Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, pp. 11895–11907, 2019a.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32, pp. 11918–11930. Curran Associates, Inc., 2019b. URL https://proceedings.neurips.cc/paper/2019/file/3001ef257407d5a371a96dcd947c7d93-Paper.pdf.

Song, Y. and Ermon, S. Improved techniques for training score-based generative models. *arXiv preprint arXiv:2006.09011*, 2020.

Tan, S., Shen, Y., and Zhou, B. Improving the fairness of deep generative models without retraining. *arXiv preprint arXiv:2012.04842*, 2020.

Terhörst, P., Kolf, J. N., Damer, N., Kirchbuchner, F., and Kuijper, A. Face quality estimation and its correlation to demographic and non-demographic bias in face recognition. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–11. IEEE, 2020.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9446–9454, 2018.

Wang, J., Liu, Y., and Levy, C. Fair classification with group-dependent label noise. *arXiv preprint arXiv:2011.00379*, 2020a.

Wang, S., Guo, W., Narasimhan, H., Cotter, A., Gupta, M., and Jordan, M. I. Robust optimization for fairness with noisy protected groups. *arXiv preprint arXiv:2002.09343*, 2020b.

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 Teh, Y. W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 681–688. Citeseer, 2011.

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.

Yang, F., Cisse, M., and Koyejo, S. Fairness with overlapping groups. *arXiv preprint arXiv:2006.13485*, 2020.

Yu, N., Li, K., Zhou, P., Malik, J., Davis, L., and Fritz, M. Inclusive gan: Improving data and minority coverage in generative models. In *European Conference on Computer Vision*, pp. 377–393. Springer, 2020.

## A. FFHQ Experiments

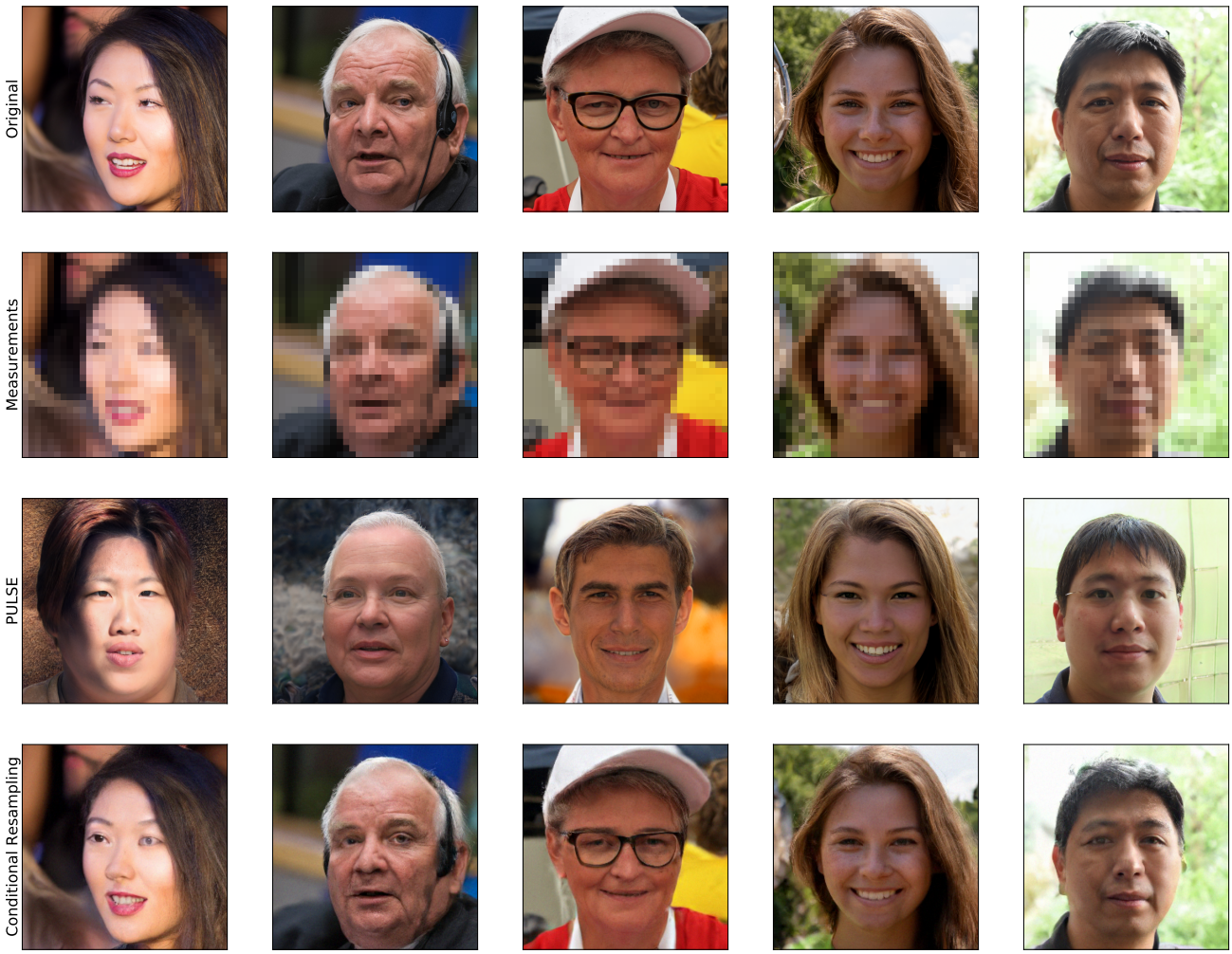



*Figure 5.* Super-resolution reconstructions on faces 69000-69004 from the FFHQ dataset. The top row shows original images, the second row shows what the algorithms observe: blurry measurements after downsampling by $32\times$ in each dimension. The third row shows reconstructions by PULSE, and the last row shows reconstructions by Posterior Sampling via Langevin dynamics, the algorithm we are advocating for.

