Breaking the Spurious Causality of Conditional Generation via Fairness Intervention with Corrective Sampling

Anonymous authors
Paper under double-blind review

Abstract

Trying to capture the sample-label relationship, conditional generative models often end up inheriting the spurious correlation in the training dataset, giving label-conditional distributions that are severely imbalanced in another latent attribute. To mitigate such undesirable correlations engraved into generative models, which we call spurious causality, we propose a general two-step strategy. (a) Fairness Intervention (FI): Emphasize the minority samples that are hard to be generated due to the spurious correlation in the training dataset. (b) Corrective Sampling (CS): Filter the generated samples explicitly to follow the desired label-conditional latent attribute distribution. We design the fairness intervention for various degrees of supervision on the spurious attribute, including unsupervised, weakly-supervised, and semi-supervised scenarios. Our experimental results show that the proposed FICS can successfully resolve the spurious correlation in generated samples on various datasets.

1 Introduction

The visual contents synthesized by deep generative models closely resemble the training data, sometimes in an undesirable way. Generative models, like classifiers, suffer from the *dataset bias* (Torralba & Efros, 2011; Buolamwini & Gebru, 2018) inherited from the training set, generating significantly fewer minority group samples (i.e., having latent attributes that are rarely combined with the label given) than the majority group samples (Zhao et al., 2018; Salminen et al., 2020; Jain et al., 2021). Such group imbalance is typically more amplified in the generated dataset than in the original training dataset; Tan et al. (2020) observe that the degree of imbalance continuously grows as the training proceeds, suggesting that the biased generation itself strengthens the bias further. This exacerbated imbalance in generated samples gives rise to various fairness issues in downstream applications, such as the underrepresentation of minority groups in content creation (Choi et al., 2020; Tan et al., 2020; Jalal et al., 2021) or harms the downstream classifier trained with generated samples (Xu et al., 2018; Sattigeri et al., 2018; Xu et al., 2019).

The imbalance in data generation introduces a slightly different (and potentially more problematic) type of fairness concerns for *conditional generative models* which synthesize samples conditioned on some given input, e.g., generating images corresponding to the given class label. Namely, each input-conditional distribution of the generated data may have an imbalanced distribution in terms of some latent attribute of the data, replicating the *spurious correlation* present in the training dataset. For instance, a BigGAN (Brock et al., 2019) trained on the CelebA (Liu et al., 2015) dataset with the “blondness” attribute as a class label may generate severely imbalanced samples in terms of “gender,” e.g., the vast majority of samples generated under the blond class condition are female (see Figure 1). It has also been recently reported that popular text-to-image generation models indeed contain such gender and cultural biases (Barr, 2022). When prompted with specific occupations like “childcare worker,” generated images are highly gender/race-imbalanced.

Such reproduction of spurious correlations by conditional generative models poses a deeper fairness concern than the imbalances in unconditional generative models. The act of conditioning on the input (i.e., giving the class label “blond”) can be viewed as an *intervention* from the perspective of counterfactual causality (Pearl, 2010). Once a spurious correlation is engraved into conditional generative models, changing the input (intervention) leads to a change in the distribution of spuriously correlated latent attributes of the conditionally

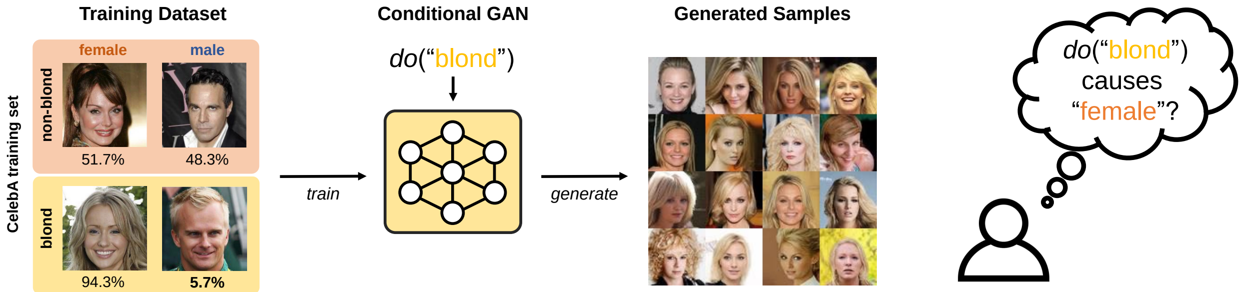


Figure 1: When trained on a dataset with spuriously correlated attributes (e.g., CelebA), conditional generative models learn to generate severely imbalanced samples when conditioned on the correlated attribute. For example, "blond"-conditioned cGAN produces only 3.88% "male" images. Such behavior is especially problematic, as once the cGAN is trained, there is a *causal relation* between the act of conditioning (intervention) and the distribution of generated samples (event), which amplifies the biased belief of users.

generated samples, forming a cause-effect relationship. This false sense of causation in conditional generation, which we call *spurious causality*,[1] is a big ethical threat as such misconception is likely to propagate to many downstream applications using generative models (Mariani et al., 2018).

However, up to our knowledge, no previous work aims to resolve the spurious causality of conditional generation. Prior results on fair generative models focus on the majority bias and address the issue by either (a) reducing the discrepancy from data distribution (Grover et al., 2019; Mo et al., 2019b; Lee et al., 2021; Yu et al., 2020; Humayun et al., 2022) or (b) following the desired fair distribution, e.g., specified by explicit labels of sensitive attributes (Tan et al., 2020) or a reference set of desiderata (Choi et al., 2020). Recall that (b) requires some additional supervision, while (a) can be done in an unsupervised manner. A naïve solution for the spurious causality is extending the prior work on majority bias to conditional generation, i.e., applying them for each conditional distribution. However, this naïve approach is suboptimal as it ignores the input condition. Instead, we aim to leverage the conditional structure to tackle the problem.

Contribution. We propose a general framework, coined *fairness intervention with corrective sampling* (FICS), to mitigate the severe yet underexplored problem of spurious causality in conditional generative models. FICS consists of two components. (1) Fairness intervention (FI): emphasize samples suffering from the spurious correlation in the dataset at the training phase. (2) Corrective sampling (CS): explicitly filter samples after the training phase so that the trained generator follows the desired distribution.

FICS provides a unified approach to resolve spurious causality for a wide range of scenarios, with various degrees of supervision available on the latent attributes and various target ratio of latent attribute distributions. To validate this point, we focus on the task of class-conditional image generation using conditional generative adversarial networks (Mirza & Osindero, 2014, cGANs) and apply FICS on unsupervised, weakly supervised, and semi-supervised setups. For the scenarios, the aim is to generate the conditional distributions whose latent attribute distribution is either identical to that of the training dataset or some desired distribution.

Our experiments show that FICS is an effective remedy for spurious causality. In particular, we train BigGAN (Brock et al., 2019) on CelebA (Liu et al., 2015) and Waterbirds (Sagawa et al., 2020), and ResNetGAN (Gulrajani et al., 2017) for Colored MNIST (Arjovsky et al., 2019). We observe that FICS consistently improves upon the conditional extensions of previous fair generation methods in both unsupervised and supervised scenarios, in terms of the sample quality and the minority attribute occurence ratio.

