A SYNTHETIC DATA GENERATION

For both Computationally Unidentifiable (CU) and Computationally Identifiable (CI) data, we sampled from two-dimensional Gaussian distributions as:

$$\mathbf{x}|y=0 \sim \mathcal{N}\left(\begin{pmatrix}0\\-2\end{pmatrix}, \begin{pmatrix}10 & 1\\1 & 3\end{pmatrix}\right), \quad \mathbf{x}|y=1 \sim \mathcal{N}\left(\begin{pmatrix}0\\2\end{pmatrix}, \begin{pmatrix}5 & 1\\1 & 5\end{pmatrix}\right). \tag{7}$$

To have perfect balance, we sample 500 samples each for each label. For CU distribution, we assign the sensitive attribute to each sample as:

$$a \sim \text{Bern}(\frac{1}{2}). \tag{8}$$

While, the sensitive attribute of CI distribution is identifiable and formulated as:

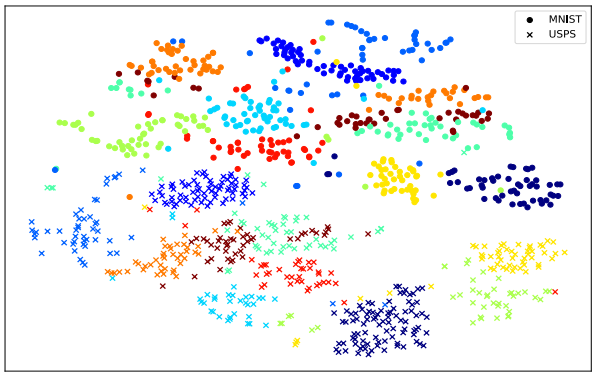
$$a = \begin{cases}1, & \text{for } x_1 \geq 0\\ 0, & \text{for } x_1 < 0,\end{cases} \tag{9}$$

where x_1 is the first entry of $\mathbf{x}$.

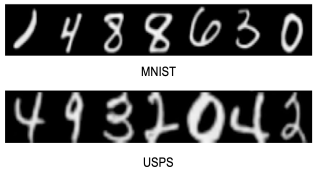
B MOTIVATING PROBLEM: GENERATIVE MODEL

Generative models usually suffer from mode collapse (Thanh-Tung & Tran, 2020) and posterior collapse (Lucas et al., 2019), which result in a lack of diversity of synthetic images. When it comes to fair generation, we want to generate samples to have similar distribution among sensitive groups. If the sensitive information is correlated to a certain attribute in the feature space, the model would be likely to generate images only from the related group. For example, when we want to generate blond hair person, it would mostly generate *female* (Hwang et al., 2020; Liu et al., 2021).

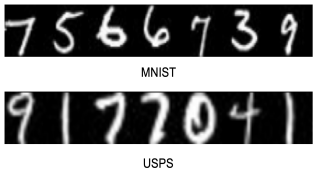
To validate this, we study simple CNN-based VAE with MNIST-USPS dataset following Li et al. (2020). Both MNIST and USPS have gray-scale handwritten digits as in Fig. 3c. We expect the latent representation to be clustered by intrinsic information (digit) regardless of the sensitive information (the source: MNIST or USPS). However, it is interesting to note that the representation is explicitly separated by sensitive information as in Fig. 3a. This indicates that when we perturb the sample in the latent space to obtain similar samples, we would only get samples from the same protected group. Generally, it is likely to be from a privileged (majority) group of the label. This would create a fatal problem in the fair generation that lacks the diversity of samples, which is the key to generating synthetic images.



(a) t-SNE visualization of the learned representation



(b) Reconstruction of MNIST-USPS



(c) Samples of MNIST-USPS

Figure 3: Qualitative analysis of the generative model (VAE) on MNIST-USPS dataset. Left figure illustrates t-SNE (Van der Maaten & Hinton, 2008) visualization of learned representation. Each color (*resp.* shape) indicates a different digit (*resp.* source: MNIST or USPS). We observe that the learned representation is clearly separated by the source (sensitive attribute). This computationally identifiable distribution can lead to an imbalance in image generation, *i.e.,* unfairness.

