Supplemental Material
When does GAN replicate? An indication on the choice of dataset size

S1. Summary

Section S2 provides extended qualitative results on the GANs replication experiments in the main paper.

Section S3 provides additional results for the effect of threshold α on the dataset ID-replication curves.

Section S4 provides additional results for the effect of training image resolution on dataset ID-replication curves.

Section S5 shows supporting empirical evidence on our practices to use downscaled images during the calculation of Intrinsic Dimensionality.

Section S6 shows replication results defined in semantic embedding space rather than the native pixel space used in the main paper.

Section S7 provides additional results on MNIST dataset, which shows the limitation of our method when the dataset is extremely simple.

Section S8 shows results comparison between the FID scores and perceptual image quality from our AMT experiment.

Section S9 describes our Amazon Mechanical Turk experiment in details.

S2. Extended Qualitative Results for GANs Replication

Figure S4 and Figure S5 show extended qualitative results of BigGAN and StyleGAN2 replication experiments for CelebA, Flower and LSUN (bedroom) datasets in the main paper. All images are randomly generated without cherry-picking. These results indicate that for a given GAN architecture and dataset, when the dataset size is small, the GAN can generate almost exact replication of training data. The replication is gradually alleviated when the dataset size increases.

S3. Effects of threshold α on GANs Replication

In the main paper, we studied the relationship between the dataset size/complexity and GANs replication where the definition of replication is given by

$$P_\alpha(G, d, \mathcal{X}) = \Pr\left(\left[\min_{\mathbf{X}\in\mathcal{X}} d(G(z), \mathbf{X})\right] \leq \alpha\right) \quad \text{(S1)}$$

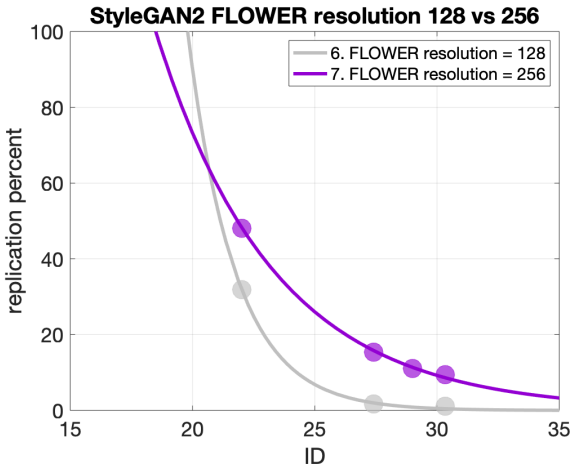

Figure S1: Replication curves for resolutions 128×128 and 256×256 for StyleGAN2 trained on Flower dataset. The trend remains exponential even with different resolutions.

with threshold α served as the acceptable noise level. Figure 3 of the main paper shows that when $\alpha = 8000$, the replication percentage shows a consistent exponential decay trend with respect to the dataset size and complexity, with a shared exponential decay factor. One may now wonder, how does the noise threshold α affect this trend? In this section, we will explore the effect of α[1].

Figure S6 shows the replication-complexity curve same as Figure 3 in the main paper but with threshold at $\alpha = 7000, 8000, 9000, 10000$, which shows that decreasing α will leads to a faster decrease of replication percentage. Table S1 shows the parameters a, c and b estimated for all the α's. The goodness-of-fit measurement is provided in Table S2.

These results show that trend discovered in the main paper still holds across different thresholds, which means the proposed one-shot prediction will perform similarly well even for different α's. This further implies that for anyone using our one-shot prediction method, the deciding factor of

[1]We also use α in the Equation 7, as part of the parameters in the modeling, which is not what we are examining here. This confusion in notation will be addressed in the final version of the main paper.