2 Related Work

Generative models. Generative models have synthesized high-fidelity samples in various domains, e.g., image (Karras et al., 2021), video (Yu et al., 2022), and language (Brown et al., 2020). Numerous methods

[1]We coined "spurious causality," as an analogy to the term spurious correlation used in classification contexts.

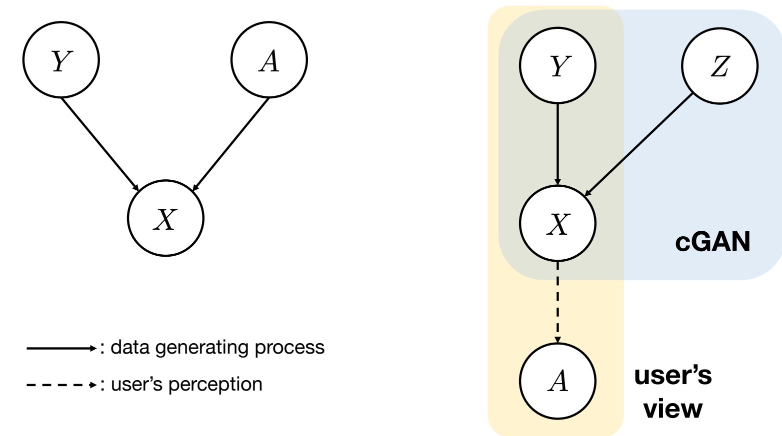
have been proposed, including generative adversarial networks (Goodfellow et al., 2014, GANs), variational autoencoders (Vahdat & Kautz, 2020, VAEs), autoregressive models (Esser et al., 2021), and diffusion models (Dhariwal & Nichol, 2021). While each method has its pros and cons, GANs are known to merit both sample quality and speed (Xiao et al., 2022). In addition, conditional GANs (cGANs) utilize additional supervision such as class labels (Mirza & Osindero, 2014), text descriptions (Ramesh et al., 2022; Ding et al., 2022), or reference images (Zhu et al., 2017; Mo et al., 2019a) as input conditions to extend the applicability further. Specifically, cGANs control the synthesized outputs and improve sample quality by restricting the output space for each condition. Despite their utility, we find that cGANs often suffer from a severe yet underexplored fairness issue. We investigate this issue, coined spurious causality, and propose a remedy for it.

Fairness in generative models. In accordance with its practical impact, the fairness (or bias) issue of the generative models has received considerable attention. In particular, most prior work on fair generation consider the majority bias, i.e., the tendency of generative models to learn the imbalance inherited from the training dataset, which is even amplified during training (Tan et al., 2020). A line of work considers an unsupervised setup, where we do not know the group identities of samples. Typical approaches in this direction handle the bias by using a discrepancy metric corrected with density ratio as the optimization objective (Grover et al., 2019; Mo et al., 2019b; Lee et al., 2021) or imposing regularizations to enforce a uniform learning of the data manifold (Yu et al., 2020; Humayun et al., 2022). Another line of work learns the fair distribution in a supervised manner, specified by sensitive labels (Xu et al., 2018; Sattigeri et al., 2018; Xu et al., 2019; Tan et al., 2020) or a reference set (Choi et al., 2020). Concretely, the former balance the occurrence of sensitive attributes by conditioning them explicitly, and the latter reweight the training samples using the reference set. However, the previous works only focus on the majority bias, and the spurious correlation over multiple attributes is not investigated.

Casuality in generative models. To achieve fairness in the generated data, there exists a line of work utilizing the underlying causal structure of the data, particularly in the tabular domain. Unlike the vision domain, where one may not directly observe all attributes, in the tabular domain, it is common for each data point to have all attribute labels, including the sensitive attribute of interest (Xu et al., 2018). Furthermore, there exists a line of research that assumes prior knowledge of the causal structure among the attributes, leveraging the parent attribute to generate its child attribute (Xu et al., 2019; van Breugel et al., 2021). Unlike the tabular domain, in the vision domain, it is not typical for each data point to have all attribute labels, due to the high cost associated with manually annotating each attribute. Our study differs from the aforementioned works, as we address more flexible scenarios with varying levels of supervision on the sensitive attribute and do not leverage underlying causal structure of the data to achieve fairness.

Relation to counterfactual fairness. Our framework shares some similarities with counterfactual fairness in terms of its capability to conduct counterfactual analysis on specific attributes. However, our framework performs do-operation on the class attribute and focuses on the distribution of the sensitive attribute in the generated images. On the other hand, counterfactual fairness aims to achieve fair prediction by performing do-operation on the sensitive attribute. As such, the methods used to achieve counterfactual fairness are not directly applicable to resolving the problem of interest in our framework.

Spurious correlation of classifiers. Spurious correlations in the training dataset lead classifiers to have poor generalization performance on out-of-distribution test sets where the spurious correlations do not hold (Geirhos et al., 2020). For the scenario where the learner has access to explicit labels on the spuriously correlated attribute, several works have been proposed to mitigate spurious correlation, including invariant learning (Arjovsky et al., 2019) and distributionally robust optimization (Sagawa et al., 2020). However, obtaining such explicit supervision on spurious attributes is expensive. Moreover, characterizing such spurious correlations in datasets requires laborious effort, as can be seen in the ablation studies on ImageNet-trained CNN having texture bias (Geirhos et al., 2018) or background bias (Xiao et al., 2020). Acknowledging the difficulty in acquiring explicit labels for spurious attributes, recent works explore weaker forms of supervision such as fair validation sets (Bahng et al., 2020; Nam et al., 2020; Creager et al., 2021; Liu et al., 2021) or a small set of attribute annotated samples (Nam et al., 2022; Sohoni et al., 2021). More specifically, this line of work (Nam et al., 2020; Creager et al., 2021; Liu et al., 2021; Nam et al., 2022; Sohoni et al., 2021) focus on how to identify poorly performing samples and reweight/resample them. Our method shares same philosophy to overcome spurious causality in conditional generation.



(a) data generating process (b) spurious causality

Figure 2: Although there is no causal relationship between the class attribute Y and the sensitive attribute A, conditioning the class to generate data can impact the sensitive attribute of the generated data. Consequently, cGAN users who observe the generated data and its sensitive attribute may falsely perceive that manipulating the class attribute alters the sensitive attribute of the generated data, as long as a spurious correlation exists between the class attribute and the sensitive attribute.