C PROOF OF THEOREM 4.4

Before we prove Theorem 4.4, we introduce Lemma C.1 as:

Lemma C.1. *Given two symmetric positive semi-definite matrices $A \in \mathbb{R}^{d\times d}$ and $B \in \mathbb{R}^{d\times d}$, the following inequality holds:*

$$0 \leq \mathrm{Tr}\left(A^{\frac{1}{2}}BA^{\frac{1}{2}}\right)^{\frac{1}{2}} \leq \sqrt{\mathrm{Tr}(A)\,\mathrm{Tr}(B)}.$$

Proof. Denote the eigen decomposition of matrix A and B as

$$A = U_A S_A U_A^\top, \quad B = U_B S_B U_B^\top,$$

where $S_A = Diag([\alpha_1, \alpha_2, \ldots, \alpha_d]) \in \mathbb{R}^{d\times d}$ and $S_B = Diag([\beta_1, \beta_2, \ldots, \beta_d]) \in \mathbb{R}^{d\times d}$ are diagonal matrices with the eigenvalues $\alpha_j|_{j=1}^d$ and $\beta_j|_{j=1}^d$ sorted in the descending order, *i.e.,* $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_d$, and $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_d$. Given that both A and B are positive semi-definite, we have $\alpha_j \geq 0, \beta_j \geq 0, j = 1, 2, \ldots, d$.

First, we prove the left-hand side of the inequality in Lemma 4.4 by showing that $A^{\frac{1}{2}}BA^{\frac{1}{2}}$ is positive semi-definite. For any vector $\mathbf{v} \in \mathbb{R}^d$, we have

$$\begin{aligned} & \mathbf{v}^\top A^{\frac{1}{2}}BA^{\frac{1}{2}}\mathbf{v} \\ =\ & (A^{\frac{1}{2}}\mathbf{v})^\top B(A^{\frac{1}{2}}\mathbf{v}) \\ =\ & (A^{\frac{1}{2}}\mathbf{v})^\top U_B S_B^{\frac{1}{2}} S_B^{\frac{1}{2}} U_B^\top (A^{\frac{1}{2}}\mathbf{v}) \\ =\ & \|(A^{\frac{1}{2}}\mathbf{v})^\top U_B S_B^{\frac{1}{2}}\|_2^2 \\ =\ & \geq 0, \end{aligned} \tag{10}$$

where the first equality is derived because matrix A is symmetric. Since $A^{\frac{1}{2}}BA^{\frac{1}{2}}$ is positive semi-definite, we have $\mathrm{Tr}\left(A^{\frac{1}{2}}BA^{\frac{1}{2}}\right)^{\frac{1}{2}} \geq 0$.

Second, we prove the right-hand side of the inequality of Lemma 4.4. We can rewrite $\mathrm{Tr}\left(A^{\frac{1}{2}}BA^{\frac{1}{2}}\right)^{\frac{1}{2}}$ as

$$\begin{aligned} & \mathrm{Tr}\left(A^{\frac{1}{2}}BA^{\frac{1}{2}}\right)^{\frac{1}{2}} \\ =\ & \mathrm{Tr}\left(A^{\frac{1}{2}}B^{\frac{1}{2}}\right) \\ \leq\ & \sum_{j=1}^{d}\sqrt{\alpha_j\beta_j}, \\ \leq\ & \sqrt{\sum_{j=1}^{d}\alpha_j}\sqrt{\sum_{j=1}^{d}\beta_j} \\ =\ & \sqrt{\mathrm{Tr}(A)\,\mathrm{Tr}(B)}, \end{aligned} \tag{11}$$

where the first inequality is obtained based on the Von Neumann trace inequality (Mirsky, 1975) and the second inequality is based on the Cauchy–Schwarz inequality. □

Now we are ready to prove Theorem 4.4.

Proof. Consider a cluster $X \in \mathbb{R}^{d\times n}$ with n samples. X is normalized to zero mean (*i.e.,* $X\mathbf{1} = \mathbf{0}$), where $\mathbf{1} \in \mathbb{R}^n$ is a vector with all elements being 1, and $\mathbf{0} \in \mathbb{R}^d$ is a vector with all elements being 0. In this cluster, suppose there are l_1 samples in the sensitive group $a = 1$, and l_0 samples in the sensitive group $a = 0$. We have $l_1 + l_0 = n$.

For a matrix X, denote $\mathbf{x}_i$ as the i-th column of X. Define a matrix $U \in \mathbb{R}^{d\times n}$ as the following:

$$\mathbf{u}_i = \begin{cases} \mathbf{x}_i, & \text{if } a_i = 1, \\ \mathbf{0}, & \text{else.} \end{cases} \tag{12}$$

where $a_i \in \{0, 1\}$ is the sensitive feature of the i-th sample.