GAN	Datasets	α	$\hat{a}$	$\hat{b}$	$\hat{c}$
BigGAN	Flower	7000	0.96	92.36	100.00
StyleGAN2	Flower	7000	0.98	248.35	100.00
BigGAN	CelebA	7000	0.99	500.00	100.00
StyleGAN2	CelebA	7000	0.99	491.60	100.00
BigGAN	LSUN	7000	0.98	78.71	100.00
StyleGAN2	LSUN	7000	0.98	95.25	100.00
BigGAN	Flower	8000	0.96	62.93	100.00
StyleGAN2	Flower	8000	0.97	116.38	100.08
BigGAN	CelebA	8000	0.98	130.00	100.00
StyleGAN2	CelebA	8000	0.97	73.65	100.00
BigGAN	LSUN	8000	0.97	36.40	101.66
StyleGAN2	LSUN	8000	0.97	51.55	100.37
BigGAN	Flower	9000	0.96	43.19	100.00
StyleGAN2	Flower	9000	0.97	61.30	100.62
BigGAN	CelebA	9000	0.98	69.72	100.30
StyleGAN2	CelebA	9000	0.97	32.27	101.81
BigGAN	LSUN	9000	0.97	25.11	101.26
StyleGAN2	LSUN	9000	0.97	31.00	101.08
BigGAN	Flower	10000	0.96	30.52	100.15
StyleGAN2	Flower	10000	0.96	38.94	100.52
BigGAN	CelebA	10000	0.97	21.65	100.71
StyleGAN2	CelebA	10000	0.96	15.21	100.20
BigGAN	LSUN	10000	0.96	20.46	100.61
StyleGAN2	LSUN	10000	0.96	17.70	100.70

Table S1: Estimated parameters of the exponential decay model for different thresholds α. Under different threshold α, parameter a,b and c estimated from exponential relationship between dataset ID and GAN replication percentages for StyleGAN2 and BigGAN trained on CelebA, LSUN-bedroom and Flower datasets. For each α value, despite with different datasets and GAN architectures, the complexity-replication curves share similar exponential decay factor a and predictor translation c. For a given dataset-GAN combination, b decreases as α increases.

GAN	Datasets	α	R^2	MAE
BigGAN	Flower	7000	0.9942	0.9029
StyleGAN2	Flower	7000	0.9999	0.0308
BigGAN	CelebA	7000	0.9704	2.6241
StyleGAN2	CelebA	7000	0.9970	1.2866
BigGAN	LSUN	7000	0.9606	1.7540
StyleGAN2	LSUN	7000	0.9992	0.5416
BigGAN	Flower	8000	0.9739	2.8855
StyleGAN2	Flower	8000	0.9994	0.2144
BigGAN	CelebA	8000	0.9388	5.1180
StyleGAN2	CelebA	8000	0.9965	1.8955
BigGAN	LSUN	8000	0.8612	6.0261
StyleGAN2	LSUN	8000	0.9930	2.4826
BigGAN	Flower	9000	0.9250	8.2542
StyleGAN2	Flower	9000	0.9990	6.2406
BigGAN	CelebA	9000	0.8638	4.5048
StyleGAN2	CelebA	9000	0.9993	17.2335
BigGAN	LSUN	9000	0.7828	13.3370
StyleGAN2	LSUN	9000	0.9733	6.5316
BigGAN	Flower	10000	0.8600	12.0084
StyleGAN2	Flower	10000	0.9969	1.4222
BigGAN	CelebA	10000	0.6800	11.6200
StyleGAN2	CelebA	10000	0.9980	0.8923
BigGAN	LSUN	10000	0.7911	14.2998
StyleGAN2	LSUN	10000	0.9339	7.9402

Table S2: R^2 and MAE of the models across different thresholds α. R^2: goodness-of-fit measurement. MAE: median absolute errors. The model at each row corresponds to the ones in Table S1. Although there is a minor decrease of R^2 with the increase of α, the overall R^2 values are still close to one, showing that the model is highly effective on approximating the relationship between dataset complexity and GAN replication, independent from the threshold α.

