Fair Latent Deep Generative Models (FLDGM) for Syntax-agnostic and Fair Synthetic Data Generation

Anonymous Authors[1]

Abstract

Deep Generative Models (DGM) for generating synthetic data with properties such as quality, diversity, fidelity, and privacy is an important research topic. Fairness is one particular aspect that has not received the attention it deserves. One difficulty is training DGM with an in-process fairness objective, which can disturb the global convergence characteristics. To address this, we propose Fair Latent Deep Generative Models (FLDGM) as enablers for more flexible and stable training of fair DGMs, by first learning a syntax-agnostic, model-agnostic fair latent vector representation of the data. This separates the fairness optimization and data generation processes thereby boosting stability and optimization performance. We conduct extensive experiments on image and tabular domains using Generative Adversarial Networks (GANs) and Diffusion Models (DM) and compare them to the state-of-the-art in terms of fairness and utility. Our proposed FLDGM achieve superior performance in generating high-quality, high-fidelity, and high-diversity fair synthetic data compared to the state-of-the-art fair generative models.

1. Introduction

Deep Generative Models (DGM) have achieved substantial progress in learning to approximate the real data distribution as closely as possible. In particular, Generative Adversarial Networks (GANs) (Goodfellow et al., 2020) and Diffusion Models (DM) (Ho et al., 2020) are the most successful among the generative models for generating high-dimensional data. Existing generation methods based on GANs and DMs have focused on properties such as fidelity, quality, diversity, and privacy. The fidelity and quality relate to how closely synthetic data captures the distribution of real data. The diversity measures how successful in generating new distribution that is covered by the real data. And, finally, privacy guarantees that synthetic data is not just a replication of real data, which is very important in sensitive domains (Xie et al., 2018).

Synthetic data fairness - generating fair data from biased data, is a much less-explored concept in the context of generative models. A few solutions to this problem such as FairGAN (Xu et al., 2018) and DECAF (van Breugel et al., 2021) have been proposed based on GANs to ensure fairness in the downstream tasks. A recent study shows that existing GAN techniques amplify the bias present in the training data resulting in more biased data in the target (Gupta et al., 2021) including differential privacy generation schemes. Therefore, protected groups or people with certain sensitive or protected characteristics like ethnicity, gender, or religion (Binns, 2018), can have biased treatments in downstream models. As a result, safeguarding against discrimination—or unfavorable outcomes as a result of a person's protected qualities (Mehrabi et al., 2021)—has become more crucial in ML.

Motivation. State-of-the-art fair generative models (Xu et al., 2018; van Breugel et al., 2021) operate on pixel (image) or attribute levels (tabular) and are highly dependent on the underlying syntax and model architectures in the high dimensional space. Also, altering training with an in-process fairness objective may disturb the quality-fairness trade-off. Reducing the computational overheads of fair DGMs without sacrificing their quality is the key to promote their accessibility.

Research gap. There is a lack of study in learning fair DGM to reach an optimal point between accessibility, fairness, quality, and flexibility (fine-tuning to various architectures, tasks, and fairness measures).

To this end, we propose Fair Latent Deep Generative Models (FLDGM), both for GANs and DMs. our FLDGM are syntax-agnostic, stable, and operate on low-dimensional continuous latent space. First, our approach starts with learning a fair compression using autoencoders that enables fast sampling from the input domain and encourages quality

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

in the target as the DGMs in the subsequent stage can focus on optimizing this compressed dimension. Second, the fair latent vectors can be used for various generative models (such as many versions of GANs and DMs) and applications independent of data-specific architectures, which makes the approach more generalizable. Third, it can be extended to impose various fairness constraints in the synthetic data given the pre-trained autoencoders for the corresponding fairness measures. Fourth, since the generation is performed by either GANS or DMs, it can produce high-quality samples which contradict the approach described in (Louizos et al., 2015). Finally, the transformation from the generated fair latent space to the fair data space can be done in a single pass. To the best of our knowledge, there are no studies involving DMs in fairness optimization.

Contributions. Our key contributions are four-fold: (i) We propose a novel formulation of a fair latent generative framework common for both GANs and Diffusion models. We show how this can be fine-tuned across various generative architectures with less computational overhead; (ii) In contrast to previous works (Xu et al., 2018; 2019; Louizos et al., 2016) which generate both fair and accurate synthetic data simultaneously, FLDGM does not require a delicate weighting factor of generation quality and fairness penalty. Therefore, our approach requires zero regularization of the latent space and ensures high-fidelity reconstructions and global convergence guarantee; (iii) The FLDGM can be generalizable to data of any category, reducing the data preprocessing and modeling overhead in DGM; (iv) Finally, we conduct extensive experiments on tabular and image domains for various generative frameworks and compare the performance to the state-of-the-art in terms of fairness and data utility. Moreover, we also analyze the fidelity, diversity, and authenticity (Alaa et al., 2022) of our proposed FLDGM.

2. Preliminaries

2.1. Algorithmic Fairness

In this section, we define disparate treatment and disparate impact measures of algorithmic fairness. Given a biased dataset $\mathbf{D} = \{X, S, Y\}$, where $X \in \mathcal{X}$, $S \in \mathcal{S}$, and $Y \in \mathcal{Y}$ respectively denote the set of non-sensitive, sensitive and target attributes. The features S and Y are categorical.

Definition 1: Fairness Through Unawareness (FTU) (Grgic-Hlaca et al., 2016) - Let h be a prediction function, $h : X \rightarrow \hat{Y}$, and $\hat{Y}$ be the prediction outcome. The function h satisfies FTU if the sensitive attributes S are not explicitly used by h to obtain $\hat{Y}$.

The above definition controls disparate treatment (Zafar et al., 2017; Barocas & Selbst, 2016), but it is susceptible to disparate impact (Feldman et al., 2015), which is caused by the proxy features that highly correlate with S. Therefore, a stronger measure is needed to control indirect discrimination, which is achieved by Demographic Parity (DP) or statistical parity.

Definition 2: Demographic Parity (DP) (Barocas & Selbst, 2016) - Suppose we have a function $f : X \rightarrow \hat{Y}, \hat{Y} = \{0, 1\}$ for binary classification, and let S splits X into a majority set $\mathcal{M}$ and a minority set $\mathcal{M}'$ ($X = \mathcal{M} \cup \mathcal{M}'$), the function f satisfies DP if $P[f(x) = 1 \mid x \in \mathcal{M}] = P\,[f(x) = 1 \mid x \in \mathcal{M}']$, where x denotes an instance of X and $P[.]$ denotes the probability of an instance. We assume the protected attribute is binary for notational convenience and can be extended to non-binary settings as well.

2.2. Fairness Objective

Most of the state-of-the-art techniques for fairness penalty computations depend on mutual information-related measures (Rodríguez-Gálvez et al., 2021; Nan & Tao, 2020; Moyer et al., 2018). These information-theoretic methods achieve fairness at the expense of data quality and utility. Another line of research is based on adversarial approaches (Madras et al., 2018; Feng et al., 2019) but it suffers from training instability since an adversary cannot be completely trained until convergence in most situations (Moyer et al., 2018). To tackle these issues, we choose distance correlation (or distance covariance) similar to (Liu et al., 2022; Guo et al., 2022) and obtain a compressed, continuous, and fair representation Z of data $\mathbf{D}$ using autoencoders. The independence between latent space Z and sensitive attribute S can be modeled as minimizing the distance correlation, $\mathcal{V}^2$ between them (Liu et al., 2022) parameterized by ϕ (more details in Appendix A).

$$\mathcal{V}^2_{\phi}(z, s) = \int_{\mathcal{Z}} \int_{\mathcal{S}} \mid p_{\phi}(z, s) - p_{\phi}(z) p_(s) \mid^2 \; dz \; ds. \quad (1)$$

2.3. Generative Models

GAN-based generation. We use two GAN architectures, namely Least Square GAN (Mao et al., 2017) and Wasserstein GAN with Gradient Penalty (WGAN-GP) (Gulrajani et al., 2017) as these are the best among the state-of-the-art GAN-based generation methods.