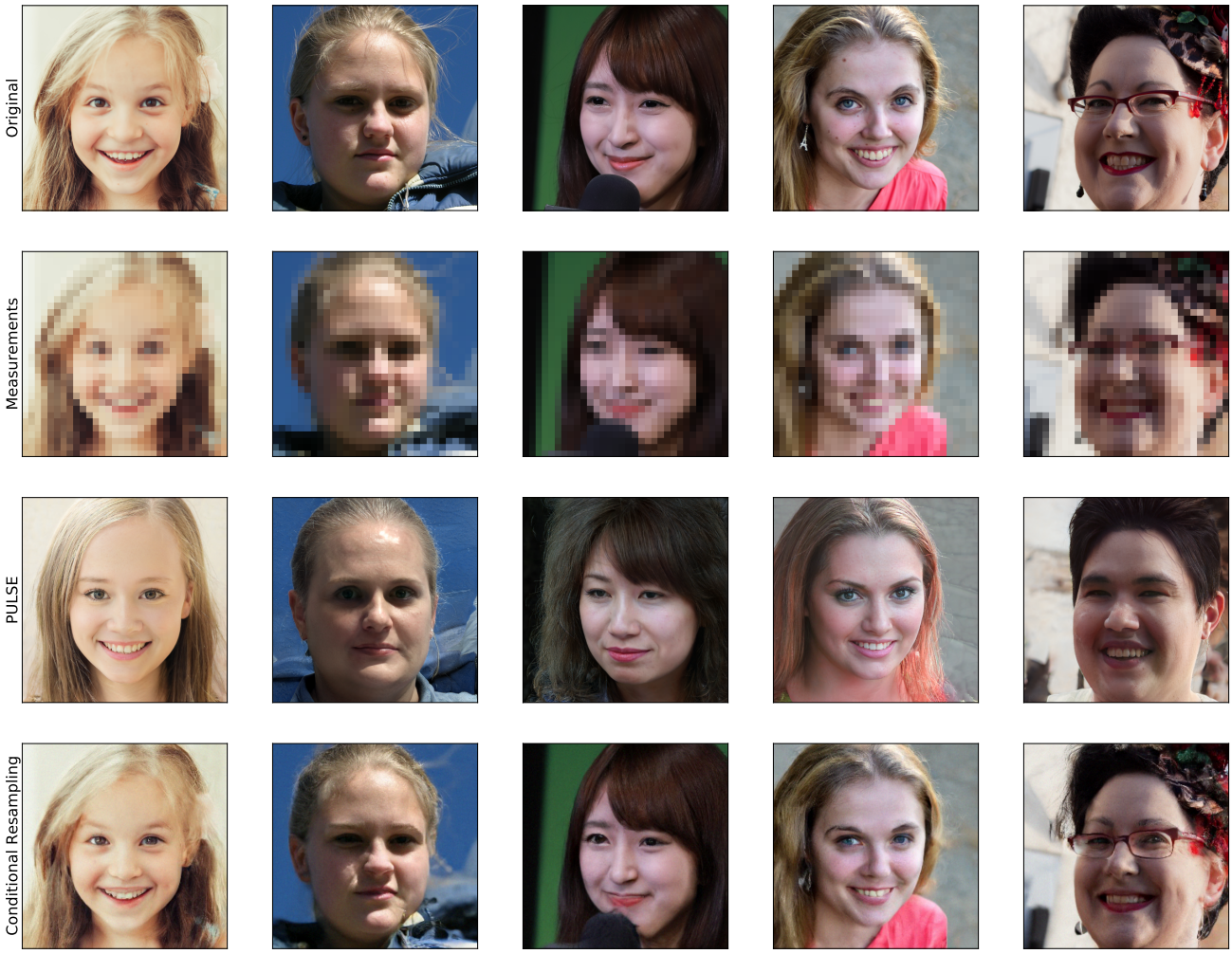



*Figure 6.* Super-resolution reconstructions on faces 69005-69009 from the FFHQ dataset. The top row shows original images, the second row shows what the algorithms observe: blurry measurements after downsampling by $32\times$ in each dimension. The third row shows reconstructions by PULSE, and the last row shows reconstructions by Posterior Sampling via Langevin dynamics, the algorithm we are advocating for.


*Figure 7.* Super-resolution reconstructions on faces 69010-69014 from the FFHQ dataset. The top row shows original images, the second row shows what the algorithms observe: blurry measurements after downsampling by $32\times$ in each dimension. The third row shows reconstructions by PULSE, and the last row shows reconstructions by Posterior Sampling via Langevin dynamics, the algorithm we are advocating for.