α should be purely driven by the strictness of the replication with minimal concerns on the performance of the one-shot predictor.

S4. Effect of dataset resolution on GANs replication

One may also wonder the relationship between the image resolution to the replication curve, as we only showed results of 128×128 in the main paper. Figure S1 shows the replication results for StyleGAN2 trained on Flower dataset in 128×128 v.s. 256×256 resolutions. As shown in the figure, the exponential trend remains when we increase the resolution to 256.

S5. Dataset ID under Different Image Sizes

To calculate Intrinsic Dimensionality (ID) of the images, we downscale images from 128×128 to 32×32 to save computational resources. In this section, we provide supporting experiments on comparing the dataset ID between these two resolutions across all the datasets used in the main paper.

Table S3 provides the results of this comparison, which shows that the dataset ID for low-resolution images does not differ significantly from their high-resolution counterparts, albeit systematically lower as expected, which means that the observed trend of exponential decay shown in the main paper with 32×32 will not change once the ID is calculated with high-resolution samples.

Datasets	Subset size	ID	
		32×32	128×128
Flower	1000	22.02	22.66
Flower	2000	24.70	25.77
Flower	4000	27.41	28.30
Flower	6000	28.99	30.30
Flower	8189	30.34	31.49
CelebA	200	11.90	12.23
CelebA	600	15.97	16.50
CelebA	1000	17.30	18.01
CelebA	4000	21.34	22.23
CelebA	8000	23.29	24.28
LSUN	200	14.87	15.33
LSUN	1000	20.80	21.64
LSUN	5000	27.06	28.31
LSUN	10000	29.57	30.94
LSUN	30000	33.60	35.26

Table S3: Comparison on Intrinsic Dimensionality (ID) calculated with 32×32 and 128×128 samples, for each dataset and subset levels used in the main paper.

S6. Replications in Semantic Spaces

In Section 5 of the main paper, the replication is defined in RGB pixel space since any replication in the RGB space will imply replication in other commonly used semantic spaces, but not *vice versa*. In this section, to further illustrate this point, we provide qualitative results for replications defined in other semantic spaces, including the InceptionV3 [4] semantic embedding and SimCLR contrastive embedding [2]. The former is widely used as for visual semantic representation for natural images and is also part of FID calculation [3]. The latter is shown to be an effective embedding (learnt with self-supervision) for a variety of downstream tasks.

S6.1. InceptionV3 Semantic Space

Our method to define the InceptionV3 semantic space is the same as in FID calculation. More specifically, for each image, synthetic or real, we pass it to the InceptionV3 network pre-trained on ImageNet after center-croping, resizing and normalization. The 2048 dimensional feature output from the last pooling layer is used as the final embedding vector. The replication of a query (generated) image is then defined as its nearest neighbour in the InceptionV3 embedding space using Euclidean distance whose distance to the query image is smaller than the threshold α.

We provide qualitative comparison of the replication in RGB space to the ones in InceptionV3 semantic space for the Flower dataset (at subset level=1000) in Figure S8. We used $\alpha = 15$ for InceptionV3 space which yields a comparable replication percentage as $\alpha = 10000$ in RGB space (87.11% and 89.75% for BigGAN, and 66.89% and 63.08% for StyleGAN2, in Inception and RGB space respectively). The replications found in RGB space gives nearly perfect matching (thus a perceptual replication), while features in InceptionV3 space can only capture part of the image feature and thus leads to unsatisfactory matching results.

S6.2. SimCLR Contrastive Space

Features in SimCLR space are acquired by passing images into SimCLR network. The network is trained on each dataset used in our experiments without pre-training. Besides real images, synthetic images are also used during the training to improve its embedding quality. We define the positive pair as two transformed versions of a single image. The transformations we allowed are Gaussian blur and small affine transformations [2]. The negative pair is defined by any two different images in a mini-batch. We use the embedding after the encoder before the projection head as our feature. Each feature is of 2048 dimensions.