The min-max optimization of LSGAN on fair latent vector Z can be defined as (Mao et al., 2017):

$$\begin{aligned}\min_{\theta_D} V_{LSGAN}(D) = &-\frac{1}{2} \times \mathbb{E}_{z \sim p_z(Z)}\big[(D(z) - b)^2\big] + \\ &\frac{1}{2} \times \mathbb{E}_{\xi \sim p_\xi(\xi)}[(D(G(\xi)) - a)^2],\end{aligned} \quad (2)$$

$$\min_{\theta_G} V_{LSGAN}(G) = \frac{1}{2} \times \mathbb{E}_{\xi \sim p_\xi(\xi)}[(D(G(\xi)) - c)^2], \quad (3)$$

where a and b are labels for fake data and real data respectively and c denotes the value that the generator wants D to believe for fake data. Also, ξ is from a uniform or Gaussian distribution $p_{\xi}(\xi)$ that maps ξ to the real data space (fair latent) through $G(\xi, \theta_G)$. The objective function for the latent vector generation of WGAN-GP is given in Appendix A.

Diffusion-based generation. A Diffusion Model (DM) (Ho et al., 2020) consists of a forward process, in which the data, Z (latent vector in our case) is progressively noised, and a reverse process, in which noise is transformed back into data from the target distribution.

The sampling chain transitions in the forward process can be set to conditional Gaussians and the Markov assumption of the forward process can be defined as (Ho et al., 2020):

$$\begin{aligned}\mathbf{q}(Z_{1:T}Z_0) &:= \prod_{t=1}^{T} \mathbf{q}(Z_t Z_{t-1}) \\ &:= \prod_{t=1}^{T} \mathcal{N}(Z_t; \sqrt{1-\beta_t} Z_{t-1}, \beta_t \mathbf{I}),\end{aligned} \tag{4}$$

where $\beta 1, \ldots, \beta T$ is the variance schedule. During the reverse process, the models learn to generate new data starting with the Gaussian noise $\mathbf{p}(Z_T) := \mathcal{N}(Z_T, \mathbf{0}, \mathbf{I})$, to the joint distribution $\mathbf{p}_{\theta}(Z_{0:T})$ as (Ho et al., 2020):

$$\begin{aligned}\mathbf{p}_{\theta}(Z_{0:T}) :=& \mathbf{p}(Z_T) \prod_{t=1}^{T} \mathbf{p}_{\theta}(Z_{t-1} Z_t) \\ &:= \mathbf{p}(Z_T) \prod_{t=1}^{T} \mathcal{N}(Z_{t-1}; \boldsymbol{\mu}_{\theta}(Z_t, t), \boldsymbol{\Sigma}_{\theta}(Z_t, t)),\end{aligned} \tag{5}$$

where the time-dependent parameters of the Gaussian transitions are learned.

3. Synthetic Data Fairness

Synthetic data fairness means generating fair synthetic data from biased data so that the downstream models trained on fair synthetic data will have fair predictions in real data [1]. In synthetic data generation, we consider $Y \in X$ if not explicitly defined.

Given a biased dataset $\mathbf{D} = \{X, S\}$, where $X \in \mathcal{X}$ and $S \in \mathcal{S}$ respectively denote the set of non-sensitive and sensitive attributes. We define a Fair Data Generation Process (FDGP) as follows:

Definition 3: Fair Data Generation Process (FDGP) - Let $\mathcal{G}$ be a generative model and $\mathcal{U}(S, Y)$ (either FTU or DP) be a definition of algorithmic fairness. The DGP is said to be fair if the $\mathcal{G}$, once optimized, learned to obtain a deterministic transformation from Multivariate Normal Distribution (MVN) to real data distribution that is maximally discriminative with respect to any downstream predictions but invariant to S, evaluated by $\mathcal{U}(S, Y)$.

Definition 4: Synthetic Data Fairness Problem (SDFP) - The Synthetic Data Fairness problem is to generate fair data $\mathbf{D}'$ from biased data $\mathbf{D}$ through Fair Data Generation Process (FDGP).

In summary, we learn to generate a distribution $p(\mathbf{D}')$ from $p(\mathbf{D})$ by removing the direct and indirect effects of malignant feature S (including proxy attributes[2]), so that the $p(\mathbf{D}')$ can be used for any downstream fair ($\mathcal{U}(S, Y)$ - *fair*) prediction tasks.

4. Fair Latent Deep Generative Models (FLDGM)

In our proposed FLDGM, the notion of FDGP is achieved by applying a sequence of operations in the biased data $\mathbf{D}$, which transforms the distribution $p(\mathbf{D})$ to $p(\mathbf{D}')$. This involves separating fairness optimization from data generation while maintaining quality, fairness, and diversity in the target, with the sub-goal of syntax and model-agnostic architectures. The generative models in FLDGM operate on a comparatively lower dimensional space than that of real data. The framework of FLDGM can thus be divided into a sequence of three stages:

1. Compressing the real data $\mathbf{D}$ into a fair representation retaining all the necessary information for any target tasks, called fair abstract compression. The output of this stage is low-dimensional fair latent continuous vectors without having any malignant information about sensitive features;
2. Fair latent vector generation, where the generative models are trained to generate high-quality fair latent vectors without focusing on syntax-related information;
3. A high fidelity reconstruction, where the data $\mathbf{D}'$ is reconstructed from the generated fair latent vectors in a single pass.

An outline of our proposed work is given in Figure 1.

4.1. Fair Abstract Compression

Removing undesired variations from data can be considered as a general compression model which relays on two sources,

[1] We assume that the model does not exhibit explicit biases and the biased outcome is caused only by the biases in the training data.

[2] features that are highly correlated with S

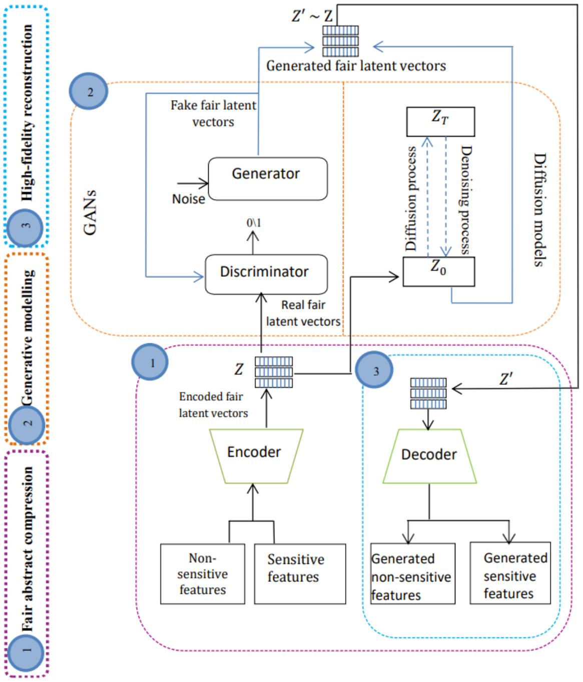

Figure 1. The proposed FLDGM architecture.

a sensitive variable S, which denotes the nuisance we want to remove, and a continuous latent vector Z which models all the remaining information from input. This fair abstract compression stage can use any of the state-of-the-art fair representation learning methods (Zemel et al., 2013; Feng et al., 2019; Oneto et al., 2020; Creager et al., 2019; Liu et al., 2022; Madras et al., 2018; Guo et al., 2022; Rodríguez-Gálvez et al., 2021; Nan & Tao, 2020; Moyer et al., 2018) that consist of an autoencoder trained for the combination of fairness loss and reconstruction loss. The choice of fairness objective in this step depends on different target fairness constraints. This ensures that the fairness optimization has a clear boundary on the attribute space, which is essential as fairness constraints are mostly defined on the independence criteria of attributes.