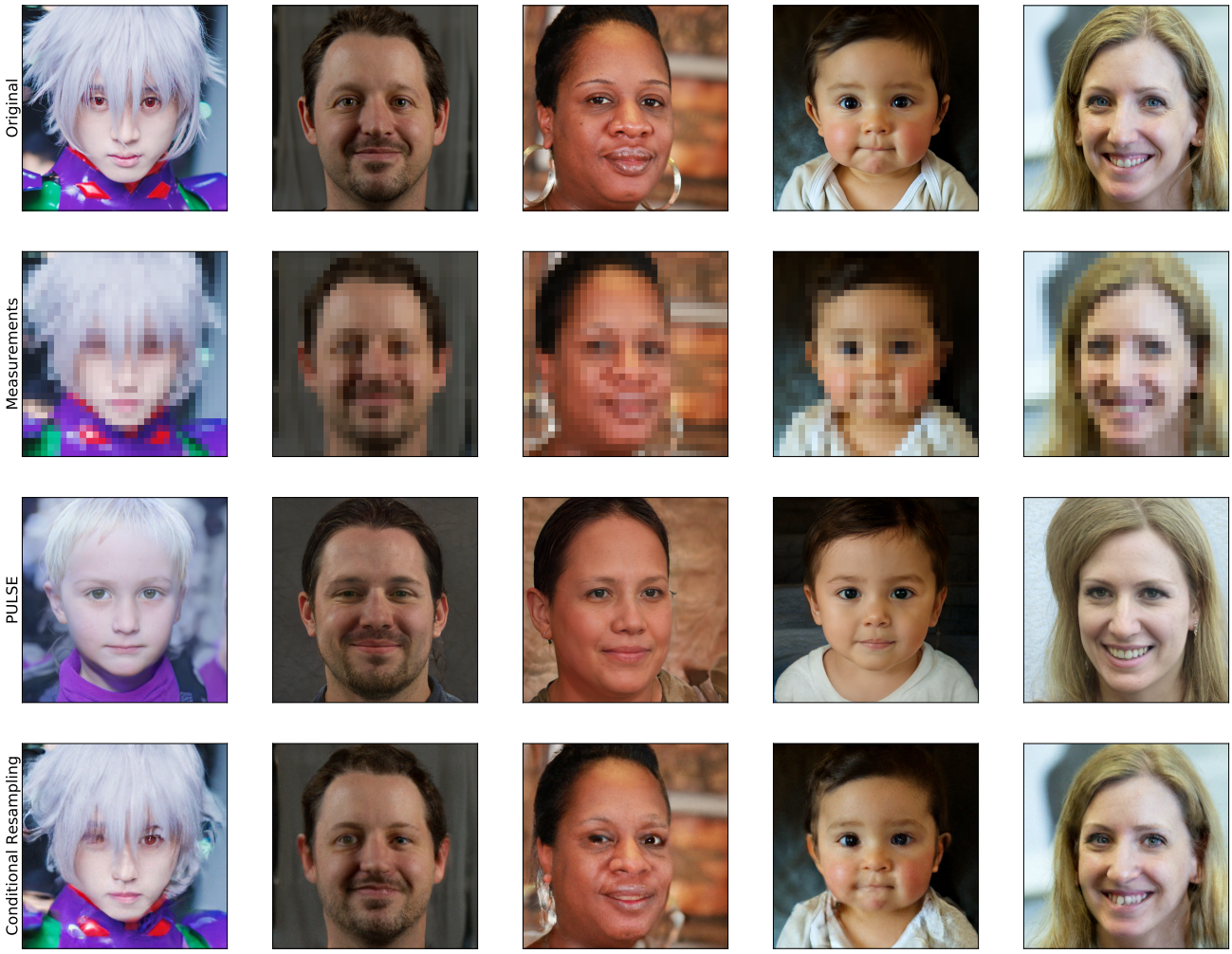
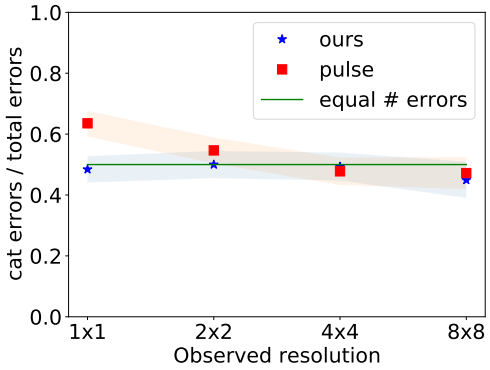



*Figure 8.* Super-resolution reconstructions on faces 69015-69020 from the FFHQ dataset. The top row shows original images, the second row shows what the algorithms observe: blurry measurements after downsampling by $32\times$ in each dimension. The third row shows reconstructions by PULSE, and the last row shows reconstructions by Posterior Sampling via Langevin dynamics, the algorithm we are advocating for.

| $m$ | PULSE | | Ours | |
|---|---|---|---|---|
| | Cats | Dogs | Cats | Dogs |
| $1 \times 1$ | 319 | 183 | 245 | 261 |
| $2 \times 2$ | 282 | 234 | 239 | 239 |
| $4 \times 4$ | 225 | 246 | 223 | 229 |
| $8 \times 8$ | 160 | 179 | 119 | 146 |

(a) Number of errors. Test set has **500** cats and **500** dogs




(b) Fraction of all errors on cats for 50% cat generator.

*Figure 9.* We use a StyleGAN2 model trained on 50% cats and report errors when reconstructing images from low-resolution measurements. The test set consists of 500 cats and 500 dogs from the AFHQ validation set to mimick the generator's training distribution (note that these correspond to all cats and dogs in the AFHQ validation set). Figure (b) shows the proportion of all errors that are on cats, along with 95% confidence intervals from a binomial test. An algorithm that satisfies SPE would have this probability=0.5 (green line). In this case where the generator is balanced, Posterior Sampling via Langevin dynamics appears to achieve SPE, PR, and RDP. PULSE also appears to satisfy SPE, PR, and RDP, except when the resolution of measurements is $1 \times 1$.

## B. AFHQ Experiments

### B.1. 50% cat generator

For this experiment, we draw $x^*$ from the validation set of the AFHQ dataset which contains 500 images of cats + 500 images of dogs. We use a generator trained on 50% cats and 50% dogs, and use it to study whether posterior sampling and PULSE satisfy RDP, SPE, and PR in practice. These results are in Figure 9.

### B.2. $x^*$ drawn from generator

In Figure 10, we show results when 200 images drawn from the 20% cat generator are reconstructed.