3 Spurious Causality of cGAN

Conditional generative models aim to approximate the ground-truth distribution $p_{\mathsf{data}}(x|y)$ of a data $x \in \mathcal{X}$ conditioned on a label $y \in \mathcal{Y}$, using a model distribution $p_{\mathsf{gen}}(x|y)$. For brevity, we assume that the models are conditional generative adversarial networks (Goodfellow et al., 2014; Mirza & Osindero, 2014, cGAN), although our framework can be applied to any generative models. cGAN trains a generator $G : (z, y) \mapsto x$ which generates the data x conditioned on a latent variable $z \sim p(z)$, drawn from a fixed prior distribution $p(z)$, and a label $y \sim p_{\mathsf{data}}(y)$. This procedure defines the generator distribution $p_{\mathsf{gen}}(x|y)$. Jointly, cGAN trains a discriminator $D : (x, y) \mapsto r \in (0, 1)$ which predicts whether the data-label pair (x, y) comes from the data ($r = 1$) or generator ($r = 0$) distributions. Formally, the training objective of cGAN is given by:

$$\mathcal{L}_{\mathsf{cGAN}}(G, D) := \mathbb{E}_{y\sim p_{\mathsf{data}}(y)}[\mathbb{E}_{x\sim p_{\mathsf{data}}(x|y)} \log D(x, y) + \mathbb{E}_{z\sim p(z)} \log(1 - D(G(z, y), y))]. \tag{1}$$

We consider training cGANs on the training dataset with *spurious correlations*. In other words, we assume that there is a sensitive attribute $a(x) \in \mathcal{A}$ corresponding to the data x, that is correlated but not causally related to the label y. In particular, we observe that the model distribution $p_{\mathsf{gen}}(x|y)$, when naïvely trained, learns to reproduce–often more severely–the spurious correlation in the training dataset. More concretely, define the *label-conditional probability of sensitive attributes* of the model distribution $p_{\mathsf{gen}}(x|y)$ as

$$p_{\mathsf{gen}}(a|y) = \int_{\mathcal{X}} \mathbf{1}[a = a(x)]\, p_{\mathsf{gen}}(x|y)\, \mathrm{d}x \tag{2}$$

and define $p_{\mathsf{data}}(a|y)$ likewise. We find that the model distribution $p_{\mathsf{gen}}(x|y)$ is trained in a way that $p_{\mathsf{gen}}(a|y)$ is more severely imbalanced than $p_{\mathsf{data}}(a|y)$.

Such exaggerated reproduction of spurious correlation by conditional generative models is a big problem, as the model itself now provides a causal link between the act of "conditioning on the label" and the "sensitive attribute of the generated sample." Indeed, the model can be viewed as a black-box on which a counterfactual experiment can be performed by intervening on the label with everything else fixed (e.g., latent variable z). Such analysis will lead to a conclusion that there exists a causal structure (Pearl, 2010) between the label and the spurious attribute of the generated samples.

This problem, which we call *spurious causality*, is potentially more problematic than the spurious correlation learned by classifiers. Namely, conditional generative models can propagate harmful stereotypes, such as the belief that the attribute "female" is caused by being "blond", to users who directly observe the effect of providing conditions to the generated samples.

We aim to mitigate this spurious causality of conditional generative models, by learning a balanced conditional distribution $p_{\mathsf{gen}}(x|y)$ which gives a desired sensitive attribute distribution $p_{\mathsf{gen}}(a|y)$. Here, the desired distribution may depend on the learners' intention:

(a) *Approximating the data distribution* $p_{\mathsf{data}}(a|y)$. As briefly discussed, standard GAN training tends give a more severely imbalanced model distribution $p_{\mathsf{gen}}(a|y)$ than the data distribution $p_{\mathsf{data}}(a|y)$. We aim to learn the model distribution that closely approximate the data distribution, i.e., $p_{\mathsf{gen}}(a|y) \approx p_{\mathsf{data}}(a|y)$.

(b) *Approximating a fair distribution* $p_{\mathsf{fair}}(a|y)$. We aim to approximate the sensitive attribute distribution specified according to some fairness criteria. For instance, one could enforce the attribute to be distributed following some label-independent population distribution, i.e., $p_{\mathsf{fair}}(a|y) = p_{\mathsf{fair}}(a)$, or even a uniform distribution, i.e., $p_{\mathsf{fair}}(a|y) = 1/|\mathcal{A}|$.

In addition to achieving desired $p_{\mathsf{gen}}(a|y)$, we also consider achieving a *fair* generative quality for each label as an auxiliary goal. For this purpose, we also report the (fair) Intra-FID for the compared methods.

4 Breaking the Spurious Causality

4.1 Overall framework

We describe the proposed FICS (Figure 3), a general strategy for mitigating the spurious causality of conditional generation. In a nutshell, FICS utilizes two mechanisms to conditionally synthesize samples with desired latent attribute distributions, where each component can be designed for the level of supervision available on the latent attributes.

- *Fairness Intervention* (FI): Conditionally generated samples that belong to the minority group are emphasized during the training, to enhance both generation quality and frequency of minority group samples. A key design challenge in FI is an identification of minority group identities of generated samples, using various degrees of supervision on latent attributes.
- *Corrective Sampling* (CS): This step performs explicit rejection sampling on the synthesized samples to achieve the desired conditional latent attribute distribution. While FI helps achieving this goal, FI alone is often not enough as it also aims to optimize the generation quality of minority samples, instead of focusing solely on the latent attribute frequencies. A key design challenge here is to devise various rejection mechanisms based on various levels of latent attribute supervisions.

We consider two different scenarios and give concrete versions of FICS for each case. First, an unsupervised setup aims to recover the data distribution by fixing the bias of the current model. Second, a supervised setup aims to follow the desired fair distribution described by additional supervision, such as a reference set of desired distribution or labels of sensitive attributes. The following subsections illustrate how to design the fairness intervention and corrective sampling for unsupervised and supervised scenarios.

4.2 Unsupervised: Learning the data distribution

In the unsupervised scenario, we encourage the generator to generate more samples that is likely to be mispredicted by a biased classifier. This approach is motivated by the observation that there is a strong correlation between the misprediction of the biased classifier and the generation frequency of the biased cGAN. We first introduce our observation, then illustrate how to fix the training and inference of cGAN.

4.2.1 Observation

Our observation (Figure 4) suggests the usefulness of having a separately trained classifier for mitigating the spurious correlation of conditional generative models. In particular, we train a classifier and a conditional generative model on the CelebA (Liu et al., 2015) dataset. We focus on the relationship between the "blondness" attribute and "gender", "wearing eyeglasses", and "smiling" attributes. When conditioned on blond, the

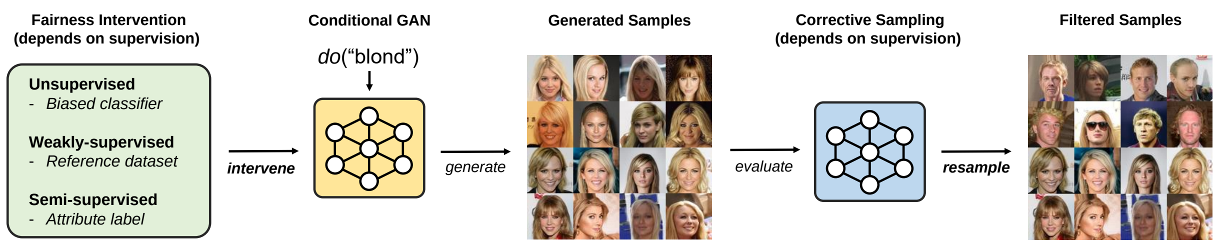


Figure 3: A visual illustration of FICS, a general two-stage framework to mitigate the spurious causality of conditional generation. Intervening the training phase to promote the generator to synthesize minority samples (FI), followed by resampling in the generation phase to further encourage the desired generator distribution (CS). Each step can be specified according to the degree of supervision provided.