Formally, given an input instance $\{(x_i, s_i)\}_{i=1}^N$, where N is the number of instances in the data, the encoder $\mathcal{E}$ in the fair abstract compression stage encodes $\{(x_i, s_i)\}_{i=1}^N$ into a continuous latent space $z \in$ Z, where $z = \mathcal{E}(x_i, s_i)$. This latent representation Z factors out all the undesired variations in the data about $s \in S \in \mathcal{S}$ and captures the remaining information for any tasks. The decoder $\mathcal{D}$ associated with the autoencoder is now responsible for reconstructing the original data points from the fair latent space, resulting $\mathcal{D}(z) = \mathcal{D}(\mathcal{E}((x_i, s_i)))$. The fair compression process can be formally defined by :

$$\mathcal{E}_{\phi_\mathcal{E}}(Z|X, S); \mathcal{D}_{\phi_\mathcal{D}}(X'|Z, S)), \tag{6}$$

where ϕ is the parameter for autoencoder and $X \sim X'$. A multivariate Gaussian has been used for posterior $\mathcal{E}_{\phi_\mathcal{E}}(Z|X, S) = \mathcal{N}_{\phi_\mathcal{E}}(Z; \mu, \sigma)$, and a standard multivariate Gaussian $\mathcal{N}(0, I)$ for $p(Z)$.

Therefore, the objective function for final fair abstract compression can be defined as a combination of reconstruction loss and fairness loss from Eq. (1) as (Liu et al., 2022):

$$\max_{\phi_\mathcal{E}\phi_\mathcal{D}} \{\log p_{\phi_\mathcal{D}(x|s)} - \alpha \mathcal{V}_\phi^2(z, s)\}, \tag{7}$$

where α is a hyperparameter. We analyze the performance of different values of α, $\alpha \in \{1, 2, ..., 10\}$ on fairness and utility and set $\alpha = 7$ for the entire training as it balances the fairness-quality tradeoff.

4.2. Fair Latent Vector Generation

Now, we have attributed with a syntax-free, fair, and continuous low-dimensional space. Thus, generative frameworks $\mathcal{G}$, such as GANs and DMs can effectively focus on generating high-quality fair latent vectors without having to deal with high-dimensional feature types (such as categorical features in tabular data and pixels in images). In this work, we use various generative models based on GANs and DMs for latent space generation as mentioned in Section 2.3.

Let Z' be a generated latent space by any of the generative models (LSGAN, WGAN-GP, and DM) then $Z' \sim Z$ for a well-optimized generative model $\mathcal{G}$.

Theorem 1 (convergence guarantee). *Assume that (i) the data generation is Markov compatible with a pre-trained autoencoder, which is optimized for a combination of fairness loss and reconstruction loss, (ii) the neural networks involved in DGM have enough capacity, and (iii) the training of all the components of DGM is iterative until optimality, then for a well-optimized FLDGM, the generated fair latent distribution $p_{Z'}$ by the generator network* **G** *in $\mathcal{G}$ always converges to the ground-truth fair latent distribution p_Z* (proof in Appendix B).

Theorem 2 (fairness guarantee). *For a well-optimized generative model $\mathcal{G}$ in FLDGM, the generated fair latent vector Z' is $\mathcal{U}(S, Y)$ - fair, given the corresponding pre-trained autoencoder* (proof in Appendix B).

4.3. High Fidelity Reconstruction

In this stage, we pass the generated latent vectors Z' to the pre-trained decoder $\mathcal{D}$ to reconstruct the features from the latent vectors in a single pass. The reconstruction model parameterized by $\phi_\mathcal{D}$ can then be represented as:

$$\mathcal{D}_{\phi_\mathcal{D}}(X'|Z', S), X' \sim X \tag{8}$$

Summary. The whole process of fair latent deep generative modeling can be formally defined as:

$$\begin{aligned}\mathcal{D}_{\phi_\mathcal{D}}(X'|Z',S) &= \mathcal{D}_{\phi_\mathcal{D}}(X'|\mathcal{G}_\theta(\xi),Z)\\ &= \mathcal{D}_{\phi_\mathcal{D}}(\underbrace{X'|(\underbrace{\mathbf{G}_{\theta_\mathbf{G}}(\xi)|\underbrace{(\mathcal{E}_{\phi_\mathcal{E}}(Z|X,S)}_{Fair-abstract-compression}))}_{Fair-latent-vector-generation},S)}_{High-fidelity-reconstruction},\end{aligned} \tag{9}$$

where we denote $\mathcal{G}_\theta$ for any generative model (GAN and DM in our case) and $\mathbf{G}_{\theta_\mathbf{G}}$ is the corresponding generator network of $\mathcal{G}_\theta$. Note that, we use ϕ for denoting autoencoder parameters and θ for the generative modeling with an appropriate subscript.

Remark. Given corresponding pre-trained autoencoders, various datasets can be generated based on different fairness constraints and output tasks. This does not add any computational overhead to the generative modeling as fairness is enforced in a separate step.

Theorem 3 (prediction fairness (Xu et al., 2018) guarantee). *Any optimal downstream models* $\mathbf{M}$ *(without any explicit biases) trained on* $\mathbf{D}'$ *will have* $\mathcal{U}(S, Y)$ *- fair predictions on* $\mathbf{D}$ *given the corresponding pre-trained autoencoder and the well-optimized generator network* $\mathbf{G}$ (proof in Appendix B).

5. Experiments

5.1. Datasets

We performed experiments on two fairness benchmark datasets, Adult Income[3](table) and Color MNIST(Lee et al., 2021) (image), where the sensitive feature S is significantly correlated with the target label and thus the proper removal of S could be challenging.

Adult Income. The Adult dataset is a tabular data containing over 65,000 instances with 11 attributes, such as age, education, gender, and income, among others. We treat gender as the sensitive attribute (as there is a known bias between gender and income) and use income as the binary output label representing whether a person earns over $50K$ or not.

Color MNIST. The color MNIST is an image database containing handwritten digits and colors for the intrinsic and biased features. Following previous studies, the color MNIST used in our experimental analysis is based on (Lee et al., 2021). More details on the analysis of fairness measures of these datasets are given in Appendix C.

5.2. Evaluation metrics

We evaluate the quality and fairness of our proposed models using the following measures:

1. Data utility - We use precision, recall, and AUROC for evaluating data utility (Flach & Kull, 2015; Kynkäänniemi et al., 2019; Sajjadi et al., 2018). We train Random Forest (RF) classifier on synthetic data and test it on real data for downstream prediction and compare it to the state-of-the-art.

2. Sample-level metric - We perform sample-level metrics analysis proposed in (Alaa et al., 2022) to measure the fidelity, diversity, and generalization of synthetic data generated by our proposed models.