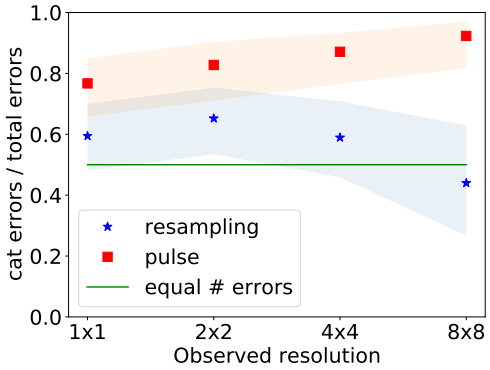
In Figure 11, we show results when 200 images drawn from the 80% cat generator are reconstructed.

### B.3. Varying training bias

We train StyleGAN2 models with 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% cats, and report the fraction of errors on cats when tested on the AFHQ validation set. The results are in Figure 12.

| $m$ | PULSE | | Ours | |
|---|---|---|---|---|
| | Cats | Dogs | Cats | Dogs |
| $1 \times 1$ | 56 | 17 | 44 | 30 |
| $2 \times 2$ | 48 | 10 | 45 | 24 |
| $4 \times 4$ | 54 | 8 | 33 | 23 |
| $8 \times 8$ | 48 | 4 | 11 | 14 |

(a) Number of errors on 20% cat generator, for each resolution. Sampled test set has **60** cats and **140** dogs. PULSE makes errors on almost all the cats and relatively few dogs, while Posterior Sampling is relatively balanced.

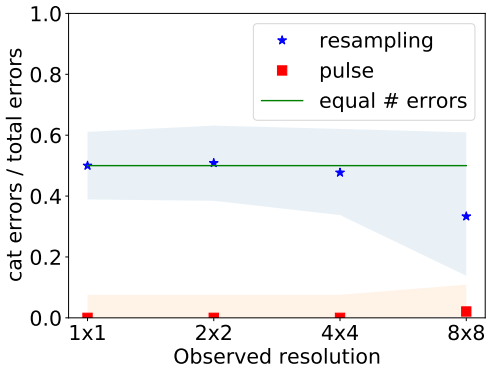



(b) Binomial hypothesis test for Symmetric Pairwise Error (SPE)

*Figure 10.* We sample 200 images from a StyleGAN2 model trained on 20% cats, and report errors when reconstructing them from low-resolution measurements. Figure (b) shows the proportion of all errors that are on cats, along with 95% confidence intervals from a binomial test. An algorithm that satisfies SPE would have this probability=0.5 (green line). PULSE is clearly biased towards the majority, while Posterior Sampling via Langevin dynamics appears to satisfy SPE (except when $m = 2 \times 2$, but one failure is unsurprising as we are performing sequential hypothesis tests.)

| $m$ | PULSE | | Ours | |
|---|---|---|---|---|
| | Cats | Dogs | Cats | Dogs |
| $1 \times 1$ | 0 | 47 | 37 | 37 |
| $2 \times 2$ | 0 | 47 | 30 | 29 |
| $4 \times 4$ | 0 | 47 | 21 | 23 |
| $8 \times 8$ | 1 | 47 | 4 | 8 |

(a) Number of errors on 80% cat generator, for each resolution. Sampled test set has **153** cats and **47** dogs. PULSE makes errors on almost all the cats and relatively few dogs, while posterior sampling is relatively balanced.




(b) Binomial hypothesis test for Symmetric Pairwise Error (SPE)

*Figure 11.* We sample 200 images from a StyleGAN2 model trained on 80% cats, and report errors when reconstructing them from low-resolution measurements. Figure (b) shows the proportion of all errors that are on cats, along with 95% confidence intervals from a binomial test. An algorithm that satisfies SPE would have this probability=0.5 (green line). PULSE is clearly biased towards the majority, while posterior sampling via Langevin dynamics appears to satisfy SPE.

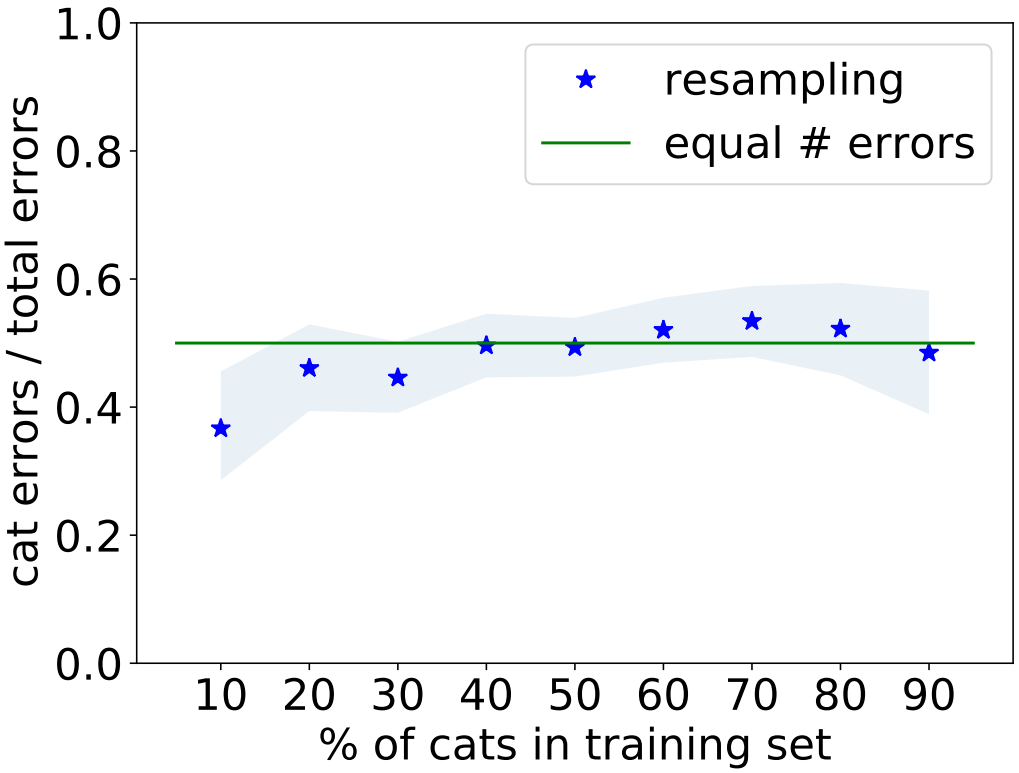



*Figure 12.* We train StyleGAN2 generators of varying bias and test SPE. The ground truth images are from the validation set, the observed measurements have resolution $4 \times 4$. Shaded areas denote 95% confidence intervals. We see that Posterior Sampling satisfies SPE. Note that the single failure in the 10% cat generator is not surprising as we are running sequential hypothesis tests on non-independent data.

