2.43	5.96
	Vanilla (Total)	29.60	34.99	29.50	5.40	22.24	19.75	7.19	2.49	19.75	22.24	7.19	2.49
	Ours (Total)	31.15	36.23	31.11	5.08	23.17	29.20	7.83	6.03	20.35	21.36	7.59	1.01
$s_0 = 0$ $s_1 = 1$	Real ( $s_0$ )	21.47*	33.81*	17.43*	12.33*	2.36	68.11	22.81	65.75	2.64	53.13	13.78	50.49
	Vanilla ( $s_0$ )	28.04	40.59	31.78	12.55	7.04	66.00	23.03	58.96	8.75	47.37	14.50	38.62
	Ours ( $s_0 \rightarrow s_1$ )	31.05	38.29	32.15	7.25	8.50	58.54	18.54	50.05	10.87	39.80	11.72	28.92
$s_0 = 1$ $s_1 = 0$	Real ( $s_0$ )	29.68*	18.19*	13.05*	11.49*	53.25	2.12	20.02	51.13	53.25	2.12	20.02	51.13
	Vanilla ( $s_0$ )	33.77	32.53	30.59	1.24	58.31	9.14	19.12	49.17	49.38	6.66	19.03	42.72
	Ours ( $s_0 \rightarrow s_1$ )	32.08	35.57	31.26	3.48	50.03	12.00	15.64	38.03	42.03	8.44	15.44	33.60

Table 3: FID comparison of generated images with a fixed start point for FairFace and CelebA.

available at <https://github.com/uzn36/AttributeSwitching>.

### Data Fairness

Table 2 presents the evaluation of error rates for the original and synthetic data, generated using the same diffusion model trained on the FairFace dataset, with different sampling methods: vanilla, mixing, editing, and attribute switching. The first column indicates the training and evaluating data, e.g., “Orig (Tr)  $\rightarrow$  Syn (Te)” represents the classifier trained on the original dataset and evaluated on synthetic data. “Real” indicates training and testing with the original dataset. The values are presented as error rates in percentage, where the column for  $S = s$  means  $P(f(X) = 1 - s | S = s) \cdot 100$  (%) for the classifier  $f$ , and BER is obtained by Equation (1). Notably, we achieved fairness in terms of both the gap and BER, for both “Orig (Tr)  $\rightarrow$  Syn (Te)” and “Syn (Tr)  $\rightarrow$  Orig (Te)” evaluations. Our approach achieved a BER similar to 50% while mixing and editing only achieved 42%, 14%, respectively. Moreover, the classifier trained with the original dataset tends to classify the generated data more often as  $S = 1$ , while this tendency is alleviated in attribute switching. This conclusion demonstrates that attribute switching has successfully achieved fairness in terms of  $\epsilon$ -fairness.

Another method to evaluate the similarity of the generated data distribution is through dimension reduction and visualization. To confirm that the proposed methods generate fair data, we performed dimension reduction (PCA) to dimension  $n = 2$  on the pre-trained ResNet embedding space. To ensure distinct distributions for each sensitive attribute, we extracted three different classes from the CIFAR-10 dataset:

bird, truck, and frog. Each class is considered a sensitive attribute, and then, we applied our method to generate distributionally equivalent images for all possible pairs. For pairwise distribution analysis, we calculated principal component vectors using true data distribution pairs. Subsequently, we projected the generated data onto these principal dimensions. The results are depicted in Figure 3. The generated data is placed in between the two true distributions while sharing similar distributions between attributes.

### Data utility

To assess data utility, we compute the FID using the training dataset and an equal number of generated samples. It’s worth noting that the FID score is highly sensitive to the dataset size, so we aimed to maintain a consistent number of samples. However, due to limitations in the FairFace dataset, we encountered insufficient data when calculating the FID with the real test data. In cases where the dataset sizes did not align, we have marked this discrepancy with an asterisk (\*).

As our method switches the attribute from  $s_0$  to  $s_1$  and labels the sensitive attribute as  $s_1$ , we conducted a comparison of FID with a fixed start point. The results are presented in Table 3. We measure the FID with a total of 12k train data, 12k train data with  $S = 0$ , and 12k train data with  $S = 1$ . As shown in the table, the total FID of attribute switching is similar to that of vanilla sampling. This suggests that the total data distribution of generated data is similar to that of vanilla data, which can ensure the data utility. Notably, the FID of the proposed method is worse than vanilla sampling, in terms of  $s_0$ . This is expected since the attribute of generated data is changed to  $s_1$ . Most importantly, the FID with

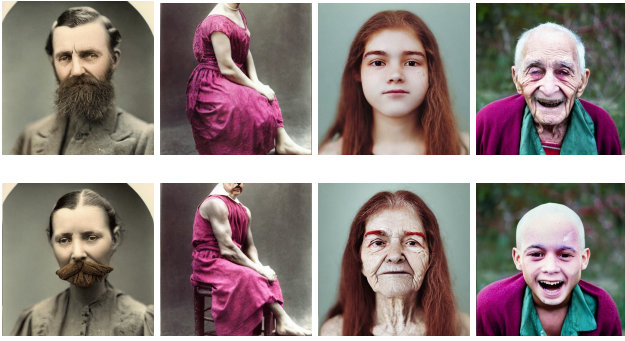


Figure 4: Sampling from stable diffusion with (Top) vanilla sampling and (Bottom) switching sensitive attribute by the proposed method. After switching, the uncontrolled features such as hairstyle, dress, and mustache are maintained.