We provide qualitative results of the replication in RGB space to the ones in SimCLR semantic space for the FLOWER dataset (at subset level=1000) in Figure S9. We used $\alpha = 0.35$ for SimCLR space which yields a comparable replication percentage as $\alpha = 10000$ in RGB space (91.99% and 89.75% for BigGAN, and 61.42% and 63.08% for StyleGAN2, in SimCLR and RGB space respectively). Similar to the results in the Inception space, the replications in SimCLR space can only capture part of the image characteristics and thus cannot be qualified as perceptual replications.

S6.3. Image2StyleGAN space

We also tested using the combination of pixel and semantic space metrics described in Image2StyleGAN paper [1] with proper weight and normalization. As shown in Figure S3, the replication results with Image2StyleGAN space is almost identical to the one obtained with the RGB-metric.

S7. Additional Results for MNIST dataset

We also conducted additional experiment on MNIST dataset with the same setup as described in the main paper. As shown in the Figure, both BigGAN and StyleGAN2 can produce exact replication even with the full dataset (the curve is approximately flat). This is due to the simplicity of the dataset (with ID=12.7 at full which is close ID=12.23 for CelebA-200). This prompts caution on applying our method for extremely simple dataset.

[2]For gaussian blur, we use kernel widith=51. For affine transformations, we use *RandomAffine* in torchvision package with degrees=10, translate=None, scale=None, shear=10. This yields an image perceptually similar to the original sample.

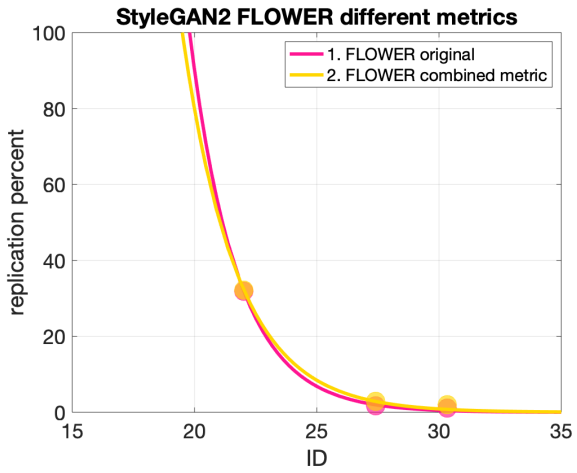

Figure S2: Replication curve when the distance metric is defined in Image2StyleGAN space (combined metric) v.s. RGB space (original).

S8. FID v.s. Perceptual Image Quality

To show the necessity of using Behavioral Experiment results instead of Fréchet Inception Distance (FID) as perceived image quality metric, Figure S7 compares the Amazon Mechanical Turk (AMT) rating with FID for StyleGAN2 and BigGAN on CelebA, LSUN-bedroom and Flower datasets under each subset level.

S9. Details for Amazon Mechanical Turk Experiment

In the main paper, we describe that the perceptual image quality is measured with a behavioral experiment on Amazon Mechanical Turk. The purpose of this experiment is to see how the perceived quality of the generated images changes with respect to subset levels.

For each generator trained at each subset level of a GAN-dataset combinations, we randomly generate 100 images. For each dataset, we also randomly select 100 real images for references.

Each of the real and synthesized images are rated by 9 AMTurkers. To ensure the quality of the rating, we limited the Workers to have HIT approval rate at least 95% and minimal 500 HIT's approved. Each HIT task contains total of 5 different images. A Worker needs to rate all 5 images to proceed to the next one. Maximum 10 minutes is allowed to complete each task. All the workers are compensated 0.05 USD (with 0.01 fee to AMT) for each task they completed.

We use 5-point Likert scale to measure the perception of image quality from human subjects. For each image, the following instruction is given,

Read the task carefully and inspect the image.