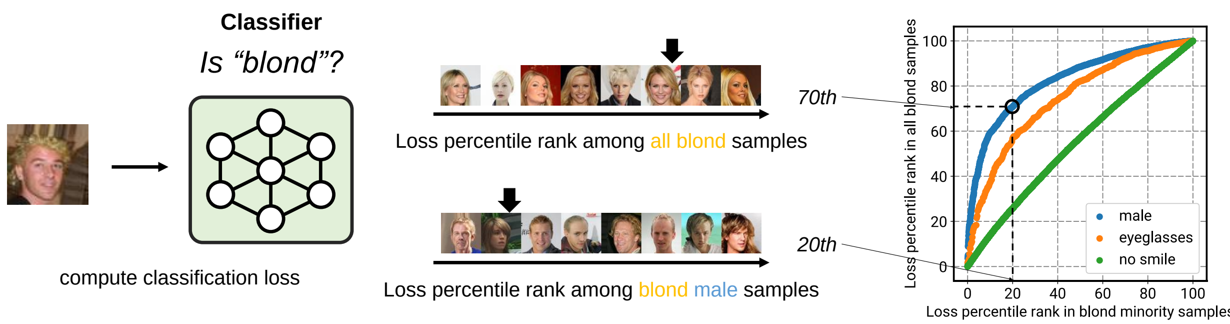


Figure 4: The relationship between the biased classifier and the generator represented through classification loss percentile ranks of generated samples. Given a point, its x, y coordinate represents the percentile rank of its classification loss among blond minority samples and all blond samples respectively. For instance, a sample having loss with 20th percentile among blond male samples has loss with around 70th percentile among blond samples. That is, blond male samples typically have higher classification loss among blond samples. "Gender" and "wearing eyeglasses" attributes are spuriously correlated with "blondness", while "smiling" attribute is not spuriously correlated with "blondness."

dataset is known to be severely imbalanced with respect to "gender" and "wearing eyeglasses" attributes. Unlike two attributes, there is no clear correlation between the "blondness" and "smiling" attributes.

From the trained model, we observe that the generated samples that are classified as the minority group are more likely to have higher classification loss. For instance, given a blond-conditionally generated sample, a sample with a higher classification loss is more likely to be male. In detail, we observe a percentile rank of the given sample among all blond samples and specific blond male samples. For example, a sample having the 20th percentile among blond male samples has a loss with around the 70th percentile among all blond samples. We observed a similar pattern for "wearing eyeglasses" attribute, which is spuriously correlated with "blondness" attribute. In contrast, "smile" attribute, which has no clear correlation with "blondness" attribute, the loss distribution of all blond samples and blond and no smile samples are not distinct. This observation suggests that the classification loss of generated samples may contain useful information to identify minority group samples, even in the case where the generator and classifier is trained separately and receive no supervision about the correlated attributes (e.g., "gender," "wearing eyeglasses").

4.2.2 Training and sampling

Fairness intervention. Motivated by the observation, we intervene the generator G to encourage synthesizing samples that are hard to be classified by the biased classifier $f_b : x \mapsto y$, which are likely to have minor attributes, i.e., low $p(a|y)$. To that end, we apply a regularizer that promotes the generated samples to maximize the classification loss (cross-entropy) of f_b. However, we found that maximizing cross-entropy is unstable since it loses convexity. Instead, we minimize the (sum of) cross-entropy over the wrong targets $y \neq y'$ to ensure convexity. Formally, our objective is:

$$\min_G \max_D \ \mathcal{L}_{\mathsf{cGAN}} + \lambda \cdot \sum_{y' \neq y} \mathsf{CE}(f_b(G(z,y)), y'), \tag{3}$$

where λ is a weight for the regularizer and CE denotes the cross-entropy loss. We simply set $\lambda = 0.01$ since it worked well for all our experiments. We pretrain the classifier f_b and fix it for the training of cGAN.

Corrective sampling. In the unsupervised scenario, the regularizer used in the first step distorts the training distribution. Thus, we correct the distribution of the generated samples utilizing the discriminator rejection sampling. Since the proposed regularizer in Eq. (3) upweights the minority samples, it loses the guarantee of generative models following the data distribution. Thus, we correct the sampling of the generated samples to recover the original data distribution. Rejection sampling technique is commonly employed to sample from a probability distribution p when the ratio p/q is known and sampling from q is feasible. Specifically, this involves sampling x from q and accepting the samples with a probability proportional to p/q. This procedure is equivalent to sampling from p. In our scenario, sampling x from p_{gen} and accepting the samples with probability proportional to $\frac{p_{\mathsf{data}}}{p_{\mathsf{gen}}}$ is equivalent to sampling from p_{data}. Here, we estimate the density ratio with an optimal discriminator $D^* = \frac{p_{\mathsf{data}}}{p_{\mathsf{data}}+p_{\mathsf{gen}}}$ for a given generator G following the discriminator rejection sampling Azadi et al. (2018). In practice, we use the discriminator D of cGAN as a proxy for D^*.

4.3 Supervised: Learning the fair distribution

In the supervised scenario, we intervene with the generator to upsample the minor samples described by a fair reference set or labels of sensitive attributes. We illustrate how to fix the cGAN given each supervision.

4.3.1 Weakly-supervised

Fairness intervention. Assume we have a reference set $\mathcal{S}_{\mathsf{ref}}$ that represents the desired fair distribution, i.e., $p_{\mathsf{fair}}(x|y)$ is the empirical distribution of $\mathcal{S}_{\mathsf{ref}}$. Here, we intervene with the generator to follow the reference set based on the density ratio between reference and data distributions. Specifically, we train a binary classifier $f_{\mathsf{ref}} : x \mapsto r \in (0,1)$ that distinguishes whether a sample belongs to the reference ($r = 1$) or data ($r = 0$) distributions in a similar manner to the density ratio trick of GANs. Formally, the (unnormalized) resampling probability is given by:

$$w(x) = \frac{p(f_{\mathsf{ref}}(x) = 1)}{M p(f_{\mathsf{ref}}(x) = 0)} \tag{4}$$

where $M > 0$ is some large constant to ensure $w(x) \leq 1$. During training, We draw samples from $p'_{\mathsf{data}}(x) := \frac{w(x)}{W} p_{\mathsf{data}}(x)$, where W is a normalizing factor that makes $p'_{\mathsf{data}}(x)$ a probability density function.

Corrective sampling. After the training, we train an additional discriminator D' which distinguishes the reference and generator distributions. Then, we apply rejection sampling (similar to the unsupervised case) to correct the samples to follow the desired fair distribution $p_{\mathsf{fair}}(x|y)$.

4.3.2 Semi-supervised

Fairness intervention. Assume we have some samples with sensitive attribute labels $\mathcal{L} := \{(x,y,a)\}$ in addition to the unlabeled (in the sense of sensitive attribute) samples $\mathcal{U} := \{(x,y)\}$. Here, we first train an attribute classifier $f_a : x \mapsto a$ using the labeled data $\mathcal{L}$, where one can also apply semi-supervised learning techniques leveraging the unlabeled data $\mathcal{U}$. Using this classifier, we estimate the population of each group $g = (y,a)$ over all samples, constructing the pseudo-labeled dataset $\mathcal{D} := \mathcal{L} \cup \hat{\mathcal{U}}$ where $\hat{\mathcal{U}} := \{(x,y,\hat{a})\}$ for $(x,y) \in \mathcal{U}$ and $\hat{a} := f_a(x)$. Here, we intervene with the generator to balance the occurrence of each group by resampling the training data with a probability inversely proportional to the group population. Formally, the (unnormalized) resampling probability is given by:

$$w(x) = \frac{1}{|\{(x', y', a') \in \mathcal{D} : (y', a') = (y, f_a(x))\}|}. \tag{5}$$

As the weakly-supervised scenario, We draw samples from $p'_{\mathsf{data}}(x) = \frac{w(x)}{W} p_{\mathsf{data}}(x)$ for training.