## C. Proofs

**Theorem 2.4.** *Let $A$ and $B$ be disjoint groups (e.g., Asian and White people), and let $A_1, A_2 \subset A$ be disjoint groups that cannot be perfectly distinguished from measurements only (e.g., South Asians and East Asians). Then Representation Demographic Parity cannot be satisfied $\{\{A, B\}, \{A_1, A_2, B\}\}$-obliviously.*

*Proof.* Let $A = A_1 \cup A_2$. We write $p_a = \Pr(x^* \in a)$, $q_{ab} = \Pr(\hat{x} \in b | x^* \in a)$. Using Representation Demographic Parity, with respect first to $\{A, B\}$, then to $\{A_1, A_2, B\}$, we have:

$$q_{AA} = q_{BB}$$
$$q_{A_1A_1} = q_{A_2A_2} = q_{BB}$$

Since $A = A_1 \cup A_2$:

$$q_{AA} = \frac{p_{A_1}(q_{A_1A_1} + q_{A_1A_2}) + p_{A_2}(q_{A_2A_1} + q_{A_2A_2})}{p_{A_1} + p_{A_2}}$$

Writing $0 < \frac{p_{A_1}}{p_{A_1}+p_{A_2}} = \alpha < 1$, and replacing $q_{AA}, q_{A_1A_1}$ and $q_{A_2A_2}$ by $q_{BB}$, we have:

$$\begin{aligned} q_{AA} &= \alpha(q_{A_1A_1} + q_{A_1A_2}) + (1-\alpha)(q_{A_2A_1} + q_{A_2A_2}) \\ q_{BB} &= \alpha(q_{BB} + q_{A_1A_2}) + (1-\alpha)(q_{BB} + q_{A_2A_2}) \\ 0 &= \alpha q_{A_1A_2} + (1-\alpha) q_{A_2A_2}. \end{aligned}$$

Therefore, an algorithm can satisfy Representation Demographic Parity $\{\{A_1 \cup A_2, B\}, \{A_1, A_2, B\}\}$-obliviously if and only if there exists no confusion between $A_1$ and $A_2$, i.e. $q_{A_1A_2} = 0 = q_{A_2A_1}$. □

**Theorem 2.5** (Representation Demographic Parity cannot be satisfied obliviously)**.** *The only way for an algorithm to satisfy Representation Demographic Parity obliviously is to achieve perfect reconstruction.*

*Proof.* Suppose there exists $x$ such that $\Pr(x) > 0$, and $x_1 \neq x$ such that $\Pr(\hat{x} = x_1) > 0$. Let us split the space into two groups $A$ and $B$, such that both $x$ and $x_1$ belong in $A$. We now further split $A$ into $A_1$ and $A_2$, such that $x_1$ belongs in $A_1$, and $x$ belongs in $A_2$. $A_1$ and $A_2$ now are not perfectly distinguishable, so using the claim above, Representation Demographic Parity is not satisfiable $\{\{A_1 \cup A_2, B\}, \{A_1, A_2, B\}\}$-obliviously, so it cannot be satisfiable obliviously. □

**Proposition 2.8.** *Whenever there exists a majority class that the measurements cannot 100% distinguish from the non-majority classes, PR and RDP are not simultaneously achievable.*

*Proof.* Suppose towards a contradiction that both PR and RDP hold, and the distribution is such that

$$\Pr(x^* \in c_1) > \frac{1}{2} > \sum_{i \neq 1} \Pr(x^* \in c_i).$$

Since PR holds, $\Pr(\hat{x} \in c_1) = \Pr(x^* \in c_1)$. However, since RDP holds and the algorithm does not reconstruct each class perfectly we have $\alpha = \Pr(\hat{x} \in c_i \mid x^* \in c_i) < 1$ for all the $i$. We now observe the following contradiction.

$$\begin{aligned} \Pr(\hat{x} \in c_1) &\leq \sum_i \Pr(\hat{x} \in c_1 \mid x^* \in c_i) \Pr(x^* \in c_i) \\ &\leq (1-\alpha) \sum_{i \neq 1} \Pr(x^* \in c_i) + \alpha \Pr(x^* \in c_1) \\ &< (1-\alpha) \Pr(x^* \in c_1) + \alpha \Pr(x^* \in c_1) \\ &= \Pr(x^* \in c_1). \end{aligned}$$

□

**Theorem 3.1.** *Posterior Sampling is the only algorithm that achieves oblivious Conditional Proportional Representation.*

*Proof.* Let $\mathcal{A}$ denote a reconstruction algorithm. Given measurements $y$, let $Q(U|y)$ denote the probability that the reconstruction from algorithm $\mathcal{A}$ lies in the measurable set $U$.