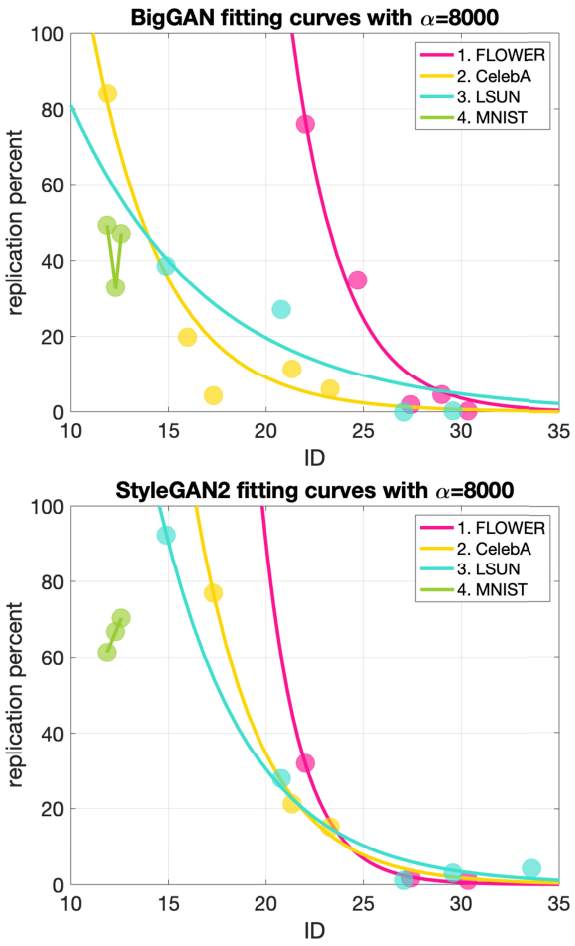

Figure S3: GAN replication curves for BigGAN and StyleGAN2 trained on MNIST comparing to other datasets.

Choose the appropriate level of quality that best suits the image:

1. ***[Excellent]*** *the image looks no different than a real image.*
2. ***[Good]*** *the image looks close to a real image but there are some issues barely noticeable.*
3. ***[Fair]*** *the image has some small issues but overall looks close to a real image.*
4. ***[Poor]*** *the image has obvious issues but I can see what is the image about.*
5. ***[Terrible]*** *the image looks terrible and I cannot see what it is about.*

The Workers can use mouse and keyboard to select the level most suitable for the image. The average completion time for each task (5 images) is 1 minute 39 seconds.

To process the data, we first use integer coding to convert the quality level to numbers, with *Excellent* at 5 and *Terri-*

ble at 1. To report the result, we aggregate the answers from the Workers by calculating the mean and its 95% confidence interval for each subset level for all the GAN-dataset combinations, which is shown in Figure 5 of the main paper.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019. 3

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[3] 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. *arXiv preprint arXiv:1706.08500*, 2017. 3

[4] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 3

Figure S4: Qualitative results of replication experiments for BigGAN-CelebA, BigGAN-Flower and BigGAN-LSUN (bedroom) combination. All images are randomly generated without cherry-picking. For the BigGAN and a given dataset, at each subset level, a BigGAN model is trained and examined for its replication. This results show that when the dataset size is small, BigGANs can generate almost exact replication of training data. The replication is gradually alleviated when the dataset size increases.