Corrective sampling. Similar to the unsupervised case, we correct the generator distribution to match the desired fair distribution $p_{\mathsf{fair}}(a|y)$ via rejection sampling. However, one cannot simply reuse the discriminator D from cGAN since D is trained with biased data distribution, resampled by Eq. (5). Thus, we train an additional discriminator D' to estimate the density ratio between (unbiased) data and generator distributions. Using D', one can recover the original data distribution $p_{\mathsf{data}}(x|y)$. Here, we apply one more rejection sampling to control the desired group ratio $p_{\mathsf{fair}}(a|y)$ using the predicted attribute $f_a(x)$. This two-stage rejection sampling: recover $p_{\mathsf{data}}(x|y)$ then control $p_{\mathsf{fair}}(a|y)$, corrects the samples to follow the desired fair distribution.

5 Experiments

5.1 Experimental setup

Datasets and models. We evaluate our framework on the CelebA (Liu et al., 2015) dataset and the Waterbirds (Sagawa et al., 2020) dataset. The CelebA dataset consists of 162,770 face images of celebrities and has 40 attribute annotations. We choose two attributes that are spuriously correlated; one is used as an input condition, and the other is used to evaluate the minority occurrence ratio. The Colored MNIST dataset (Arjovsky et al., 2019) consists of 40,000 digit images, having binary labels (0 for digits 0-4, 1 for digits 5-9). Each digit is colored by either red or green to impose a spurious correlation between binary labels and colors. Unlike the original construction, we do not flip the final label (no label noise). The Waterbirds dataset consists of 4,795 bird images with two types of birds (waterbirds and landbirds) on two types of backgrounds (water background and land background); the bird type is used as an input condition, and the background is used to evaluate minority occurrence ratio. We use BigGAN (Brock et al., 2019) for the conditional generative models and ResNet-50 (He et al., 2016) for the classifiers in all our experiments. We use ResNetGAN (Gulrajani et al., 2017) for the conditional generative models and 4-layer convolutional network for the classifiers in our Colored MNIST experiments. See Appendix A for additional details.

Evaluation metrics. We evaluate the conditional generative models in two aspects: (a) sample quality, i.e., $p(x|y)$, and (b) attribute balancing, i.e., $p(a|y)$. For (a) on unsupervised scenarios, we use the intra-class Fréchet Inception distance (Intra-FID) (Heusel et al., 2017), which measures both sample quality and diversity of conditional distribution. Specifically, Intra-FID measures the distance between the feature distributions of data and generated samples. For supervised scenarios, we use a fair version of Intra-FID (Fair Intra-FID) that measures the distance between the fair and generated samples. For (b), we report the occurrence ratio of minority (or sensitive) attributes, which should follow the ratio of data and fair distributions for unsupervised and supervised scenarios, respectively. Here, we also report the estimated occurrence ratio for the (training/reference) dataset of the classifier as the prediction can be under- (or over-) estimated.

Baselines. We extend the prior (unconditional) fair generation methods as the baselines for the fair conditional generation, i.e., modify their formula from $p(x)$ to $p(x, y)$. We consider three representative methods: for unsupervised (Lee et al., 2021), weakly-supervised (Choi et al., 2020) scenarios, and supervised (Tan et al., 2020), respectively. For the unsupervised method, we consider Dia-GAN (Lee et al., 2021), which resamples training data based on an estimated log-density ratio to emphasize minority samples. For the weakly supervised method, we consider Choi et al. (2020), giving different weights for each sample based on the density ratio classifier trained to distinguish the training dataset and the reference dataset. For the supervised method, we consider FairGen (Tan et al., 2020), a latent variable modification method based on the attribute classifier trained on the latent variable space. See Appendix B for a detailed explanation.

Computation time. Training of the original BigGAN on CelebA for 200k iteration takes 30 hours on a single TITAN Xp GPU and 40 CPU cores (Intel Xeon CPU E5-2630 v4). FICS takes a longer training time due to an additional classifier; takes ×1.6 times for our training setup.

Table 1: Comparison of unsupervised fair conditional generation methods under CelebA, using (a) "blondness" or (b) "wearing lipstick" attributes as a class condition. We report the Intra-FID (↓) and minority occurrence ratio (%) of generated samples. We report the minority occurrence ratio of the training dataset and the estimated values by attribute classifier. Our method achieves the best of both worlds: almost matching the minority occurrence ratio of the original dataset (estimated by the classifier) while producing high-quality samples. The lowest intra-FID and the closest minority occurrence ratio to the estimated are marked in bold.

(a) Blondness

	Intra-FID (↓)		Minor in Blond (%)		
	Blond	Non-blond	Male	Eyeglasses	No lipstick
Dataset			5.72	1.66	18.78
Estimated			5.85	1.85	17.61
Unsupervised					
cGAN (Mirza & Osindero, 2014)	1.86	1.64	3.97	0.88	11.04
Dia-GAN (Lee et al., 2021)	1.88	1.84	5.06	1.49	14.10
FICS (ours)	**1.78**	**1.38**	**6.13**	**1.90**	**16.34**

(b) Wearing lipstick

	Intra-FID (↓)		Minor in Lipstick (%)		
	Lipstick	No lipstick	Male	Eyeglasses	Hat
Dataset			0.58	1.05	1.25
Estimated			0.80	1.21	1.87
Unsupervised					
cGAN (Mirza & Osindero, 2014)	1.36	1.67	0.28	0.76	1.04
Dia-GAN (Lee et al., 2021)	1.36	1.66	0.30	0.86	1.09
FICS (ours)	**1.24**	**1.41**	**0.33**	**1.02**	**1.53**

Table 2: Comparison of unsupervised fair conditional generation methods under (a) Colored MNIST using "digit" as a class condition and (b) Waterbirds using "bird type" as a class condition. We report the average value over classes for Colored MNIST and the values of each class for Waterbirds. The lowest intra-FID and the closest minority occurrence ratio to the estimated are marked in bold.

	Colored MNIST		Waterbirds		Landbirds	
	Intra-FID (↓)	Minor (%)	Intra-FID (↓)	Minor (%)	Intra-FID (↓)	Minor (%)
Dataset		20.00		5.03		5.00
Estimated		20.00		8.72		8.96
Unsupervised						
cGAN (Mirza & Osindero, 2014)	10.31	0.16	44.68	5.52	24.73	4.69
Dia-GAN (Lee et al., 2021)	10.20	0.05	44.32	5.54	24.80	4.66
FICS (ours)	**2.78**	**23.50**	**41.80**	**6.75**	**23.46**	**4.89**

5.2 Unsupervised experiments

We first demonstrate the results in an unsupervised scenario, where the goal is to reduce the discrepancy between the generator and data distributions. For CelebA, we use "blondness" and "wearing lipstick" as input conditions. For each condition, we choose {male, eyeglasses, no lipstick} and {male, eyeglasses, hat} as corresponding minority attributes. We report Intra-FID and minority occurrence ratio (compared to the data distribution) of the baselines and FICS in Table 1 and Table 2.