3. Synthetic data fairness - We use both FTU and DP (Section 2.1) for analyzing downstream fairness using a Random Forest classifier.

Furthermore, we perform explainability (Lundberg & Lee, 2017) and bias amplification (Wang et al., 2019) analysis to substantiate our study. We generated synthetic data using our generative models WGAN-GP, LSGAN, and DM on the datasets mentioned above and computed the metrics by taking an average of over 10 repetitive runs. We have the following variants: FLD-WGAN-GP, FLD-LSGAN, and FLD-DM each with FTU and DP for the fairness definitions in $\mathcal{U}(\mathcal{S}, \mathcal{Y})$. The neural network architectures and implementation details are given in Appendix D.

Competing Methods. The methods we benchmark against for Adult Income data are FairGAN and DECAF (as these models are designed for tabular data). Also, we compare the results of WGAN-GP without fairness to analyze the importance of FDGP. We follow the results from (van Breugel et al., 2021)[4] as it is difficult to reproduce the results of DECAF as studied in (Wang et al., 2022). For Color MNIST, we perform visualization analysis and explainability for verifying the utility, fairness, and quality of our proposed models.

6. Results

6.1. Data Utility and Fairness

6.1.1. Adult Income

We list the utility and fairness measures in Table 1. The precision of our proposed GAN and DM models is far better than FairGAN, GAN, and WGAN-GP, whereas we obtain almost the same score with DECAF. We improve all the state-of-the-art methods in terms of recall score. The AUROC score of our FLDGM models is better (around 10

[3] https://archive.ics.uci.edu/ml/datasets/adult

[4] The results are taken from the paper directly

percent improvement) compared to the corresponding fair generation methods. Note that DECAF-ND is simply a causal GAN without any fairness optimization. The FTU and DP of our proposed models are superior to the corresponding state-of-the-art fair generation methods.

6.1.2. Color MNIST

We perform visualization analysis on Color MNIST data. Following (Liu et al., 2022), we set color as the sensitive attribute and the generated images are de-correlated from color. Then, we control the color intensity to generate digits with a single color, meaning that the color is disentangled from the digits (this is done in the fair abstract compression stage). By changing the value of color intensity, the generated digits can be either blue, green, or red with approximately the same digit style (Figure 2). Note that, digit generation with fairness does not degrade the image quality, since the generative models could focus on image quality in the DGP.

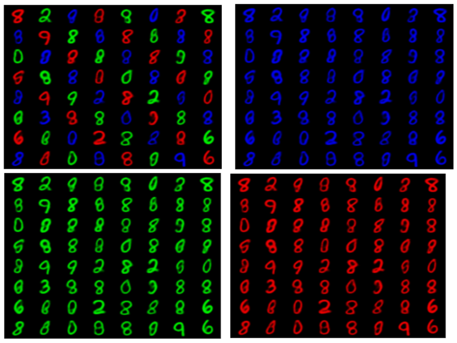

Figure 2. Generated digits with similar styles, color as a sensitive attribute.

To verify the quality and fairness of downstream tasks, we train a classifier on the generated digits and tested it on real data, that contains various color biases (by changing the Standard Deviation (SD)) (Table 2.) (details in Appendix C). We get an accuracy of 0.951 for color digit classification, whereas the accuracy of the classifier trained on real data is 0.931. This implies that the proposed FLDGM helped to improve the classification performance of digit prediction.

6.2. Sample-level Metric Analysis

Motivated by a recent study (Alaa et al., 2022) on evaluating the faithfulness of synthetic data, we perform sample-level metrics analysis on our proposed variants. This is to measure the quality of synthetic data generation in terms of fidelity, diversity, and authenticity. The results are given in Figure 3. Note that, our proposed models are highly authentic as per (Alaa et al., 2022), which shows the significance of our FLDGM.

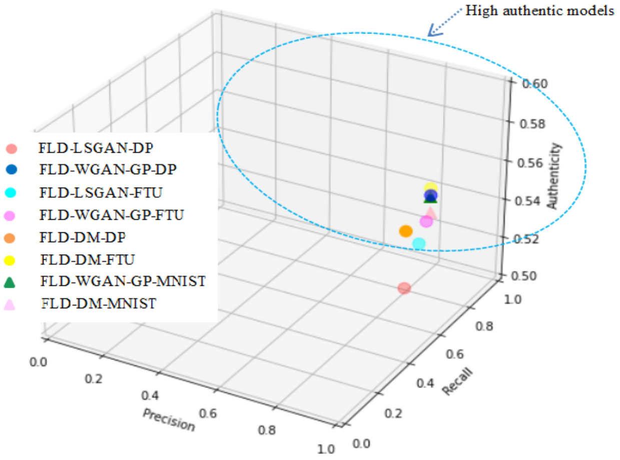

Figure 3. Sample level metrics analysis

6.3. Data Leakage Analysis

To further evaluate the data leakage, $\lambda_{\mathbf{D}}$ and model leakage, $\lambda_{\mathbf{M}}$ (Wang et al., 2019) of our proposed model, we train an attacker (which is a 'gender' classifier on Adult Income data) on the ground truth labels and model prediction by a Random Forest classifier. The $\lambda_{\mathbf{M}}$ trained on different data are given in Table 3. It shows that the leakage is controlled in data generation by all the proposed models as the bias amplification $\Delta(\lambda_{\mathbf{M}} - \lambda_{\mathbf{D}})$ is less than 0 for all the models with both FTU and DP as target fairness constraints.

6.4. Explainability Analysis

In order to analyze the difference in predictions of a Random Forest classifier trained on both real data and synthetic data, generated by our proposed model (we consider FLD-WGAN-GP-DP), we explain the predictions using Shapely Additive Explanations. For Adult income data, the contribution of the feature 'sex' is reduced, whereas the contribution of 'Relationship', 'Educational-Num', and 'Hours per week' (these are intrinsic features for income prediction) is slightly increased as given in Figure 4. For color MNIST, the most important areas of digits are concentrated on the shape in generated data (Figure 5 on right), whereas in real data, it is distributed over the entire area including the background (Figure 5 on the left).

7. Related Works

We focus on the related literature in terms of (i) non-parametric generative models and (ii) fair synthetic data generation. We refer to Appendix E for an overview of comparing various generative models with respect to our key areas of interest.