If $\mathcal{A}$ satisfies CPR, then for all measurable $U \subset \mathbb{R}^n$, and all $y \in \mathbb{R}^m$, we have

$$Q(U|y) = P(U|y).$$

By the definition of the total variation distance, we have

$$TV(Q(\cdot|y), P(\cdot|y)) = \sup_{U \in \mathcal{B}(\mathbb{R}^n)} Q(U|y) - P(U|y).$$

Since we have $Q(U|y) = P(U|y)$ for all measurable $U \in \mathcal{B}(\mathbb{R}^n)$ and almost all measurements $y \in \mathbb{R}^m$, we have $TV(Q(\cdot|y), P(\cdot|y)) = 0$ for almost all $y \in \mathbb{R}^m$.

This shows that the output distribution of $\mathcal{A}$ must exactly match the posterior distribution $P(\cdot|y)$, and hence posterior sampling is the only algorithm that can satisfy obliviousness and CPR. □

**Theorem 3.3.** *In the setting of Definition 2.1, Conditional Proportional Representation implies Symmetric Pairwise Error.*

*Proof.* We want to show that if $\Pr(\widehat{x} \in c_i|y) = \Pr(x^* \in c_i|y), \forall c_i \in C$, for almost all $y \in \mathbb{R}^m$, then we have $\Pr(\widehat{x} \in c_i, x^* \in c_j) = \Pr(\widehat{x} \in c_j, x^* \in c_i), \forall c_i, c_j \in C$.

Consider the term $\Pr(\widehat{x} \in c_i, x^* \in c_j)$. We can write this as an average over $y$, to get:

$$\Pr(\widehat{x} \in c_i, x^* \in c_j) = \mathop{\mathbb{E}}_{y} \Pr(\widehat{x} \in c_i, x^* \in c_j|y).$$

Note that $\widehat{x} \& x^*$ are conditionally independent given $y$. This is because $\widehat{x}$ is purely a function of $y$. This gives

$$\Pr(\widehat{x} \in c_i, x^* \in c_j) = \mathop{\mathbb{E}}_{y}[\Pr(\widehat{x} \in c_i|y) \Pr(x^* \in c_j|y)].$$

If we have CPR with respect to $c_i$ and $c_j$, then we can rewrite the above equation as

$$\Pr(\widehat{x} \in c_i, x^* \in c_j) = \mathop{\mathbb{E}}_{y}[\Pr(x^* \in c_i|y) \Pr(\widehat{x} \in c_j|y)].$$

Using the conditional independence of $x^*, \widehat{x}$ given $y$, we now have

$$\begin{aligned}\Pr(\widehat{x} \in c_i, x^* \in c_j) &= \mathop{\mathbb{E}}_{y} \Pr(x^* \in c_i, \widehat{x} \in c_j|y), \\ &= \Pr(\widehat{x} \in c_j, x^* \in c_i).\end{aligned}$$

This completes the proof.

□

**Corollary 3.4.** *Posterior Sampling achieves symmetric pairwise error for any pair of sets $U, V \subset \mathbb{R}^n$.*

*Proof.* The proof follows directly from Theorem 3.1 and Theorem 3.3 □

**Theorem 3.5.** *Let $C = \{c_1, \ldots, c_k\}$ be a partition. There exists a choice of weights $\lambda_i > 0$ with $\sum \lambda_i = 1$ such that Posterior Sampling with respect to the reweighted distribution*

$$p_\lambda(x) = \sum_i \lambda_i p(x \mid x \in c_i)$$

*satisfies RDP with respect to $C$.*

In the special case of 2 classes, the reweighting is very simple: $\lambda_1 = \lambda_2 = \frac{1}{2}$.

*Proof.* We will prove this theorem by contradiction. Before we start, observe that if we scale the mass of $c_i$ by $\lambda_i \geq 0$, we have,

$$\begin{aligned}\alpha_i &:= \Pr(\hat{x} \in C_i \mid x^* \in C_i) \\ &= \mathop{\mathbb{E}}_{y|x^* \in C_i}\left[\frac{\lambda_i \Pr(x^* \in C_i \mid y)}{\sum_j \lambda_j \Pr(x^* \in C_j \mid y)}\right]\end{aligned}$$

WLOG, assume $\sum_i \lambda_i = 1$, this can be done by rescaling the $\lambda_i$'s by their sum. RDP is achieved if all the $\alpha_i$ are equal. Let the smallest $\alpha_i$ when all the $\lambda_i$'s are equal be $\epsilon$. Consider the set $T := \{\vec{\lambda} \mid \sum_i \lambda_i = 1, \forall i\ \alpha_i \geq \epsilon\}$.

Towards a contradiction, suppose no assignment of $\vec{\lambda} \in T$ achieves $f(\vec{\lambda}) := \frac{\max_i \alpha_i}{\min_j \alpha_j} = 1$. Let $r := \min_{\lambda_1,\ldots,\lambda_k} f(\vec{\lambda}) > 1$, and let $\vec{\lambda}^*$ be a point which achieves this. $\vec{\lambda}^*$ exists since $f(\vec{\lambda})$ is continuous over $T$, which is compact.

