A. Dataset Creation

As discussed in Sec. 2, existing datasets designed for the binary discrimination of real vs. synthetic samples are not suitable for the task of model attribution (*i.e*. discriminating among multiple different generative models) in two aspects: (i) the diversity of generative models (GMs) is limited, and (ii) the variability of the models' training datasets makes the study of model fingerprint – independent of their training datasets – difficult. Rather, a proper benchmark dataset for model attribution should satisfy the following desiderata:

1. It should include GMs from various families, covering VAEs, GANs, Flows and Score-based (*i.e*. Diffusion) models
2. It should contain state-of-the-art models, in addition to the more standard models that existing works have focused on (*e.g*. ProGAN, CycleGAN, StyleGAN)
3. The generative models in the dataset should be trained on the same training set in order for the analysis on the fingerprint features to be directly attributed to the characteristics of the generative models, without confounding effects from the variability in the models' training datasets.

To this end, we designed three new datasets (GM-CIFAR10, GM-CelebA, GM-CHQ and GM-FFHQ; See Tbl. 1) that carefully satisfy these three desiderata. GM-CIFAR10 contains images from generative models trained on CIFAR-10 [32]. GM-CelebA contains images from generative models trained on CelebA [34], GM-CHQ from models trained on CelebA-HQ (256) [24], and GM-FFHQ from models trained on FFHQ (256) [24]. To complement the existing datasets, our datasets include GMs that achieve state-of-the-art results on unconditional image synthesis, such as DDGAN [63] and StyleSwin [66] for GAN, NVAE [58] and Efficient-VDVAE [19] for VAE, and LDM [50] and LSGM [59] for diffusion models.

A.1. Details on dataset creation

GM-CelebA dataset We construct a dataset of real and GM-generated images by collecting real images from the original CelebA (image-aligned and resized to 64x64) [34] and generating 100k samples from GMs trained on CelebA-64. We consider 4 VAE models, 5 GAN models, 1 Flow and 1 Score-based model, based on the best availability of the released code and model checkpoints. We trained the models on our own when no official pretrained model was released. See Tab. 1 for the list of GMs used for this dataset.

GM-CHQ dataset To study the fingerprints of more advanced generative models, we collect samples from state-of-the-art models such as NVAE [58], Efficient VDVAE [19], VQ-GAN [10], StyleGAN2 [27], Denoising Diffusion GAN (DDGAN) [63], DDPM++ [22], NCSN++ [55] and Latent Score-based Generative Model (LSGM) [59]. All the models are trained on CelebA-HQ 256 [24]. From each model, we collect 100k samples. See Tab. 1 for the full list of GMs and the supplementary materials for details on the sampling procedure from each generative model. d

Fig. 7, Fig. 8 and Fig. 9 show samples from our GM-CHQ dataset. The images are randomly sampled from each GMs following the process detailed in each work or codebase.

B. Details on baseline fingerprinting methods

Tab. 6 summarizes baseline fingerprinting methods that we compared against our definitions proposed in Sec. 4.

C. Feature space analysis

C.1. Fréchet Distance Ratio (FDR)

We measure the separability of a fingerprint feature space using the ratio of Fréchet Distance. This measure was also used in Yu et al. [64] to evaluate the learned feature space for GAN fingerprints. In our work, we use it to evaluate fingerprints in a more generalized sense in that they are to identify more diverse set of GMs (not just GANs) including many state-of-the-art models.

FDR is computed as the ratio of inter-class and intra-class Fréchet Distance [7]:

$$FDR = \frac{\text{inter-class FD}}{\text{intra-class FD}} \tag{19}$$

Intra-class FD aims to capture the average tightness of a feature distribution per class, and can be measured as the FD between two disjoint sets of images in the same class. As in Yu et al. [64], we split, for each class, the fingerprint features into two disjoint sets of equal size, compute their Fréchet Distance, and then average it over each class.

Inter-class FD aims to capture the average distance between feature distributions of different classes. To compute this distance, we measure the FD between two feature sets from different classes and take the average over every possible pair of (different) classes.

D. Experiment: characterization of generative models

We further study the clustering structure of the set of GM artifacts and explore if it is possible to relate the clustering patterns to the hyperparameters that govern the model design of generative models, such as the type of sampling layers and the type of loss functions. To do so, we take insights from the experiments in [1, 8, 9], and group the model hyperparameters into the following categories: Type of upsampling, Type of non-linearity in the last layer, Type of normalization, Use of downsampling, Use of skip connection, and Type of loss function. For example, we categorize the loss functions

Paper	Input domain	Representation	Classifiers	Metric(best)	Datasets
McCloskey18 [40]	RGB	Histogram of saturated, under-exposed pixels	SVM	AUC (0.7)	NIST MFC2018
Nataraj19 [43]	RGB	Co-occurrence matrix of pixels	CNN	EER (12.3%)	100k-Faces (StyleGAN)
Durall20 [8]	Freq.	1D power spectrum (azimuthal integral)	SVM	Binary Acc (96%)	Own (DCGAN, DRAGAN, SGAN, WGAN-gp)
Dzanic20 [9]	Freq.	Fourier spectrum (norm. by DC gain)	KNN	Binary Acc (99.2%)	Own (StyleGAN,StyleGAN2, PGGAN,VQ-VAE2,ALAE)
Wang20 [61]	Freq.	2D average spectra	CNN	LOMO, Binary Acc (84.7%)	Own (10 GANs)
Marra18 [37]	Learned	Supervised	Pretrained CNN + Finetuned (Inception-v3/XceptionNet)	LOMO[1], Binary Acc (94.49%)	Own (Real, CycleGAN per category)
Marra19 [39]	Learned	Supervised	CNN + IL	Binary Acc (99.3%)	Own (4 GANs, 1 Flow)
Yu19 [64]	Learned	Supervised	CNN	Multi Acc (98.58%)	Own (ProGAN, SNGAN CramerGAN, MMDGAN)