Method	**Data Quality**			**Fairness**	
	Precision (↑)	Recall (↑)	AUROC (↑)	FTU (↓)	DP (↓)
Real data	0.920 ± 0.006	0.936 ± 0.008	0.807 ± 0.004	0.116 ± 0.028	0.180 ± 0.010
GAN	0.607 ± 0.080	0.439 ± 0.037	0.567 ± 0.132	0.023 ± 0.010	0.089 ± 0.008
WGAN-GP	0.683 ± 0.015	0.914 ± 0.005	**0.798 ± 0.009**	0.120 ± 0.014	0.189 ± 0.024
FairGAN	0.681 ± 0.023	0.814 ± 0.079	0.766 ± 0.029	0.009 ± 0.002	0.097 ± 0.018
DECAF-ND	0.780 ± 0.023	0.920 ± 0.045	0.781 ± 0.007	0.152 ± 0.013	0.198 ± 0.013
DECAF-FTU	0.763 ± 0.033	0.925 ± 0.040	0.765 ± 0.010	0.004 ± 0.004	0.054 ± 0.005
DECAF-CF	0.743 ± 0.022	0.875 ± 0.038	0.769 ± 0.004	0.003 ± 0.006	0.039 ± 0.011
DECAF-DP	0.781 ± 0.018	0.881 ± 0.050	0.672 ± 0.014	0.001 ± 0.002	0.001 ± 0.001
FLD-LSGAN-FTU (ours)	0.762 ± 0.002	**0.998 ± 0.023**	0.762 ± 0.012	0.002 ± 0.001	**0.000 ± 0.001**
FL-LSGAN-DP (ours)	0.763 ± 0.001	0.941 ± 0.002	0.771 ± 0.010	**0.000 ± 0.001**	**0.000 ± 0.000**
FL-WGAN-GP-FTU (ours)	0.772 ± 0.034	0.918 ± 0.001	0.763 ± 0.023	0.001 ± 0.001	**0.000 ± 0.001**
FL-WGAN-GP-DP (ours)	0.782 ± 0.001	0.951 ± 0.001	0.762 ± 0.013	**0.000 ± 0.000**	**0.000 ± 0.000**
FL-DM-FTU (ours)	**0.791 ± 0.011**	0.912 ± 0.002	0.795 ± 0.001	**0.000 ± 0.000**	0.001 ± 0.000
FL-DM-DP (ours)	0.786 ± 0.002	0.905 ± 0.001	0.787 ± 0.011	**0.000 ± 0.001**	**0.000 ± 0.001**

Table 1. Data quality and fairness analysis of the proposed FLDGM with real data as a reference.

	SD=0.020	SD=0.025	SD=0.030	SD=0.035	SD=0.040	SD=0.045	SD=0.050
($\alpha = 0$)	.476±.005	.576±.001	.664±.007	0.720±.010	.785±.003	0.838±0.002	.931±.001
($\alpha = 0.5$)	.901±.001	0.927±.003	.950±.020	0.812±.002	.950±.001	.951±.001	951±.001

Table 2. Experiment on CMNIST data with FLDGM ($\alpha = 0.5$) and no fairness ($\alpha = 0$)

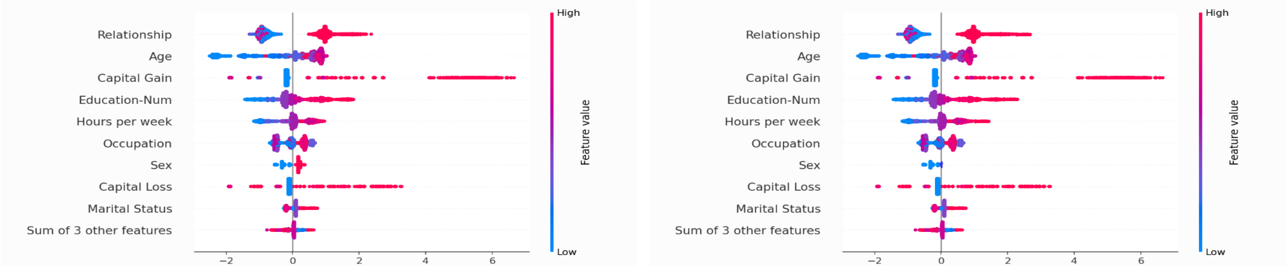

Figure 4. SHAP analysis on Adult Income

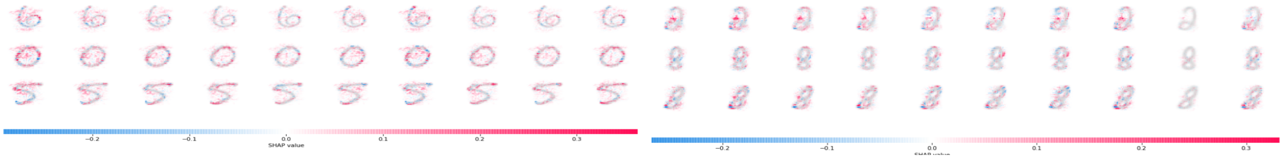

Figure 5. SHAP analysis on color MNIST

7.1. Non-parametric Generative Models

The state-of-the-art methods in synthetic data generation are either based on GANs (Gulrajani et al., 2017; Yoon et al., 2020; Xie et al., 2018) or Variational Auto Encoders (VAE) (Welling & Kingma, 2014). Recently, diffusion models have shown many improvements in high-quality synthetic data generation, particularly in images. The models above are well known for synthetic data generation, having trade-offs in various properties such as quality, diversity, etc, but unable to generate fair data (except (Xu et al., 2018) as discussed below).

Method	λ_M (↓)	Δ(↓)
Real data	0.642	0.022
WGAN-GP	0.712	0.092
FairGAN	0.583	-0.037
FLD-LSGAN-FTU	0.535	-0.085
FLD-LSGAN-DP	0.503	-0.117
FLD-WGAN-GP-FTU	0.505	-0.115
FLD-WGAN-GP-DP	0.502	-0.118
FLD-DM-FTU	**0.500**	**-0.120**
FLD-DM-DP	0.511	-0.109

Table 3. Bias amplification by models trained on different data. The reference dataset leakage is 0.620 with an approximate F1 score of 0.86.

7.2. Fair Synthetic Data Generation

The methods under this category (Xu et al., 2018; Choi et al., 2020) range from training generative models with a combination of generative loss and fairness loss, adversarial training, and post-processing schemes.

In FairGAN (Xu et al., 2018), an adversarial approach is proposed to predict the sensitive attributes from the generated data. This encourages the generator to produce data that is independent of the sensitive features. One main problem with this approach is that the adversary cannot be trained until convergence in every epoch which in turn degrades the performance. Also, this method is designed for binary-sensitive attributes. A fair data generation method by giving access to a small reference fair data is introduced in (Choi et al., 2020). The motivation of this work is not aligned with downstream fairness and explicit notions of fairness (van Breugel et al., 2021). A post-processing de-biasing method based on causal knowledge is proposed in (van Breugel et al., 2021), where the de-biasing has done at the inference time after the sequential generation of features by individual generators. This approach is designed for tabular data and is strictly based on the causal relationship between features. The computational complexity of this approach is very high, though the de-biasing is flexible to various fairness constraints at the target domain. Another approach based on VAE is proposed in (Louizos et al., 2016), where fairness is introduced by an additional regularization based on Maximum Mean Discrepancy (MMD) to get complete independence between data and sensitive attributes. This method imposes additional overhead for optimizing the MMD in the DGP.

8. Disadvantages and Societal Implications

Disadvantages. The fairness and quality of synthetic data generated by our proposed Fair Latent Deep Generative Models are limited by the performance of the Fair abstract compression stage. Thus the choice of auto-encoder architecture and corresponding fair compression should be designed in a way to balance the tradeoffs between quality and fairness. However, we have succeeded in reducing the computational overhead of syntax-specific generation (high-dimensional) and prevented quality loss when optimizing for fairness in the DGP, with very high flexibility in fine-tuning to various architectures and tasks, which is a great improvement in this context.

Societal Implications. Adversarial attacks on GANs can reveal training instances (Carlini et al., 2021; Tinsley et al., 2021), which is a hot topic of research. However, the extent to which it applies to diffusion models is under-explored. Moreover, generative models tend to exacerbate biases that are present in the training data (Gupta et al., 2021). In our proposed approach, the training instances are continuous fair latent vectors that do not directly reveal personal information in adversarial attacks (as it is encoded). Therefore, in an environment where privacy is of great concern, it is advisable to have an authentic human-in-the-loop who keeps the details of the autoencoder and shares other components for downstream applications, thereby having proper control over data privacy. For the second problem, de-biasing data happens in the fair compression stage, thereby the subsequent generative modeling could not access the bias information in the data, which greatly controls the bias amplification in downstream models.