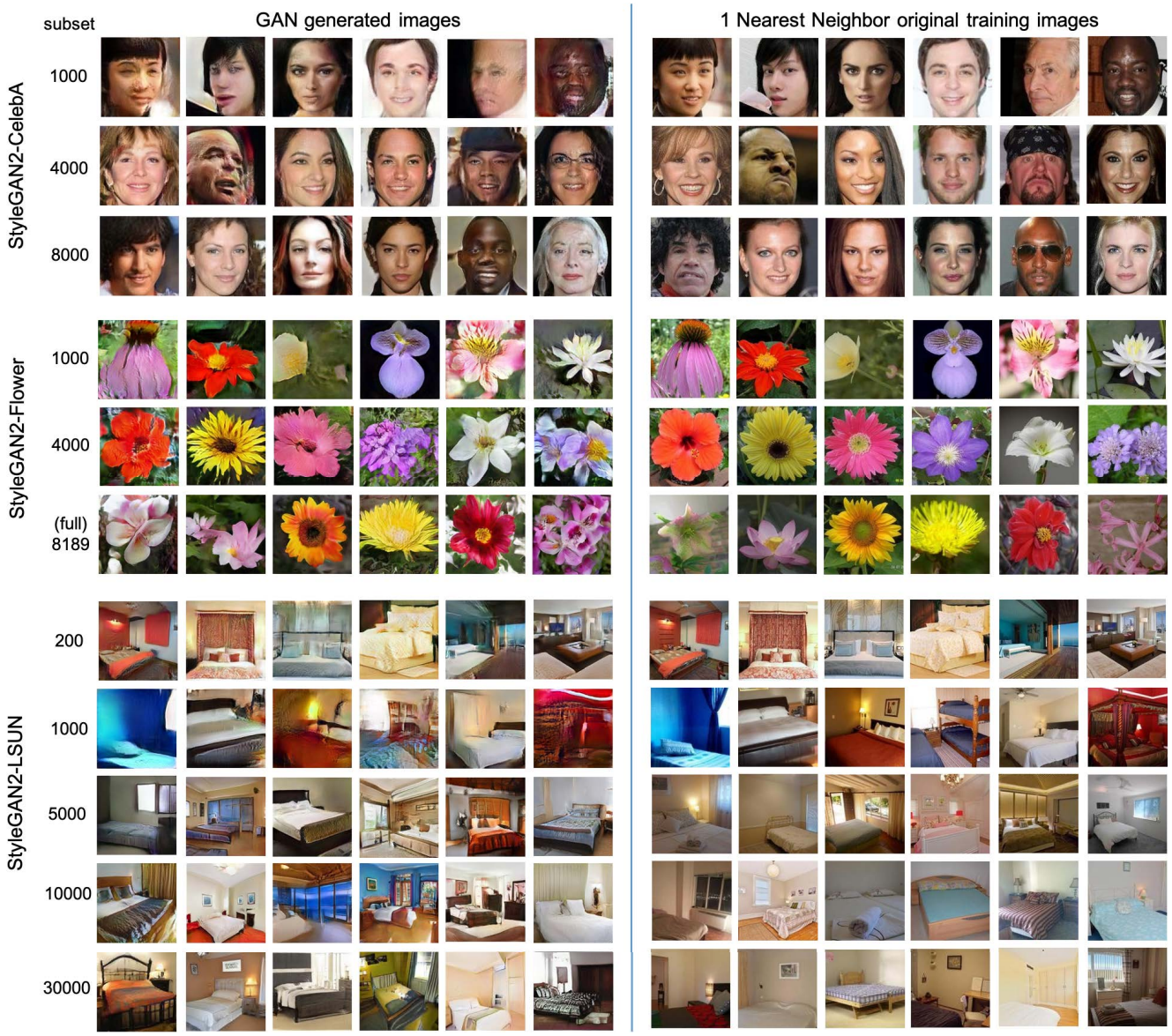

Figure S5: Qualitative results of replication experiments for StyleGAN2-CelebA, StyleGAN2-Flower and StyleGAN2-LSUN (bedroom) combination. All images are randomly generated without cherry-picking. For the StyleGAN2 and a given dataset, at each subset level, a StyleGAN2 model is trained and examined for its replication. This results show that when the dataset size is small, StyleGAN2 can generate almost exact replication of training data. The replication is gradually alleviated when the dataset size increases.