We will show that there exists $\vec{\lambda}' \in T$ such that $f(\lambda') < r$, which contradicts our hypothesis. Let $S := \{i \in [k] \mid \alpha_i \leq \sqrt{r}\min_i \alpha_i\}$ and,

$$\lambda'_i = \begin{cases} r^{1/4}\lambda^*_i, & \text{if } i \in S \\ \lambda^*_i, & \text{otherwise} \end{cases}$$

Let $\alpha'_i$ be $\Pr(\hat{x} \in C_i \mid x^* \in C_i)$ where the probability is with respect to the modified distribution. For $i \in S$,

$$\begin{aligned}\alpha'_i &= \mathop{\mathbb{E}}_{y|x^* \in C_i}\left[\frac{\lambda'_i \Pr(x^* \in C_i \mid y)}{\sum_j \lambda'_j \Pr(x^* \in C_j \mid y)}\right] \\ &= r^{1/4} \mathop{\mathbb{E}}_{y|x^* \in C_i}\left[\frac{\lambda_i \Pr(x^* \in C_i \mid y)}{\sum_j \lambda'_j \Pr(x^* \in C_j \mid y)}\right] \\ &\leq r^{1/4} \mathop{\mathbb{E}}_{y|x^* \in C_i}\left[\frac{\lambda_i \Pr(x^* \in C_i \mid y)}{\sum_j \lambda_j \Pr(x^* \in C_j \mid y)}\right]\end{aligned}$$

Where the last line follows from the fact that $\lambda'_i \geq \lambda_i$ for all $i$, since $r > 1$, which means the denominator only increases. A similar calculation shows that if $i \notin S$, then each $\alpha_i$ is multiplied by a factor between $r^{-1/4}$ and 1.

We notice that, by definition of $S$, all the $j$ such that $\alpha_j = \min_i \alpha_i$ are in $S$, and all the $k$ such that $\alpha_k = \max_i \alpha_i$ are not in $S$. This ensures that $\max_i \alpha'_i < \max_i \alpha_i$ and $\min_i \alpha'_i > \min_i \alpha_i \geq \epsilon$. Which, in turn, contradicts the hypothesis that $r$ was the smallest achievable ratio with the original constraints, since we can always renormalize $\lambda'_i$ without affecting the $\alpha_i$.

$$\frac{\max_i \alpha'_i}{\min_i \alpha'_i} < \frac{\max_i \alpha_i}{\min\{r^{-1/4}\sqrt{r}\min_i \alpha_i, \min_i \alpha_i\}} \leq r.$$

□

**Theorem 3.7.** *Let $C = \{c_1, \cdots, c_k\}$ form a disjoint partition of $\mathbb{R}^n$. An algorithm minimizes Representation Cross-Entropy on $C$ iff the algorithm satisfies CPR on $C$.*

*Proof.* The proof follows from Lemma C.1. Note that $H(U|Y)$ is a function of $x^*\&y$ and hence has no dependence on the reconstruction algorithm. By the non-negativity of $KL$ divergence, the representation cross-entropy is minimized when $Q(U_i|y) = P(U_i|y)$ for each $i \in [N]$, almost surely over $y$. □

**Lemma C.1.** *Let $U : \mathbb{R}^n \to \{c_1, c_2, \cdots, c_k\}$ be a function that encodes which group contains an image, and assume that the groups $c_1, \cdots, c_k \subset \mathbb{R}^n$ are disjoint and form a partition of $\mathbb{R}^n$.*

*For a reconstruction algorithm $\mathcal{A}$, let $Q(c_i|Y)$ denote the probability that the reconstruction lies in the set $c_i$ given measurements $y$. Let $P(c_i|y)$ denote the probability that $x^*$ lies in $U_i$ conditioned on $y$.*

*Then we have*

$$RCE(\mathcal{A}) = H_P\left(U|y\right) + \mathop{\mathbb{E}}_{y}\left[KL\left(P(U|y)\|Q(U|y)\right)\right],$$

*where*

$$H_P(U|y) := -\mathop{\mathbb{E}}_{y}\left[\sum_{i\in[k]} P(c_i|y)\log P(c_i|y)\right],$$
$$KL(P(U|y)\|Q(U|y)) := \sum_{i\in[k]} P(c_i|y)\log\left(\frac{P(c_i|y)}{Q(c_i|y)}\right).$$

**Remark:** There is a slight abuse of notation in the lemma. Since $U$ is a function of $x^*$, when treating $x^*$ as a random variable, we also treat $U$ as a random variable.

*Proof.* By the definition of $RCE$ and the tower property of expectations, we have

$$\begin{aligned}
-RCE(\mathcal{A}) &= \mathop{\mathbb{E}}_{x^*,y}\log\Pr\left[\widehat{x}\in U(x^*)|y\right] = \mathop{\mathbb{E}}_{y}\mathop{\mathbb{E}}_{x^*|y}\left[\log\left(\Pr\left[\widehat{x}\in U(x^*)|y\right]\right)\right],\\
&= \mathop{\mathbb{E}}_{y}\mathop{\mathbb{E}}_{x^*|y}\left[\sum_{i\in[N]}\mathbf{1}\left\{x^*\in U_i\right\}\log\left(\Pr\left[\widehat{x}\in U(x^*)|y\right]\right)\right],\\
&= \mathop{\mathbb{E}}_{y}\mathop{\mathbb{E}}_{x^*|y}\left[\sum_{i\in[N]}\mathbf{1}\left\{x^*\in U_i\right\}\log\left(\Pr\left[\widehat{x}\in U_i|y\right]\right)\right],\\
&= \mathop{\mathbb{E}}_{y}\mathop{\mathbb{E}}_{x^*|y}\left[\sum_{i\in[N]}\mathbf{1}\left\{x^*\in U_i\right\}\log\left(Q(U_i|y)\right)\right],\\
&= \mathop{\mathbb{E}}_{y}\left[\sum_{i\in[N]}P(U_i|y)\log\left(Q(U_i|y)\right)\right].(*)
\end{aligned}$$

where the second line follows because the $U_i$s form a partition, the third line follows since $\widehat{x}\in U(x^*)$ is equivalent to $\widehat{x}\in U_i$ if we know that $x^*\in U_i$, the fourth line follows from the definition of $Q(U_i|y)$ and the last line follows from linearity of expectation.