9. Conclusion

We have proposed Fair Latent Deep Generative Models (FLDGM), a syntax-agnostic and model-agnostic generative framework that enables an efficient way to significantly improve both the fairness and quality of synthetic data generation using Diffusion models and Generative Adversarial Networks on image and tabular data. Based on our experimental analysis and evaluation, we did demonstrate favorable results in terms of quality, fidelity, diversity, authenticity, and fairness compared to state-of-the-art schemes across a wide range of proposed models in the absence of task-specific architectures.

Future Directions. One interesting future research direction could be to extend this framework for de-biasing hate speech detection and replace the biased contents in the tweets or speech with another, that could be generated by any underlying Natural Language Generation (NLG) methods. This area has not been explored but is very important as it has applications in text summarization, question generation, hate speech detection and removal, and text-to-image generation. Also, in contexts, where multi-modal data contains various biases, it could be interesting to first learn a common representation without biases and then build downstream models on top of it.

References

Alaa, A., Van Breugel, B., Saveliev, E. S., and van der Schaar, M. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International Conference on Machine Learning*, pp. 290–306. PMLR, 2022.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Azadi, S., Olsson, C., Darrell, T., Goodfellow, I., and Odena, A. Discriminator rejection sampling. *arXiv preprint arXiv:1810.06758*, 2018.

Barocas, S. and Selbst, A. D. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

Bell, C. Mutual information and maximal correlation as measures of dependence. *The Annals of Mathematical Statistics*, pp. 587–595, 1962.

Binns, R. Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*, pp. 149–159. PMLR, 2018.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Choi, K., Grover, A., Singh, T., Shu, R., and Ermon, S. Fair generative modeling via weak supervision. In *International Conference on Machine Learning*, pp. 1887–1898. PMLR, 2020.

Creager, E., Madras, D., Jacobsen, J.-H., Weis, M., Swersky, K., Pitassi, T., and Zemel, R. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*, pp. 1436–1445. PMLR, 2019.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268, 2015.

Feng, R., Yang, Y., Lyu, Y., Tan, C., Sun, Y., and Wang, C. Learning fair representations via an adversarial framework. *arXiv preprint arXiv:1904.13341*, 2019.

Flach, P. and Kull, M. Precision-recall-gain curves: Pr analysis done right. *Advances in neural information processing systems*, 28, 2015.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Grgic-Hlaca, N., Zafar, M. B., Gummadi, K. P., and Weller, A. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS symposium on machine learning and the law*, volume 1, pp. 2. Barcelona, Spain, 2016.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

Guo, D., Wang, C., Wang, B., and Zha, H. Learning fair representations via distance correlation minimization. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Gupta, A., Bhatt, D., and Pandey, A. Transitioning from real to synthetic data: Quantifying the bias in model. *arXiv preprint arXiv:2105.04144*, 2021.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., and Aila, T. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 32, 2019.

Lee, J., Kim, E., Lee, J., Lee, J., and Choo, J. Learning debiased representation via disentangled feature augmentation. *Advances in Neural Information Processing Systems*, 34:25123–25133, 2021.

Liu, J., Li, Z., Yao, Y., Xu, F., Ma, X., Xu, M., and Tong, H. Fair representation learning: An alternative to mutual information. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1088–1097, 2022.

Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. S. The variational fair autoencoder. In *ICLR*, 2016.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. Learning adversarially fair and transferable representations. In

International Conference on Machine Learning, pp. 3384–3393. PMLR, 2018.

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., and Paul Smolley, S. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2794–2802, 2017.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

Moyer, D., Gao, S., Brekelmans, R., Galstyan, A., and Ver Steeg, G. Invariant representations without adversarial training. *Advances in Neural Information Processing Systems*, 31, 2018.

Nam, J., Mo, S., Lee, J., and Shin, J. Breaking the spurious causality of conditional generation via fairness intervention with corrective sampling. *arXiv preprint arXiv:2212.02090*, 2022.

Nan, L. and Tao, D. Variational approach for privacy funnel optimization on continuous data. *Journal of Parallel and Distributed Computing*, 137:17–25, 2020.

Oneto, L., Donini, M., Pontil, M., and Maurer, A. Learning fair and transferable representations with theoretical guarantees. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 30–39. IEEE, 2020.

Rényi, A. On measures of dependence. *Acta mathematica hungarica*, 10(3-4):441–451, 1959.

Rodríguez-Gálvez, B., Thobaben, R., and Skoglund, M. A variational approach to privacy and fairness. In *2021 IEEE Information Theory Workshop (ITW)*, pp. 1–6. IEEE, 2021.

Sajjadi, M. S., Bachem, O., Lucic, M., Bousquet, O., and Gelly, S. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 2018.

Tinsley, P., Czajka, A., and Flynn, P. This face does not exist... but it might be yours! identity leakage in generative models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1320–1328, 2021.

van Breugel, B., Kyono, T., Berrevoets, J., and van der Schaar, M. Decaf: Generating fair synthetic data using causally-aware generative networks. *Advances in Neural Information Processing Systems*, 34:22221–22233, 2021.

Wang, S., Verhagen, P., Zhuge, J., and Shulev, V. Replication study of decaf: Generating fair synthetic data using causally-aware generative networks. In *ML Reproducibility Challenge 2021 (Fall Edition)*, 2022.

Wang, T., Zhao, J., Yatskar, M., Chang, K.-W., and Ordonez, V. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5310–5319, 2019.

Welling, M. and Kingma, D. P. Auto-encoding variational bayes. *ICLR*, 2014.

Xie, L., Lin, K., Wang, S., Wang, F., and Zhou, J. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*, 2018.

Xu, D., Yuan, S., Zhang, L., and Wu, X. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pp. 570–575. IEEE, 2018.

Xu, D., Yuan, S., Zhang, L., and Wu, X. Fairgan+: Achieving fair data generation and classification through generative adversarial nets. In *2019 IEEE International Conference on Big Data (Big Data)*, pp. 1401–1406. IEEE, 2019.

Yoon, J., Drumright, L. N., and Van Der Schaar, M. Anonymization through data synthesis using generative adversarial networks (ads-gan). *IEEE journal of biomedical and health informatics*, 24(8):2378–2388, 2020.

Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, pp. 962–970. PMLR, 2017.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *International conference on machine learning*, pp. 325–333. PMLR, 2013.

Zhang, L., Wu, Y., and Wu, X. A causal framework for discovering and removing direct and indirect discrimination. *arXiv preprint arXiv:1611.07509*, 2016.

A. Background

A.1. Distance correlation

State-of-the-art studies (Bell, 1962; Rényi, 1959) point that a dependence measure should satisfy seven properties such as symmetry, boundedness, monotonicity, etc. Both mutual information and distance correlation satisfy five properties out of seven mentioned in (Bell, 1962). One problem with mutual information-based dependence measure is that it should need an adversary to approximate the upper bound, which is unstable. Distance correlation accounts for this and can better balance the fairness-quality tradeoff. The complexity of computing distance correlation is dependent on the discrete values of S, latent space dimension, and batch size. In (Liu et al., 2022), the authors used matrix algebra and parallel computation to compute this fairness penalty.