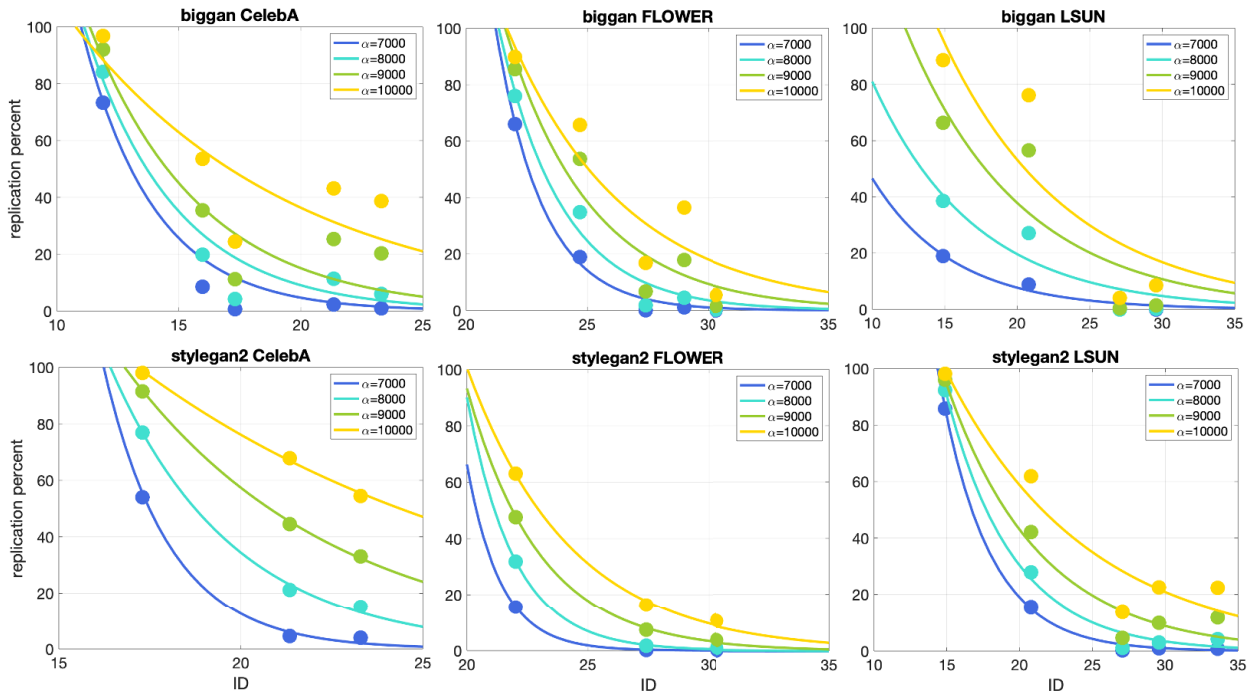

Figure S6: Scatter plots and curve fitting for dataset ID vs GAN replication percentage at each subset level for BigGAN and StyleGAN2 trained on CelebA, Flower and LSUN-bedroom, with different thresholds α. For a given α, regardless of GAN architectures, datasets or α values, the results show a common exponential decay trend. For each plot with same dataset and GAN architecture but increasing threshold values α, the fitted curves gets farther away from the origin, indicating a decreasing scaling b value.