Similarly we define a matrix $V \in \mathbb{R}^{d \times n}$ as the following:

$$\mathbf{v}_i = \begin{cases} \mathbf{x}_i, & \text{if } a_i = 0, \\ \mathbf{0}, & \text{else.} \end{cases} \tag{13}$$

Based on the definition in equation 12 and equation 13 we have $U + V = X$.

Assume that samples in $a = 1$ and $a = 0$ groups are drawn from multivariate Gaussian distributions, respectively. According to Definition 1 in the main paper, the Fréchet distance (FD) between U and V is defined as:

$$FD^2(U, V) = \|\mu_U - \mu_V\|_2^2 + \text{Tr}\left(\Sigma_U + \Sigma_V - 2(\Sigma_U^{\frac{1}{2}}\Sigma_V\Sigma_U^{\frac{1}{2}})^{\frac{1}{2}}\right), \tag{14}$$

where $\mu_U = \frac{1}{n}\sum_i \mathbf{u}_i$ and $\mu_V = \frac{1}{n}\sum_i \mathbf{v}_i$ are the means of U and V, respectively; Σ_U and Σ_V are the covariance matrices:

$$\Sigma_U = \tfrac{1}{n-1}\sum_i (\mathbf{u}_i - \mu_U)(\mathbf{u}_i - \mu_U)^\top, \tag{15}$$

$$\Sigma_V = \tfrac{1}{n-1}\sum_i (\mathbf{v}_i - \mu_V)(\mathbf{v}_i - \mu_V)^\top. \tag{16}$$

We can rewrite equation 15 as:

$$\begin{aligned} \Sigma_U &= \frac{1}{n-1}\sum_i (\mathbf{u}_i - \frac{1}{n}U\mathbf{1})(\mathbf{u}_i - \frac{1}{n}U\mathbf{1})^\top \\ &= \frac{1}{n-1}\left(\sum_i \mathbf{u}_i\mathbf{u}_i^\top - \frac{1}{n}U\mathbf{1}\mathbf{1}^\top U^\top - \frac{1}{n}U\mathbf{1}\mathbf{1}^\top U^\top + \frac{n}{n^2}U\mathbf{1}\mathbf{1}^\top U^\top\right) \\ &= \frac{1}{n-1}\left(UU^\top - \frac{1}{n}U\mathbf{1}\mathbf{1}^\top U^\top\right) \\ &= \frac{1}{n-1}UHU^\top, \end{aligned} \tag{17}$$

where the matrix H in equation 17 is defined as

$$H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top.$$

Following the same procedure we can rewrite equation 16 as:

$$\Sigma_V = \frac{1}{n-1}VHV^\top. \tag{18}$$

Based on equation 17 and equation 18, we can rewrite the FD definition in equation 14 as:

$$\begin{aligned} &FD^2(U, V) \\ =&\frac{1}{n^2}\|U\mathbf{1} - V\mathbf{1}\|_2^2 + \frac{1}{n-1}\text{Tr}\left(UHU^\top + VHV^\top - 2\big((UHU^\top)^{\frac{1}{2}}VHV^\top(UHU^\top)^{\frac{1}{2}}\big)^{\frac{1}{2}}\right). \end{aligned} \tag{19}$$

Next we will prove two properties of $UHU^\top$ and $VHV^\top$: 1) symmetry; 2) positive semi-definite.

1) For symmetry, it can be easily verified that both $UHU^\top$ and $VHV^\top$ are symmetric matrices since $H = H^\top$.

2) For the positive semi-definite property, for any $\mathbf{v} \in \mathbb{R}^d$, we have

$$\mathbf{v}^\top UHU^\top\mathbf{v} = ||HU^\top\mathbf{v}||_2^2 \geq 0, \tag{20}$$

where the equality is derived based on the property that $H^\top H = H$. We can derive a similar inequality for $VHV^\top$. Thus both $UHU^\top$ and $VHV^\top$ are positive semi-definite.