A.2. Latent vector generation using WGAN-GP

In (Arjovsky et al., 2017), it has shown that divergences that GANs minimize are not continuous with the generator, leading to training instability. They introduced the concept of the Earth-Mover (also called Wasserstein-1) distance, which is the minimum cost of transporting mass in order to transform from one distribution to the other. They have also added a gradient penalty to the GAN training to avoid the vanishing gradient problem. Therefore the final objective function for the fair latent vector generation is:

$$L = \underbrace{\underset{\tilde{\boldsymbol{z}}\sim\mathbb{P}_g}{\mathbb{E}}[D(\tilde{\boldsymbol{z}})] - \underset{\boldsymbol{z}\sim\mathbb{P}_z}{\mathbb{E}}[D(\boldsymbol{z})]}_{\text{Original critic loss}} + \underbrace{\lambda \underset{\hat{\boldsymbol{z}}\sim\mathbb{P}_{\hat{\boldsymbol{z}}}}{\mathbb{E}}\left[(\|\nabla_{\hat{\boldsymbol{z}}} D(\hat{\boldsymbol{z}})\|_2 - 1)^2\right]}_{\text{gradient penalty}}, \tag{10}$$

The $P^{\tilde{z}}$ denotes sampling uniformly between fair latent distribution P_z and the generator distribution P_g. The penalty coefficient $\lambda = 10$ is set according to (Arjovsky et al., 2017).

B. Theorems and proofs

Lemma 1. Given an optimal discriminator D in GAN and a fixed diffusion process in DM, the global minimum of $\mathbf{G}$ is achieved if and only if $p_{\mathbf{G}} = p_{\mathbf{R}}$, where $p_{\mathbf{G}}, p_{\mathbf{R}}$ respectively denote the generated latent distribution and ground truth fair latent distribution.

Proof. For GAN, we refer to Theorem 1 from (Goodfellow et al., 2020). The GAN min-max optimization is performed by updating the generator G and the discriminator D with corresponding losses. Note that we have not altered the GAN training in FLDGM. So, all the theoretical results for GANs can be transferred to FLDGM. Therefore, the global minimum of G in GAN-based FLDGM can be achieved if and only if $p_G = p_{\mathbf{R}}$. We could not find any theoretical convergence results for DMs, but if exists, can be transferred to FLDGM as we have made zero changes to the DM training. Therefore the global minimum of a generator neural network $\mathbf{G}$ in FLDGM is achieved if and only if $p_{\mathbf{G}} = p_{\mathbf{R}}$. We denote $\mathbf{G}$, as a general representation of a generator network in FLDGM.

Theorem 1. *Assume that (i) the data generation is Markov compatible with a pre-trained autoencoder, which is optimized for a combination of fairness loss and reconstruction loss, (ii) the neural networks involved in DGM have enough capacity, and (iii) the training of all the components of DGM is iterative until optimality, then for a well-optimized FLDGM, the generated fair latent distribution $p_{Z'}$ by the generator network $\mathbf{G}$ in $\mathcal{G}$ always converges to the ground-truth fair latent distribution p_Z.*

Proof for Theorem 1. Given the adequate capacity of $\mathbf{G}$, by the convex training of DGM and the existence of global optimum as stated in Lemma 1, the Z' always converges to Z.

Theorem 2. *For a well-optimized generative model $\mathcal{G}$ in FLDGM, the generated fair latent vector Z' is $\mathcal{U}(S,Y)$ - fair, given the corresponding pre-trained autoencoder.*

Proof for Theorem 2. Let $\mathbf{G}$ be a generator neural network in $\mathcal{G}$. For a fixed $\mathbf{G}$, the $\mathcal{G}$ will converge to a true fair latent distribution Z and once optimized $\mathbf{G}$ can generate synthetic fair latent vectors Z' similar to Z (using Theorem 1). Thus the fairness, $\mathcal{U}(S,Y)$ contained in Z will be replicated in Z' for a well-optimized $\mathcal{G}$. Here, $\mathcal{U}(S,Y)$ is a definition of algorithmic fairness enforced in the fair abstract compression stage(autoencoder). Therefore the generated fair latent vector Z' is $\mathcal{U}(S,Y)$ *- fair.*

Theorem 3. *Any optimal downstream models $\mathbf{M}$ (without any explicit biases) trained on $\mathbf{D}'$ will have $\mathcal{U}(S,Y)$ - fair* *predictions on* $\mathbf{D}$ *given the corresponding pre-trained autoencoder and the generator network* $\mathbf{G}$.

Proof for Theorem 3. According to theorem 2, the generated fair latent vector Z' is $\mathcal{U}(S, Y)$ *- fair*. Given the autoencoder and $\mathbf{G}$, we have access to the decoder $\mathcal{D}$. The transition from fair latent Z' to fair data $\mathbf{D}'$ can be done in a single pass through $\mathcal{D}$ without retraining. Suppose, we have an optimal downstream predictor $\mathbf{M}$, which can be any universal function approximator (e.g., MLP), that is trained on a sufficiently large quantity of the synthetic data $\mathbf{D}'$ generated by $\mathbf{G}$, and passed through $\mathcal{D}$, then the optimal prediction is $\mathcal{U}(S, Y)$ *- fair* as we have not altered the training of decoder while reconstruction. Therefore, the definition of algorithmic fairness $\mathcal{U}(S, Y)$ will be satisfied by $\mathbf{D}'$.

Note that, we can achieve maximum fair performance when the model does not have any explicit biases. Therefore, for the sake of simplicity, we assume that the model does not add any biases during training.

C. Fairness analysis on Adult Income and CMIST data

C.1. Adult Income

For Adult income data, we used two fairness measures namely FTU and DP respectively for measuring direct and indirect discrimination. We selected 'gender' as a sensitive attribute as per studies (Feldman et al., 2015; Zhang et al., 2016) as there is a bias between 'gender' and 'income'. Around 68 percent of the gender population are men. One issue with this is that the model parameters may tend to be skewed toward the majority. For example, the trends will vary between the female and male populations. Trends mean the associations between attributes and the target variables. This may cause unfairness in the female population as the model maximizes accuracy across the entire population. In this case, we can define a privileged group as the male population and an unprivileged group as the female population.

Fairness Through Unawareness (FTU). It is used to analyze the direct influence of sensitive features, 'gender' in our case on income prediction. It can be calculated by the difference in predictions of a classifier for setting $'gender' = 1$ and $'gender' = 0$ (1 for male and 0 for female) such that the difference should be zero if the 'gender' has no direct influence on 'income'. We refer to Table 1 as the FTU for real data is 0.116 which is not fair. This means that the feature gender has a direct influence on income prediction. The metric FTU does account for direct discrimination as it only measures the direct influence.

Demographic Parity (DP). DP is based on the Predicted as Positive (PPP) rates. This means that DP measures the percentage of individuals who have either been correctly (TP) or incorrectly (FP) predicted as positive. In summary, this is the percentage of individuals who have benefited from the model prediction. Therefore, it measures the indirect influence of 'gender' on 'income' prediction. This is important in measuring all the features which contributed to the positive prediction, which means that if there is any other feature that has a strong correlation with 'gender', termed as a proxy attribute can also push some individuals to positive prediction. Therefore, it is a strong measure of indirect discrimination through proxy attributes. For a fair prediction, the DP should be close to zero, which means that both the unprivileged and privileged groups should have the same positive predictions. From Table 1, the DP of real data is 0.180 which is illegal.

C.2. CMIST data

The CMNIST data (Lee et al., 2021) is designed with seven standard deviation (SD) values (equally spaced between 0.02 and 0.05): the lower the value, the more difficult for the model to perform the task, since the model can fit the training set by recognizing colors instead of shapes. The classification performance has been improved in FLDGM which substantiates the importance of debiasing CMNIST data. An example of a prediction is given in Figure 6.