In the CelebA dataset, the vanilla cGAN suffers from the spurious causality issue, significantly amplifying the imbalance of minority attributes, e.g., only creates 3.97% of blond males while the original dataset contains

Table 3: Comparison of supervised fair conditional generation methods where "blondness" attribute is used as a class condition. We report the Fair Intra-FID (↓) and minority occurrence ratio (%) of generated samples. The lowest fair intra-FID and the closest minority occurrence ratio to the estimated are marked in bold.

	Target Male ratio: 30%		Target Male ratio: 50%	
	Fair Intra-FID (↓) Blond	Occurrence of Male in Blond (%)	Fair Intra-FID (↓) Blond	Occurrence of Male in Blond (%)
Reference		30.00		50.00
Estimated		28.60		47.15
Weakly-supervised				
Choi et al. (2020)	10.71	7.60	10.70	18.34
FICS (ours)	**10.36**	**15.65**	**10.24**	**27.04**
Semi-supervised				
FairGen (Tan et al., 2020)	12.71	**29.54**	11.98	49.76
FICS (ours)	**10.56**	25.72	**11.68**	**47.11**

(a) Vanilla BigGAN (Brock et al., 2019)

(b) FICS (ours)

Figure 5: Blond-conditionally generated samples from the original BigGAN and our method, under the semi-supervised scenario with a target male ratio of 50%. Our method creates diverse blond male images, unlike the original BigGAN is biased toward female images.

5.85%. Dia-GAN relieves this amplification and increases the minority occurrence; however, it often decreases the Intra-FID, i.e., losing the diversity as a cost of the boosted minority. In contrast, our proposed method, FICS, achieves the best of both worlds: it improves the Intra-FID of all considered cases and recovers the minority occurrence of the data distribution, e.g., producing 6.13% of blond males, almost matching the 5.85% of the training dataset.

We observe a similar tendency in the Colored MNIST and Waterbirds datasets. In the Colored MNIST dataset, both vanilla cGAN and Dia-GAN fails to generate sufficient number of minority group samples. In contrast, FICS achieves a minority occurrence ratio close to the training dataset. In the Waterbirds dataset, the vanilla cGAN suffers from the spurious causality issue, providing only 5.52% and 4.69% of minority samples of waterbirds and landbirds, respectively. Dia-GAN shows similar results to the vanilla cGAN, resulting from the almost identical sample weights. FICS shows improved Intra-FID and the closest minority occurrence to that of the training set for both classes.

5.3 Supervised experiments

We demonstrate the results for the supervised scenario, where the goal is to follow fair distribution. We conduct experiments on the CelebA dataset, using "blondness" as an input condition and "gender" as a sensitive attribute. Recall that the original data distribution has only a few blond males; we aim to learn a fair distribution that produces {30%, 50%} of male images conditioned on blond. We report the Fair Intra-FID and minority occurrence ratio (should be similar to the desired fair distribution) of the baselines

Table 4: Comparison of (a) unsupervised and (b) semi-supervised fair conditional generation methods with or without each component of FICS. For (b), we use "blondness" attribute is used as a class condition and 50% male in blond as the desired distribution. The lowest fair intra-FID and the closest minority occurrence ratio to the estimated are marked in bold. Both fairness intervention (FI) and corrective sampling (CS) benefits.

(a) Unsupervised

	Intra-FID (↓)	Male in Blond (%)
cGAN	1.86	3.97
FI only	1.80	6.02
CS only	1.85	5.02
FICS (ours)	**1.78**	**6.13**

(b) Semi-supervised

	Fair Intra-FID (↓)	Male in Blond (%)
cGAN	13.16	3.97
FI only	12.30	18.23
CS only	12.91	38.49
FICS (ours)	**11.68**	**47.11**

and our proposed framework. Table 3 shows the results on weakly- and semi-supervised scenarios: we will discuss the details and observations in the remaining section.

Weakly-supervised. We use the same 4,000 samples (but without sensitive attribute annotations) as the reference set for the weakly-supervised setup. We compare our method with Choi et al. (2020), and use the same binary classifier that distinguishes the reference and data distributions for both methods; recall that we resample training data with the classifier while Choi et al. reweights the loss. The results show that our method outperforms Choi et al. for all considered cases, both Fair Intra-FID (i.e., generation quality) and minority occurrence ratio (i.e., attribute balancing). Here, both methods still have a gap with the semi-supervised methods since they rely on the weaker form of supervision. Still, we remark that our method significantly reduces the gap between semi-supervised and weakly-supervised scenarios, particularly in terms of the minority occurrence ratio.

Semi-supervised. We use 4,000 labeled (sensitive attribute) samples with the same attribute occurrence ratio of the desired fair distribution for the semi-supervised setup. We compare our method with FairGen, and use the same attribute classifier for both methods, which is trained by a semi-supervised learning technique called FixMatch (Sohn et al., 2020). The results show that our method shows better Fair Intra-FID than FairGen, while both methods show comparable minority occurrence ratio. Recall that FairGen modifies the prior distribution to generate minority samples; it distorts the prior to the low-density area of the original Gaussian distribution. Thus, FairGen often generates low-fidelity samples (though satisfying the desired attribute), leading to the lower FID.

Visualization. Figure 5 visualizes the generated samples conditioned on blond from vanilla BigGAN and debaised one by FICS. We use the semi-supervised version with a target male ratio of 50% for our method. Unlike the original BigGAN mostly creates female images, our method creates diverse blond male images.

5.4 Ablation study

We conduct an ablation study to see the contribution of each component in our semi-supervised method. The result is shown in Table 4. With fairness intervention during the training (based on pseudo-labeling and group balancing), we could observe gain in both fair intra-FID and minority occurence ratio, still there exists large gap between the occurrence ratio achieved and the desired occurrence ratio. With corrective sampling after the training, we could observe reasonable occurrence ratio for male, still there is room to be improved. With full components of our semi-supervised method, we can see further improvement in both minority occurrence ratio and fair intra-FID.

6 Conclusion

We raise the issue of spurious causality, which occurs from conditional generative models trained on the dataset with spurious correlation. To alleviate this, we propose the FICS framework that intervene with the training and correct the sampling of cGAN. Our experiments show that the FICS framework outperforms the conditional extension of previous fair generation methods, in both scenarios of recovering data distribution and following fair distribution.

Broader Impact Statement

Although we alleviate some spurious causality of conditional generative models, the model can still retain unrecognized biases. Since the algorithmic debiasing cannot remove all possible biases from the model, the users should carefully check the model on their usage.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Samaneh Azadi, Catherine Olsson, Trevor Darrell, Ian Goodfellow, and Augustus Odena. Discriminator rejection sampling. In *International Conference on Learning Representations*, 2018.

Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *International Conference on Machine Learning*, 2020.

Kyle Barr. AI image generators routinely display gender and cultural bias. *Gizmodo*, November 2022. URL https://gizmodo.com/ai-dall-e-stability-ai-stable-diffusion-1849728302.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 2018.

Kristy Choi, Aditya Grover, Trisha Singh, Rui Shu, and Stefano Ermon. Fair generative modeling via weak supervision. In *International Conference on Machine Learning*, 2020.

Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning*, 2021.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, 2021.

Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217*, 2022.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2018.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673, 2020.