Acc (%)	Smile ( $\uparrow$ )	Male (-)		Smile	Male
Vanilla	88.38	93.04	Ours	88.14	55.71

Table 4: Classification accuracy trained with synthetic data.

$s_1$  is significantly better even when the gap between the two attributes is reduced. Given that FID measures the difference between two distributions, we can conclude that this result pertains not only to data utility but also to fairness considerations. This finding further supports the notion that attribute switching contributes to making the distribution of generated images more similar.

Attribute switching can also ensure data utility for downstream tasks, particularly in training classifiers where sensitive attributes, such as “Male,” are removed. For example, consider the “Smiling” classification task with the CelebA dataset. We conditioned a diffusion model on both the class “Smiling” and the sensitive attribute “Male”. The classification results are in Table 4. The classifier trained with synthetic data from vanilla sampling achieved high test accuracy in both “Smiling” and “Male”, indicating both attributes are predictable by the synthetic data. In contrast, attribute switching maintains “Smiling” accuracy while achieving fair “Male” accuracy, close to 50%. These results indicate that attribute switching only modifies the sensitive attribute, without significantly affecting other attributes. This implies that attribute switching can be employed across various downstream tasks without raising fairness concerns.

### Sampling with Text-conditioning Models

As we mathematically prove in Theorem 1, we now demonstrate that the proposed method can be easily applied to any pre-trained model, regardless of whether the model is conditioned with multi-class labels or text information, just as (Balaji et al. 2022). We used widely used pre-trained diffusion models with text conditional embedding, i.e., stable diffusion models (Rombach et al. 2022). For the sensitive attributes  $s_0$  and  $s_1$ , we use guided text to generate images as follows: from “a color photo of the face of  $s_0$ ” to “a color photo of the face of  $s_1$ ”. Fig-



Figure 5: Generated pairwise images with the same random seeds. In each row, images with red borderlines denote vanilla sampling images. Beginning with these vanilla images, we progressively increase the value of  $\tau$ , moving further away from the initial vanilla image.

ure 4 illustrates the generated images using text conditioning: {“man”, “woman”} and {“young person”, “old person”} with  $\tau$  determined by averaging four results using a batch size of 5, as in Figure 2(a). In Figure 4, each image sampled with a switching mechanism has characteristics that were not present with the vanilla sampling generated (e.g., a woman with a mustache and a man with a pink dress). More generated images of Figures with different text guidance and results of mixing embedding can be found in the Appendix.

Figure 5 shows pair-wise images with same random seed but altering the  $s_0 \rightarrow s_1$  and  $s_1 \rightarrow s_0$  during sampling. We use the text condition of {“young person”, “old person”} and plot  $\tau \in \{500, 600, 700, 800, 1000\}$ . By choosing the initial sensitive condition  $s_0$  and switching condition  $s_1$ , we can control what property would be retained and what property would be modified. For example, the top image pairs show similar ages of intermediate images, while  $s_0 \rightarrow s_1$  and  $s_1 \rightarrow s_0$  keep female and male properties, respectively.

### Conclusion & Limitations

In this paper, we investigated the fair sampling in diffusion models, focusing on  $\epsilon$ -fairness to generate distributionally similar images. To address the limitations of existing fairness frameworks, we present a novel switching sampling method that generates data satisfying both fairness and utility. We theoretically prove and experimentally show the effectiveness of the proposed methods in various settings, including pre-trained diffusion models with text embeddings.

While attribute switching preserves high-level features of images, it allows to generate the images with similar distributions. Therefore, it can generate images that were previously challenging, such as bald women. However, elements with significant contextual impact tend to be removed during the process, which should be addressed later carefully.

## Ethics Statement

This study focuses on addressing fairness issues in generative models, particularly in the context of image generation using diffusion models. Generative models have a broad spectrum of applications ranging from data augmentation and synthetic data creation for privacy preservation to generating novel content for creative and entertainment purposes. However, there is an inherent risk that these models may perpetuate or even exacerbate fairness issues, such as reinforcing stereotypes or introducing biases in their applications.

Our proposed method aims to enable pre-trained generative models to produce images, leveraging similar distributions across the groups defined by sensitive attributes such as race or gender. This approach avoids the need for additional classifier training, making it computationally efficient and readily applicable to existing models.