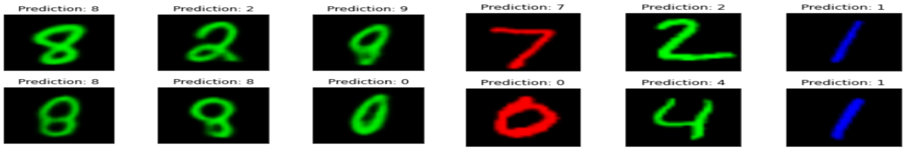

Figure 6. Downstream prediction of a classifier on color MNIST.

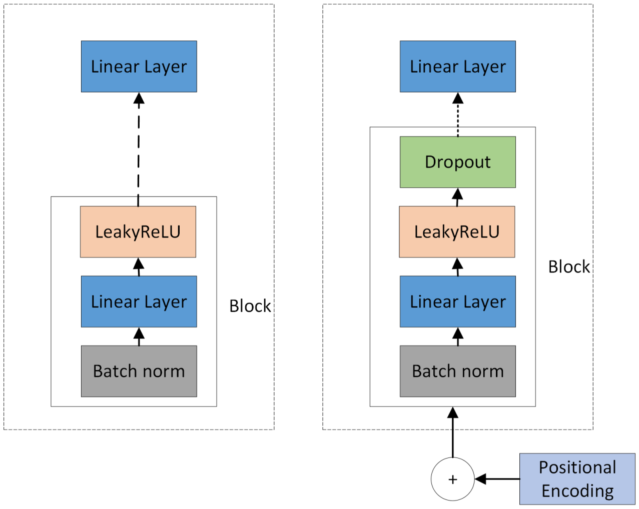

Figure 7. Architectures for GAN (left) and diffusion (right)

D. Architecture and implementation details

D.1. Generative model architecture

In our Fair Latent Deep Generative Models (FLDGM) framework, we used GAN-based and Diffusion-based architectures. For the GAN-based generative model, we used Least Square GAN (Mao et al., 2017) and Wasserstein GAN with gradient penalty (Gulrajani et al., 2017) to generate fair latent space. For the generator and discriminator of both GAN architectures, we used a series of blocks consisting of batch normalization followed by linear layer and leaky relu activation functions. In the last layer, we used a linear layer.

For the diffusion architecture, we used the gaussian diffusion model (Ho et al., 2020) which adds noise to the real data over some time steps T and makes the data a gaussian noise in the forward process, and then in the reverse process, we learn a neural network to approximate equation 5. For the neural network architecture, we used two blocks consisting of batch normalization, linear layer, relu activation functions, and lastly a dropout layer. Also, to get the data in any timesteps $t \in T$, we used positional encoding before passing the data to the blocks. Figure 7 shows both the architecture for GAN and the diffusion model.

D.2. Autoencoder architecture

The autoencoder architecture we used consists of linear layer-ReLU-linear layer for both the encoder and decoder for Adult income data. For CMNIST, we used 5 pairs of Convsd-ReLU with a flatten at the output for the encoder. For the decoder, we used 5 pairs of ConvTranspose2d-ReLU with a sigmoid at the output. We used a batch size of 2048 with a learning rate of 1e-3 and training epochs of 1000 for both data.

D.3. Implementation details and Hyperparameters

Architecture	batch size	optimizer	learning rate	epochs	timesteps (T)	number of critic
FLD-WGAN-GP	2048	Adam	2e-3	20000	×	5
FLD-LSGAN	2048	Adam	2e-3	20000	×	×
FLD-DM	2048	Adam	1e-4	5000	1000	×

Table 4. Hyperparameters for our models

In order to generate fair latent space, first we need the ground truth. For that, we train both the Adult Income dataset and MNIST dataset with the autoencoder(Liu et al., 2022). We followed the same hyperparameters used in the original

Architecture	Layer dimension
Generator	$64 \rightarrow 128$ $128 \rightarrow 256$ $256 \rightarrow 128$ $128 \rightarrow 64$ $64 \rightarrow 8$
Discriminator	$8 \rightarrow 128$ $128 \rightarrow 64$ $64 \rightarrow 32$ $32 \rightarrow 1$
Diffusion	$8 \rightarrow 256$ $256 \rightarrow 256$ $256 \rightarrow 8$

Table 5. Layer Dimension for the Generator, Discriminator

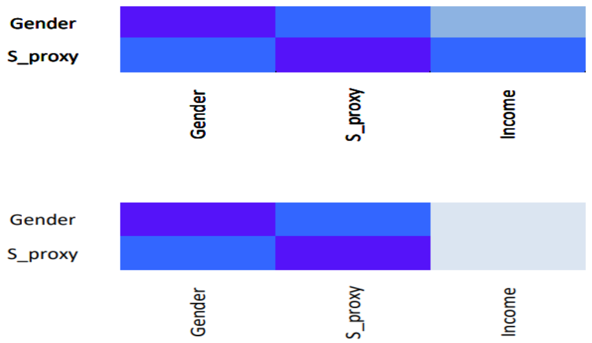

Figure 8. FLDGM accounts for proxy attributes (dark blue indicates high correlation)

experiment.

Once the ground truth fair latent space has been obtained, we trained GAN and diffusion models to generate a synthetic version of it. Hyperparameters for our models can be found in Table 4. In the FLD-WGAN-GP and FLD-LSGAN architectures, we used four blocks of network mentioned in Section D for the generator and three blocks for the discriminator. The layer dimension used in each block for FLD-WGAN-GP, FLD-LSGAN, and FLD-DM can be found in Table 5.

E. Related work summary

An overview of related works in terms of different key areas of interest is given in Table 5.

Model	Reference	(i)	(ii)	(iii)	(iv)	Goal
VFAE	(Louizos et al., 2016)	✓	×	×	↑	synthetic data
DM	(Ho et al., 2020)	×	×	✓	↑	synthetic data
FairGAN	(Xu et al., 2018)	✓	×	×	↑	fair synthetic data
DECAF	(van Breugel et al., 2021)	✓	×	✓	↑	fair synthetic data
Fair latent deep generative models	ours	✓	✓	✓	↓	fair synthetic data

Table 6. Overview of related works. The key areas of interest are (i) provide fairness, (ii) syntax-agnostic generation, (iii) fairness optimization is separated from DGP, and (iv) computational overhead (↑ - high, ↓ - low).

F. Comparison of conditional generation and FLDGM

We conducted an additional experiment to see the difference between the fairness objectives based on conditional generation and distance correlation minimization. Conditional generation is done by generating balanced samples conditioned on the attribute 'gender'. In order to assess the effect of a proxy attribute, we created an extra feature S_{proxy} that is strongly correlated with 'gender' in the Adult Income dataset. For the male sub-group, we set $S_{proxy} = 1$ for 95 percent of all cases, and for the remaining $S_{proxy} = 0$. For the female group, the above values are swapped. A correlation plot in Figure 8 shows that there is a strong correlation between S_{proxy} and income in the conditional generation. In FLDGM, this correlation has been minimized to a reasonable extent. Therefore the distance correlation-based fairness objective in FLDGM is superior in debiasing including proxy attributes, which proves the study in (Wang et al., 2019). Also, conditional generation is not advisable in situations where there are more sensitive or proxy attributes, and keeping the balance between all these attributes is not easy while generation. Another line of research has been done in (Nam et al., 2022), where the spurious correlation of proxy attributes is tackled by corrective sampling, which involves retraining the generative model followed by a discriminator rejection sampling (Azadi et al., 2018).