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

Aditya Grover, Jiaming Song, Ashish Kapoor, Kenneth Tran, Alekh Agarwal, Eric J Horvitz, and Stefano Ermon. Bias correction of learned generative models using likelihood-free importance weighting. In *Advances in Neural Information Processing Systems*, 2019.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in Neural Information Processing Systems*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

Ahmed Imtiaz Humayun, Randall Balestriero, and Richard Baraniuk. Magnet: Uniform sampling from deep generative network manifolds without retraining. In *International Conference on Learning Representations*, 2022.

Niharika Jain, Alberto Olmo, Sailik Sengupta, Lydia Manikonda, and Subbarao Kambhampati. Imperfect imaGANation: Implications of GANs exacerbating biases on facial data augmentation and snapchat selfie lenses. In *ICLR Workshop on Synthetic Data Generation – Quality, Privacy, Bias*, 2021.

Ajil Jalal, Sushrut Karmalkar, Jessica Hoffmann, Alex Dimakis, and Eric Price. Fairness for image generation with uncertain sensitive attributes. In *International Conference on Machine Learning*, 2021.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Advances in Neural Information Processing Systems*, 2021.

Jinhee Lee, Haeri Kim, Youngkyu Hong, and Hye Won Chung. Self-diagnosing GAN: Diagnosing underrepresented samples in generative adversarial networks. In *Advances in Neural Information Processing Systems*, 2021.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, 2021.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *IEEE/CVF International Conference on Computer Vision*, 2015.

Giovanni Mariani, Florian Scheidegger, Roxana Istrate, Costas Bekas, and Cristiano Malossi. Bagan: Data augmentation with balancing gan. *arXiv preprint arXiv:1803.09655*, 2018.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Instagan: Instance-aware image-to-image translation. In *International Conference on Learning Representations*, 2019a.

Sangwoo Mo, Chiheon Kim, Sungwoong Kim, Minsu Cho, and Jinwoo Shin. Mining gold samples for conditional gans. In *Advances in Neural Information Processing Systems*, 2019b.

Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze the discriminator: a simple baseline for fine-tuning gans. *arXiv preprint arXiv:2002.10964*, 2020.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. In *Advances in Neural Information Processing Systems*, 2020.

Junhyun Nam, Jaehyung Kim, Jaeho Lee, and Jinwoo Shin. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. In *International Conference on Learning Representations*, 2022.

Judea Pearl. An introduction to causal inference. *International Journal of Biostatistics*, 2010.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

Joni Salminen, Soon-gyo Jung, Shammur Chowdhury, and Bernard J. Jansen. Analyzing demographic bias in artificially generated facial pictures. In *Extended Abstracts of the 2020 CHI Conference on HUman Factors in Computing Systems*, 2020.

Prasanna Sattigeri, Samuel C Hoffman, Vijil Chenthamarakshan, and Kush R Varshney. Fairness gan. *arXiv preprint arXiv:1805.09910*, 2018.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, 2020.

Nimit Sohoni, Maziar Sanjabi, Nicolas Ballas, Aditya Grover, Shaoliang Nie, Hamed Firooz, and Christopher Ré. Barack: Partially supervised group robustness with guarantees. *arXiv preprint arXiv:2201.00072*, 2021.

Shuhan Tan, Yujun Shen, and Bolei Zhou. Improving the fairness of deep generative models without retraining. *arXiv preprint arXiv:2012.04842*, 2020.

Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2011.

Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in Neural Information Processing Systems*, 2020.

Boris van Breugel, Trent Kyono, Jeroen Berrevoets, and Mihaela van der Schaar. Decaf: Generating fair synthetic data using causally-aware generative networks. 2021.

Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *International Conference on Learning Representations*, 2020.

Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. In *International Conference on Learning Representations*, 2022.

Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks. In *IEEE International Conference on Big Data (Big Data)*, 2018.

Depeng Xu, Yongkai Wu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Achieving causal fairness through generative adversarial networks. In *International Joint Conferences on Artificial Intelligence*, 2019.

Ning Yu, Ke Li, Peng Zhou, Jitendra Malik, Larry Davis, and Mario Fritz. Inclusive GAN: Improving data and minority coverage in generative models. In *European Conference on Computer Vision*, 2020.

Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks. In *International Conference on Learning Representations*, 2022.

Shengjia Zhao, Hongyu Ren, Arianna Yuan, Jiaming Song, Noah Goodman, and Stefano Ermon. Bias and generalization in deep generative models: An empirical study. In *Advances in Neural Information Processing Systems*, 2018.

Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. *Advances in Neural Information Processing Systems*, 2020.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE/CVF International Conference on Computer Vision*, 2017.

A Implementation details

A.1 Classifiers

Network architectures. For CelebA and Waterbirds, we use the `torchvision` implementation of ResNet-50 starting from ImageNet-pretrained weights. For Colored MNIST, we use 4-layer convolutional network with channel sizes of 32, 32, 64, 64, kernel size of 5, 3, 3, 3, and stride of 2, 1, 2, 1.

Biased classifiers for FICS. It is well-known that ERM trained classifier trained with a biased dataset inherits spurious correlations existing in the dataset (Sagawa et al., 2020; Liu et al., 2021). Thus, we did not perform any additional technique to amplify the spurious correlation. To handle a class imbalance in the dataset, we use group DRO (Sagawa et al., 2020) to ensure low worst-class error. We use the SGD optimizer with a learning rate of 0.001, momentum 0.9 for 100 epochs with batch size 64. In addition, we use ℓ_2 regularization of 0.0005 and early stopping for the best validation accuracy.

Attribute classifiers for evaluation. For the attribute classifiers for evaluation, we train the models with group DRO to ensure low worst-group error. We use the SGD optimizer with a learning rate of 0.0001, momentum 0.9 for 15 epochs with batch size 64. In addition, we use ℓ_2 regularization of 0.01 and early stopping for the best validation accuracy. Table 5, 6 reports the accuracy of attribute classifiers on the CelebA and Waterbirds dataset, respectively. The attribute classifier for classifying color on the Colored MNIST dataset showed 100% accuracy for all groups.

Table 5: Accuracy of the attribute classifiers used for evaluation on the CelebA dataset.

Attribute of interest	Blond		Non-blond	
	Present	Absent	Present	Absent
Male	86.11	97.18	97.20	96.08
Eyeglasses	96.77	99.30	92.91	99.56
Lipstick	82.47	83.79	86.83	92.10

Table 6: Accuracy of the attribute classifiers used for evaluation on the Waterbirds dataset.

Attribute of interest	Waterbirds		Landbirds	
	Water	Land	Water	Land
Background	94.74	90.98	89.70	95.93

Reference set classifiers for weak-supervision. For the reference set classifiers, we follow the details from the original paper (Choi et al., 2020). We use the Adam optimizer with a learning rate of 0.0001 without any ℓ_2 regularization for 15 epochs with batch size 64. We use the same reference set classifiers for the weakly-supervised baseline Choi et al. (2020) proposed and our weakly-supervised version. We use $\alpha = 0.3$ for target male ratio 30% and $\alpha = 0.5$ for target male ratio 50% for the weakly-supervised baseline.

Attribute classifiers for semi-supervision. We use 4,000 samples with the sensitive attribute annotation and the rest of the samples in the training set without the sensitive attribute annotation to train the sensitive attribute classifier. We use FixMatch (Sohn et al., 2020), a recent SOTA semi-supervised learning technique, to train the sensitive attribute classifier. We only use random horizontal flip for data augmentation and use a fixed threshold of 0.95 for τ. We use the SGD optimizer with a learning rate of 0.001, momentum 0.9 for 100 epochs with batch size 64. In addition, we use ℓ_2 regularization of 0.005 and early stopping for the best validation accuracy.