Then based on Lemma 4.4, we can derive that:

$$\operatorname{Tr}\left((UHU^\top)^{\frac{1}{2}}VHV^\top(UHU^\top)^{\frac{1}{2}}\right)^{\frac{1}{2}} \geq 0, \tag{21}$$

$$\operatorname{Tr}\left((UHU^\top)^{\frac{1}{2}}VHV^\top(UHU^\top)^{\frac{1}{2}}\right)^{\frac{1}{2}} \leq \sqrt{\operatorname{Tr}(UHU^\top)\operatorname{Tr}(VHV^\top)}. \tag{22}$$

Denote

$$\mathbf{s}_U = U\mathbf{1}, \quad \mathbf{s}_V = V\mathbf{1}, \quad \mathbf{s} = X\mathbf{1}.$$

We have $\mathbf{s}_U + \mathbf{s}_V = \mathbf{s} = \mathbf{0}$ since X is normalized to zero mean. From equation 19 and equation 21 we can derive that

$$\begin{aligned}
& FD^2(U,V) \\
\leq\ & \frac{1}{n^2}\|U\mathbf{1} - V\mathbf{1}\|_2^2 + \frac{1}{n-1}\operatorname{Tr}\left(UHU^\top + VHV^\top\right) \\
=\ & \frac{1}{n^2}\|U\mathbf{1} - V\mathbf{1}\|_2^2 + \frac{1}{n-1}\operatorname{Tr}\left(UHU^\top + VHV^\top\right) \\
& \quad - \frac{2}{n-1}\sqrt{\operatorname{Tr}(UHU^\top)\operatorname{Tr}(VHV^\top)} + \frac{2}{n-1}\sqrt{\operatorname{Tr}(UHU^\top)\operatorname{Tr}(VHV^\top)} \\
=\ & \|\frac{\mathbf{s}_U}{n} - \frac{\mathbf{s}_V}{n}\|_2^2 + \frac{1}{n-1}\left(\|UH\|_F - \|VH\|_F\right)^2 + \frac{2}{n-1}\sqrt{\operatorname{Tr}(UHU^\top)\operatorname{Tr}(VHV^\top)},
\end{aligned} \tag{23}$$

where the last row of equation 23 is derived based on the property that $HH^\top = H$.

Further, we can derive that

$$\begin{aligned}
& \sqrt{\operatorname{Tr}(UHU^\top)\operatorname{Tr}(VHV^\top)} \\
=\ & \sqrt{\operatorname{Tr}(UU^\top - \frac{1}{n}\mathbf{s}_U\mathbf{s}_U^\top)\operatorname{Tr}(VV^\top - \frac{1}{n}\mathbf{s}_V\mathbf{s}_V^\top)} \\
=\ & \sqrt{\operatorname{Tr}(UU^\top - \frac{1}{n}\mathbf{s}_U\mathbf{s}_U^\top)\operatorname{Tr}(XX^\top - UU^\top - \frac{1}{n}\mathbf{s}_U\mathbf{s}_U^\top)} \\
\leq\ & \sqrt{\operatorname{Tr}(UU^\top + \frac{1}{n}\mathbf{s}_U\mathbf{s}_U^\top)\operatorname{Tr}(XX^\top - UU^\top - \frac{1}{n}\mathbf{s}_U\mathbf{s}_U^\top)} \\
\leq\ & \frac{1}{2}\operatorname{Tr}(XX^\top).
\end{aligned} \tag{24}$$

In the third row of equation 24, the fact that $\operatorname{Tr}\left(UU^\top + VV^\top\right) = \operatorname{Tr}(XX^\top)$ follows from the definition of U and V in equation 12 and equation 13, and the fact that $\operatorname{Tr}(\mathbf{s}_U\mathbf{s}_U^\top) = \operatorname{Tr}(\mathbf{s}_V\mathbf{s}_V^\top)$ is due to the property that $\mathbf{s}_U + \mathbf{s}_V = \mathbf{0}$. The inequality in the fourth row of equation 24 is derived because $VHV^\top$ is positive semidefinite and $\operatorname{Tr}(\mathbf{s}_U\mathbf{s}_U^\top) = \|\mathbf{s}_U\|_2^2 \geq 0$. The inequality in the last row of equation 24 is derived based on the Cauchy–Schwarz inequality.

From equation 23 and equation 24 we can derive the following for FD(U,V):

$$\begin{aligned}
FD^2(U,V) \leq\ & \|\frac{\mathbf{s}_U}{n} - \frac{\mathbf{s}_V}{n}\|_2^2 + \frac{1}{n-1}\left(\|UH\|_F - \|VH\|_F\right)^2 + \frac{1}{n-1}\operatorname{Tr}(XX^\top) \\
=\ & FFDC^2(U,V),
\end{aligned} \tag{25}$$

where $FFDC^2(U,V)$ is defined in Definition 2 of the main paper.