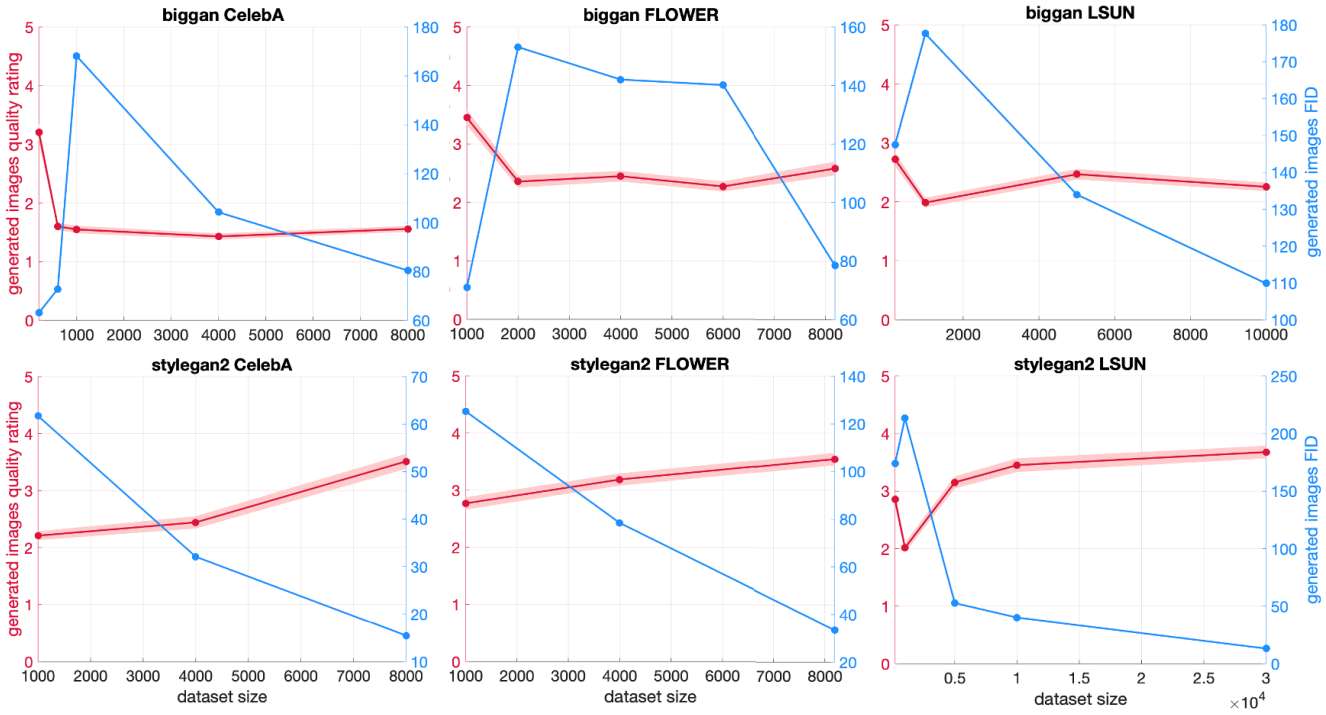

Figure S7: Comparison between perceptual image quality and FID. Each figure shows the curves of perceptual image quality (red) and FID (blue) with respect to the subset levels for each GAN-dataset combination. Although in most experiments, the FID correlates (anticorrelates) with the perceptual image quality acquired from AMT behavioral experiments, the two shows different trends in BigGAN-CelebA experiment, where the image quality improves significantly from subset level 1000 to 8000 according to FID, but only slightly for perceptual quality rating. Note that for FID, the smaller the better and for perceptual image quality, the higher the better.

Figure S8: Comparison on image replications in RGB and InceptionV3 space. FLOWER dataset with subset 1000 images are used. RGB space gives nearly perfect matching, while InceptionV3 space can only capture parts of the image feature and thus leads to matching results inconsistent with human perceptions. When graded by a single human rater, 59.86% of the InceptionV3 replications for BigGAN and 94.01% replications for StyleGAN2 are unsatisfactory.

Figure S9: Comparison on image replications in RGB and SimCLR space. FLOWER dataset with subset 1000 images are used. RGB space gives nearly perfect matching, while SimCLR space can only capture parts of the image feature and thus leads to matching results inconsistent with human perceptions. When graded by a single human rater, 10.51% of the SimCLR replications for BigGAN and 39.43% for StyleGAN2 are unsatisfactory.