Now we can multiply and divide $P(U_i|y)$ within the $\log$ term above. This gives

$$\begin{aligned}
(*) &= \mathop{\mathbb{E}}_{y}\left[\sum_{i\in[N]}P\left(U_i|y\right)\log\left(\frac{Q(U_i|y)P(U_i|y)}{P(U_i|y)}\right)\right],\\
&= \mathop{\mathbb{E}}_{y}\left[\sum_{i\in[N]}P\left(U_i|y\right)\log\left(P(U_i|y)\right)\right]\\
&\quad + \mathop{\mathbb{E}}_{y}\left[\sum_{i\in[N]}P\left(U_i|y\right)\log\left(\frac{Q(U_i|y)}{P(U_i|y)}\right)\right],\\
&= -H\left(U|y\right) - \mathop{\mathbb{E}}_{y}\left[KL\left(P(U|y)\|Q(U|y)\right)\right].
\end{aligned}$$

This concludes the proof.

□

# D. Langevin Dynamics

## D.1. StyleGAN2

We want to sample from the distribution $p(x|y)$ induced by a StyleGAN2. Note that sampling from the marginal distribution $p(x)$ of a StyleGAN2 is achieved by sampling a latent variable $z \in \mathbb{R}^{512}$, and 18 noise variables $n_i \in \mathbb{R}^{d_i}$ of varying sizes, and setting $x = G(z, n_1, \cdots, n_{18})$. Hence, we can sample from $p(x|y)$ by sampling $\widehat{z}, \widehat{n}_1, \cdots, \widehat{n}_{18}$, from $p(z, n_1, \cdots, n_{18}|y)$, and setting $\widehat{x} = G(\widehat{z}, \widehat{n}_1, \cdots, \widehat{n}_{18})$.

The prior of the latent and noise variables is a standard Gaussian distribution. Since we know the prior distribution of these variables, if we know the distribution of the meaurement process, we can write out the posterior distribution.

For the measurement process we consider, we have $y = Ax^*$, where $A$ is a blurring matrix of appropriate dimension. Note that in the absence of noise, posterior sampling must sample solutions that exactly satisfy the measurements. However, this is difficult to enforce in practice, and hence we assume that there is some small amount of Gaussian noise in the measurements. In this case, the posterior distribution becomes:

$$p(z, n_1, \cdots, n_{18}|y) \propto p(y|z, n_1, \cdots, n_{18}) p(z, n_1, \cdots, n_{18}), \tag{1}$$

$$\Leftrightarrow \log p(z, n_1, \cdots, n_{18}|y) = -\frac{\|y - AG(z, n_1, \cdots, n_{18})\|^2}{2\sigma^2/m} - \|z\|^2/2 - \sum_{i=1}^{1} 8\|n_i\|^2/2 + c(y), \tag{2}$$

where $c(y)$ is an additive constant which depends only on $y$.

Now, Langevin dynamics tells us that if we run gradient ascent on the above log-likelihood, and add noise at each step, then we will sample from the conditional distribution asymptotically. Please note that we sample $z$ and *all noise variables* $n_1, \cdots, n_{18}$.

In our experiments, we do 1500 gradient steps. In practice, we replace the $\sigma$ in the equation above with $\sigma_t$, where $t$ is the iteration number. When the measurements have resolution $8 \times 8$ or $4 \times 4$, we find that $\sigma_1 = 1.0, \sigma_{1500} = 0.1$ works best. When the resolution of the measurements is $2 \times 2$ or $1 \times 1$, we find that $\sigma_1 = 1.0, \sigma_{1500} = 0.01$ works best. We change the value of $\sigma_t$ after every 3 gradient steps, such that $\sigma_1, \sigma_4, \sigma_7, \cdots, \sigma_{1497}$ form a geometrically decreasing sequence. The learning rate $\gamma_{1500}$ is also tuned to be a decreasing geometric sequence, such that $\gamma_t = 5 \cdot 10^{-6}$. Please see (Song & Ermon, 2019a) for the equations specifying the learning rate tuning, and the logic behind it.

We also find that adding a small amount of noise corresponding to $\sigma_{1500}$ in the measurements helps Langevin mix better.

We note that our approach is different from prior work (Karras et al., 2020b; Menon et al., 2020a), which optimizes a function of our variable $z$, and a subset of the noise variables.

**NCSNv2** The NCSNv2 model (Song & Ermon, 2020) has been designed such that sampling from the marginal distribution requires Langevin dynamics. This model is given by a function $s(x; \sigma)$, which outputs $\nabla \log p_\sigma(x)$, where $p_\sigma(x)$ is the distribution obtained by convolving the distribution $p(x)$ with Gaussian noise of variance $\sigma^2$. That is, $p_\sigma(x) = (p * \mathcal{N}(0, \sigma^2))(x)$.

It is easy to adapt the NCSNv2 model to sample from posterior distributions, see (Song & Ermon, 2020) for inpainting examples. In our super-resolution experiments, one can compute the gradient of $p(y|x)$, to get the following update rule for Langevin dynamics:

$$x_{t+1} \leftarrow x_t + \gamma_t(s(x_t; \sigma_t) - A^T(Ax_t - y)/\sigma_t^2) + \sqrt{2\gamma_t}\xi_t, \tag{3}$$

where $A$ is the blurring matrix, and $\xi_t \sim \mathcal{N}(0, I_n)$ is i.i.d. Gaussian noise sampled at each step. We use the default values of noise and learning rate specified in `https://github.com/ermongroup/ncsnv2/blob/master/configs/ffhq.yml`. That is, $\sigma_1 = 348, \sigma_{6933} = 0.01$, and $\gamma_{6933} = 9 \cdot 10^{-7}$. Note that the value of $\sigma, \gamma$ changes every 3 iterations, and both $\gamma_t$ and $\sigma_t$ decay geometrically. See (Song & Ermon, 2020) for specific details on how these are tuned.

# E. Code

All code and generative models, along with hyperparameters and README are available at `https://github.com/ajiljalal/code-cs-fairness`.