Based on equation 19, equation 22 and equation 25 we further have:

$$FFDC^2(U,V) - \frac{1}{n-1}\operatorname{Tr}(XX^\top) \quad \leq FD^2(U,V) \leq FFDC^2(U,V). \tag{26}$$

Given the multiple clusters $U_k|_{k=1}^c$ and $V_k|_{k=1}^c$ defined after Definition 2 in the main paper , we substitute U_k and V_k to equation 26 and get the following:

$$FFDC^2(U_k,V_k) - \frac{1}{n_k-1}\operatorname{Tr}(X_kX_k^\top) \quad \leq FD^2(U_k,V_k) \leq FFDC^2(U_k,V_k), k = 1,2,\ldots,c. \tag{27}$$

Taking the max operation on equation 27 we get

$$\begin{aligned}
&\max_k \left(FFDC^2(U_k, V_k) - \frac{1}{n_k - 1} \operatorname{Tr}(X_k X_k^\top) \right) \\
&\leq FFD^2(\{X_1, X_2, ..., X_k\}) - \max_k \frac{1}{n_k - 1} \operatorname{Tr}(X_k X_k^\top) \\
&\leq \max_k FD^2(U_k, V_k) \\
&\leq \max_k FFDC^2(U_k, V_k) \\
&= FFD^2(\{X_1, X_2, ..., X_k\}),
\end{aligned}$$

where $FFD^2(\{X_1, X_2, ..., X_k\})$ is defined in Definition 3 of the main paper. □

D Additional Experiment: Adult Dataset

We conduct an experiment on the famous fairness tabular benchmark: Adult (Kohavi, 1996). **Adult** data from the UCI repository (Kohavi, 1996) is a tabular dataset that contains 48,842 instances with the features such as workclass, education, and sex. The goal is to predict whether income exceeds 50K USD per year. The feature sex is used as the sensitive feature and $c = 2$ clustering problem.

Table 4 summarizes the experimental results. Similar to the results in the main paper, we could achieve significantly lower FFD while achieving similar or better accuracy and balance comparing with other methods.

	Adult			
	Acc (Diff)	NMI	Balance	FFD
Perfect	1.0 (0.0)	1.0	0.481	2.01
k-means++	0.714 (0.249)	0.170	0.111	1.95
ScFC (Backurs et al., 2019)	0.690 (0.042)	0.105	0.350	1.98
ALG (Bera et al., 2019)	0.684 (0.054)	0.095	0.356	1.98
VFC (Ziko et al., 2019)	0.690 (0.042)	0.105	0.478	2.00
DFC (Li et al., 2020)	0.686 (0.172)	0.152	0.182	12.6
Ours (only $\mathcal{L}_{cls}$)	0.686 (0.180)	0.152	0.171	1.37
Ours	0.697 (0.178)	0.157	0.186	0.46

Table 4: Evaluation of clustering methods on Adult dataset. For accuracy and NMI, it is higher the better. Balance is better if it is closer to perfect clustering *i.e.,* original data statistic. For accuracy difference, FFD, and the lower bound of FD, the lower, the better.

E Additional Experiment: Fair Downstream Task

To validate the claim that representation with lower FFD helps to improve fairness in downstream tasks, we evaluate with learned representation from DFC (Li et al., 2020) and ours, respectively. We train logistic regression (LR) for classification and k-means++ for clustering. In Table 5, we observe that representation with lower FFD consistently achieves lower fairness violations for both clustering and classification without any sacrifice of utility. This shows how distributional independence of the sensitive attribute is critical for fair downstream tasks.

		Representation	Clustering (k-means++)		Classifiaction (LR)	
		FFD	Acc (diff)	Balance	Acc (diff)	EOD
MNIST-USPS	Ours	1.82	0.831 (0.016)	0.090	0.881 (0.079)	0.069
	DFC	14.13	0.824 (0.160)	0.044	0.882 (0.091)	0.078
MTFL	Ours	48.64	0.726 (0.038)	0.105	0.802 (0.003)	0.129
	DFC	67.84	0.733 (0.158)	0.138	0.768 (0.030)	0.375
Adult	Ours	0.68	0.688 (0.188)	0.172	0.813 (0.138)	0.322
	DFC	7.15	0.685 (0.172)	0.181	0.813 (0.150)	0.306

Table 5: Evaluation of downstream tasks with representation from different deep fair methods on various datasets.