However, it's important to note that our study primarily focuses on distributional fairness, one facet of the multifaceted concept of fairness in generative models. Moreover, many existing studies primarily explore fairness within specific downstream tasks, neglecting broader considerations. A comprehensive understanding of fairness in generative models demands the counterfactual investigation of multiple factors but has been somewhat unexplored.

We hope that our work will highlight the importance of evaluating and addressing fairness issues in generative models more comprehensively. Future research should expand upon this foundation, exploring various fairness factors in depth to ensure more equitable and unbiased applications of generative AI technologies.

## Acknowledgements

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. RS-2023-00272502, No. NRF-2022R1F1A1065171, No. 2022R1A5A6000840) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2020-0-01336, Artificial Intelligence graduate school support (UNIST)).

## References

- Amini, A.; Soleimany, A. P.; Schwarting, W.; Bhatia, S. N.; and Rus, D. 2019. Uncovering and mitigating algorithmic bias through learned latent structure. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 289–295.
- Anderson, B. D. 1982. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3): 313–326.
- Balaji, Y.; Nah, S.; Huang, X.; Vahdat, A.; Song, J.; Kreis, K.; Aittala, M.; Aila, T.; Laine, S.; Catanzaro, B.; et al. 2022. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*.
- Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S. W.; Fidler, S.; and Kreis, K. 2023. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22563–22575.
- Brack, M.; Friedrich, F.; Hintersdorf, D.; Struppek, L.; Schramowski, P.; and Kersting, K. 2023. Sega: Instructing diffusion using semantic dimensions. *arXiv preprint arXiv:2301.12247*.
- Choi, J.; Lee, J.; Shin, C.; Kim, S.; Kim, H.; and Yoon, S. 2022. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11472–11481.
- Choi, K.; Grover, A.; Singh, T.; Shu, R.; and Ermon, S. 2020. Fair generative modeling via weak supervision. In *International Conference on Machine Learning*, 1887–1898. PMLR.
- Couairon, G.; Verbeek, J.; Schwenk, H.; and Cord, M. 2022. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*.
- Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.
- Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.
- Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 259–268.
- Friedrich, F.; Schramowski, P.; Brack, M.; Struppek, L.; Hintersdorf, D.; Luccioni, S.; and Kersting, K. 2023. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*.
- Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.
- Ho, J.; and Salimans, T. 2021. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.
- Humayun, A. I.; Balestriero, R.; and Baraniuk, R. 2022. Polarity sampling: Quality and diversity control of pre-trained generative networks via singular values. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10641–10650.
- Ilyas, A.; Santurkar, S.; Tsipras, D.; Engstrom, L.; Tran, B.; and Madry, A. 2019. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32.

- Karkkainen, K.; and Joo, J. 2021. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1548–1558.
- Karras, T.; Aittala, M.; Aila, T.; and Laine, S. 2022. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35: 26565–26577.
- Kenack, P. J.; Arapov, D. D.; Hussain, R.; Kazmi, S. A.; and Khan, A. 2021. On the fairness of generative adversarial networks (gans). In *2021 International Conference "Nonlinearity, Information and Robotics" (NIR)*, 1–7. IEEE.
- Kingma, D.; Salimans, T.; Poole, B.; and Ho, J. 2021. Variational diffusion models. *Advances in neural information processing systems*, 34: 21696–21707.
- Kong, Z.; Ping, W.; Huang, J.; Zhao, K.; and Catanzaro, B. 2020. DiffWave: A Versatile Diffusion Model for Audio Synthesis. In *International Conference on Learning Representations*.
- Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.
- Kwon, M.; Jeong, J.; and Uh, Y. 2023. Diffusion Models Already Have A Semantic Latent Space. In *The Eleventh International Conference on Learning Representations*.
- Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2018. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018): 11.
- Long, M.; Cao, Y.; Wang, J.; and Jordan, M. 2015. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, 97–105. PMLR.
- Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35: 5775–5787.
- Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.
- Sattigeri, P.; Hoffman, S. C.; Chenthamarakshan, V.; and Varshney, K. R. 2019. Fairness GAN: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development*, 63(4/5): 3–1.
- Sinha, A.; Song, J.; Meng, C.; and Ermon, S. 2021. D2c: Diffusion-decoding models for few-shot conditional generation. *Advances in Neural Information Processing Systems*, 34: 12533–12548.
- Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.
- Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.
- Verma, S.; and Rubin, J. 2018. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, 1–7.
- von Platen, P.; Patil, S.; Lozhkov, A.; Cuenca, P.; Lambert, N.; Rasul, K.; Davaadorj, M.; and Wolf, T. 2022. Diffusers: State-of-the-art diffusion models. <https://github.com/huggingface/diffusers>.
- Xu, D.; Yuan, S.; Zhang, L.; and Wu, X. 2018. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, 570–575. IEEE.
- Xu, D.; Yuan, S.; Zhang, L.; and Wu, X. 2019. Fairgan+: Achieving fair data generation and classification through generative adversarial nets. In *2019 IEEE International Conference on Big Data (Big Data)*, 1401–1406. IEEE.