A.2 Generative models

Network architectures. For generative model, we implement our code based on `PyTorch-StudioGAN` repository[2]. We use BigGAN for the CelebA and Waterbirds experiments, and ResNetGAN with spectral normalization (Miyato et al., 2018) for the Colored MNIST experiments. For CelebA and Colored MNIST, we train our generative model from scratch. For Waterbirds, we train our model starting from ImageNet-pretrained weights provided by `PyTorch-StudioGAN` repository.

Training details. For all the baselines and our method on the CelebA dataset, we train BigGAN with the same configurations. We train the models using the Adam optimizer with learning rate of 0.0002, $\beta_1 = 0$, $\beta_2 = 0.999$, for total 200k iterations with batch size 32. We update the discriminator 4 times per one generator step. For our Waterbirds experiments, we train models using the Adam optimizer with learning rate of 0.0001 for the generator and 0.0002 for the discriminator, $\beta_1 = 0$, $\beta_2 = 0.999$, for total 20k iteration with batch size 128. We update the discriminator 2 times per one generator step. We use DiffAug (Zhao et al., 2020) to supplement the limited number of training data. For the Colored MNIST experiments, we train the models using the Adam optimizer with learning rate of 0.0002, $\beta_1 = 0.5$, $\beta_2 = 0.999$, for total 20k iterations with batch size 64. We update the discriminator 5 times per one generator step.

Dia-GAN. For the CelebA dataset, we train 150k iterations for phase 1 and use log density ratios from 140k to 150k iterations to reweight samples for phase 2. For the Waterbirds dataset, we train 5k iterations for phase 1 and use log density ratios from 500 to 5 iterations to reweight samples for phase 2. For the Colored MNIST dataset, we train 4k iteration for phase 1 and use log density ratios from 100 to 4k iterations to reweight samples for phase 2. Following the original paper, we clip the sample weights from 0.01 to 0.5.

FICS. For the CelebA dataset, we first pretrain 100k iterations with the vanilla configurations and then finetune for the next 100k iterations. For finetuning, we freeze the discriminator until 4 layers using FreezeD (Mo et al., 2020) and add the proposed regularizer. We also start to resample after 100k iterations when supervision is provided. For the Waterbirds and Colored MNIST dataset, we did not perform any additional finetuning before FICS training. For the λ of FICS regularizer, we tuned over $\{0.002, 0.005, 0.01, 0.02\}$ and use the values with the best FID.

B Baselines

We extend the prior (unconditional) fair generation methods as the baselines for the fair conditional generation, i.e., modify their formula from $p(x)$ to $p(x, y)$. We consider three representative methods for unsupervised (Lee et al., 2021), supervised (Tan et al., 2020), and weakly-supervised (Choi et al., 2020) scenarios.

Unsupervised. Unsupervised fair generation aims to promote underrepresented samples and reduce the gap between data and generator distributions. For instance, Dia-GAN (Lee et al., 2021) resamples the training data based on the log-density ratio $\log p_{\mathsf{data}}(x)/p_{\mathsf{gen}}(x)$ estimated by the discriminator.

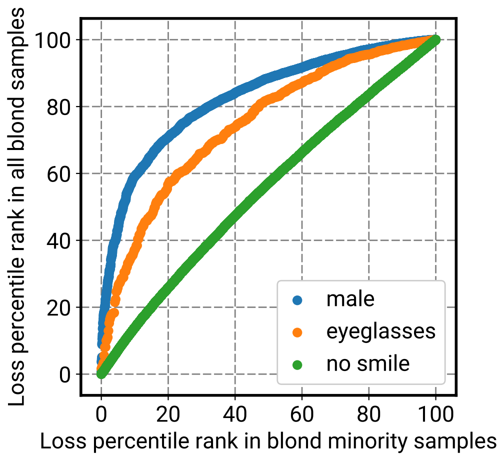
Supervised. Given annotations of sensitive attribute $a \in \mathcal{A}$, a straightforward solution for a fair conditional generation is conditioning both y and a, as one can explicitly control the generated attributes using $p(x|y, a)$. However, this joint-condition approach is hard to train when the data is imbalanced; too scarce for the minority group (y, a). As an alternative, FairGen (Tan et al., 2020) propose a latent modification approach that modifies an unconditional generative model to a conditional model by learning a prior distribution $p_a(z)$ that restricts the samples under $p(x|a)$, unlike the original prior $p(z)$ sampling $p(x)$. Note that the supervised method performs well if the target attribute a is known; however, it requires heavy annotation costs and cannot prevent the unknown biases.

Weakly-supervised. Choi et al. (2020) consider a weaker form of the fairness supervision, assuming a reference set $\mathcal{S}_{\mathsf{ref}}$ with empirical distribution $p_{\mathsf{ref}}(x) \approx p_{\mathsf{fair}}(x)$. Specifically, they reweighted the training data with density ratio $p_{\mathsf{ref}}(x)/p_{\mathsf{data}}(x)$ estimated by a classifier distinguishing the reference and data distributions. Note that the weakly-supervised method heavily depends on the collection of the reference set; it is often unrealistic to collect the oracle reference set in practice.

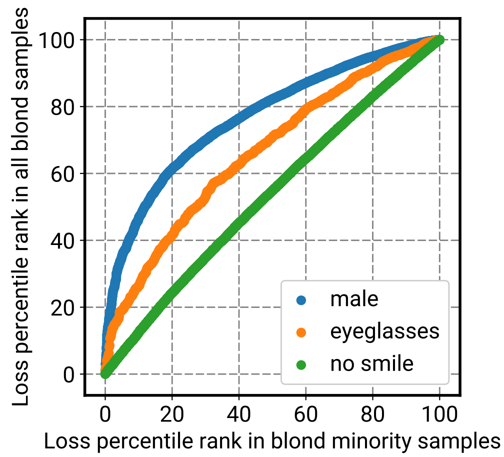
[2] https://github.com/POSTECH-CVLab/PyTorch-StudioGAN

C Additional experiments

C.1 The relationship between the biased classifier and the generator



(a) Vanilla BigGAN (Brock et al., 2019)



(b) FICS (ours)

Figure 6: The relationship between the biased classifier and the generator represented through classification loss percentile ranks of generated samples. Given a point, its x, y coordinate represents the percentile rank of its classification loss among blond minority samples and all blond samples respectively. The correlation between the classification loss and the chance to be minority still remains after the FICS training.

To validate the effectiveness of our proposed regularization approach, we conducted an investigation into the correlation between the classification loss and the probability of a sample being a minority at the end of the FICS training. We present the results in Figure 5, where we plot the classification loss percentile ranks as shown in Figure 3. As the number of minority samples generated by the FICS training increased, we observed that a sample with a loss ranking in the 20th percentile among blond male samples had a relatively lower rank in the FICS-trained cGAN compared to the vanilla cGAN. However, we found that the correlation between the classification loss and the likelihood of a sample being a minority persisted even after the FICS training. Based on these findings, we can conclude that our regularizer, which is based on the observed correlation, consistently encouraged the cGAN to generate minority samples throughout the entire training process.