Table 6. **Features and datasets used in the baseline methods**

	Model Params. (NMI ↑)					Optim. Params
Methods	***Upsampling***	Non-linearity	Normalization	Downsampling	Use skip	***Loss Type***
$ManiFPT_{\text{RGB}}$	0.625	0.453	0.647	0.432	0.541	0.563
$ManiFPT_{\text{FREQ}}$	0.654	0.354	0.534	0.692	0.317	0.631
$ManiFPT_{\text{SL}}$	0.613	0.452	0.481	0.546	0.434	0.677
$ManiFPT_{\text{SSL}}$	0.680	0.477	0.465	0.615	0.357	0.573
Average	**0.643**	0.434	0.465	0.532	0.571	<u>0.611</u>

Table 7. **Clustering structure in GM-CHQ.** We measure the alignment of clustering in Normalized Mutual Information (NMI) on our feature spaces (using RGB, Frequency, Supervised-learning (SL), Self-supervised learning (SSL) representations to clusterings based on model design parameters (*e.g.* type of upsampling, type of non-linearity in the last layer, type of normalization later, type of loss function). NMI is bounded to [0,1]. Higher index indicates closer agreement between two cluster assignments.

in our datasets into three types (likelihood-based (VAEs), implicit density matching (GANs), and score-matching (Score-based models)), and the type of non-linearity in the last layer into ReLU, Tanh, and Sigmoid. See the supplementary for more details on our categorization of the hyperparameters and specific values each GM in our datasets take.

Metric. We use Normalized Mutual Information (NMI) [41] to measure the clustering alignment between the clustering in a fingerprint representation ($\mathcal{C}_f$) and the clustering on the assignment of a model design choice (*e.g.* type of upsampling operation) as the ground-truth cluster labels ($\mathcal{C}_h$). For instance, to measure how well the clustering in a fingerprint space coincides with the clusters according to the type of loss function, we set as $\mathcal{C}_h$ the result of clustering datapoints based on the type of their source GM's loss function. If the loss type is a proper criterion to categorize different generative models, the two clusterings ($\mathcal{C}_f$ based on the fingerprint representations and $\mathcal{C}_h$ on the labels of loss type) will have a strong agreement, and their clustering index will be high.

Results. Table 5 reports NMI between a feature space and each category of model hyperparameters, reflecting which criterion in grouping the generative models (*e.g.* the type of upsampling vs. the type of non-linearity vs. the type of loss function) agrees well with the grouping in the fingerprint representation space. Note that NMI is bounded to $[0, 1]$, and a higher index indicates a closer agreement between two cluster assignments. The last row (Avg_{ours}) shows the NMI averaged over our methods in RGB, frequency, supervised-learning based and unsupervised-learning based representation space. First of all, we observe that the clustering on our fingerprint space aligns the best with the clustering by the GMs' upsampling types and loss types. In other words, our result suggests that the two hyperparameters (Type of upsampling and Type of loss function) show the most similar clustering patterns with our fingerprint representations.

The high NMI for the type of upsampling supports previous experiments that identified the upsampling operation of generator networks as a cause of the high-frequency discrep-

ancies in the GM-generated images [4, 8, 9]. Additionally, the high NMI for the type of loss function confirms the general consensus in the research community that the training objective of a generative model is one of the key factors that affect their characteristics.

Therefore, our findings confirms the general intuition in the research community about distinct sources of limitations in generative models and shows the utility of our definitions.

E. Visualization: artifacts of generative models

We visualize more examples of artifacts of generative models in GM-CelebA and GM-CHQ, computed under our proposed definition in Sec. 3.1. Fig. 5 shows the triplets of (generated images (x_G), its closest point to the data manifold in RGB ($x^\star$) and the artifact (a)). Fig. 6 visualize the artifacts in frequency domain from the GM-CHQ dataset.

E.1. Artifacts in RGB space (GM-CelebA)

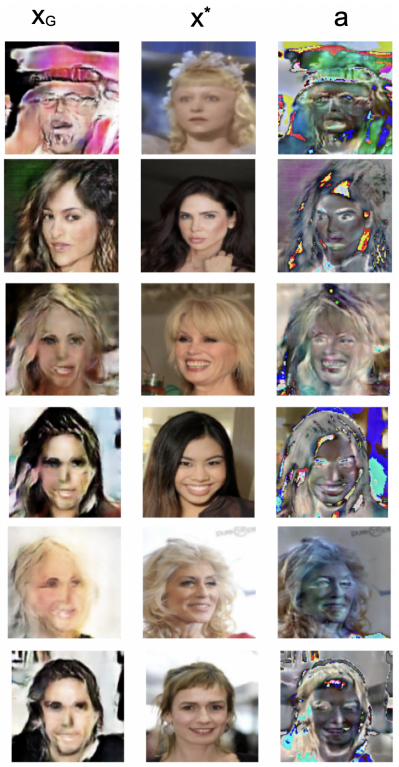


Figure 5. **Visualization of artifacts in the RGB space (GM-CelebA).** Each column corresponds to the generated images (x_G), their closest points on the data manifold ($x^\star$), and the artifacts (a). Each artifact is computed as the different between $x^\star$ and x_G following the definition and algorithm in Sec. 3.1.

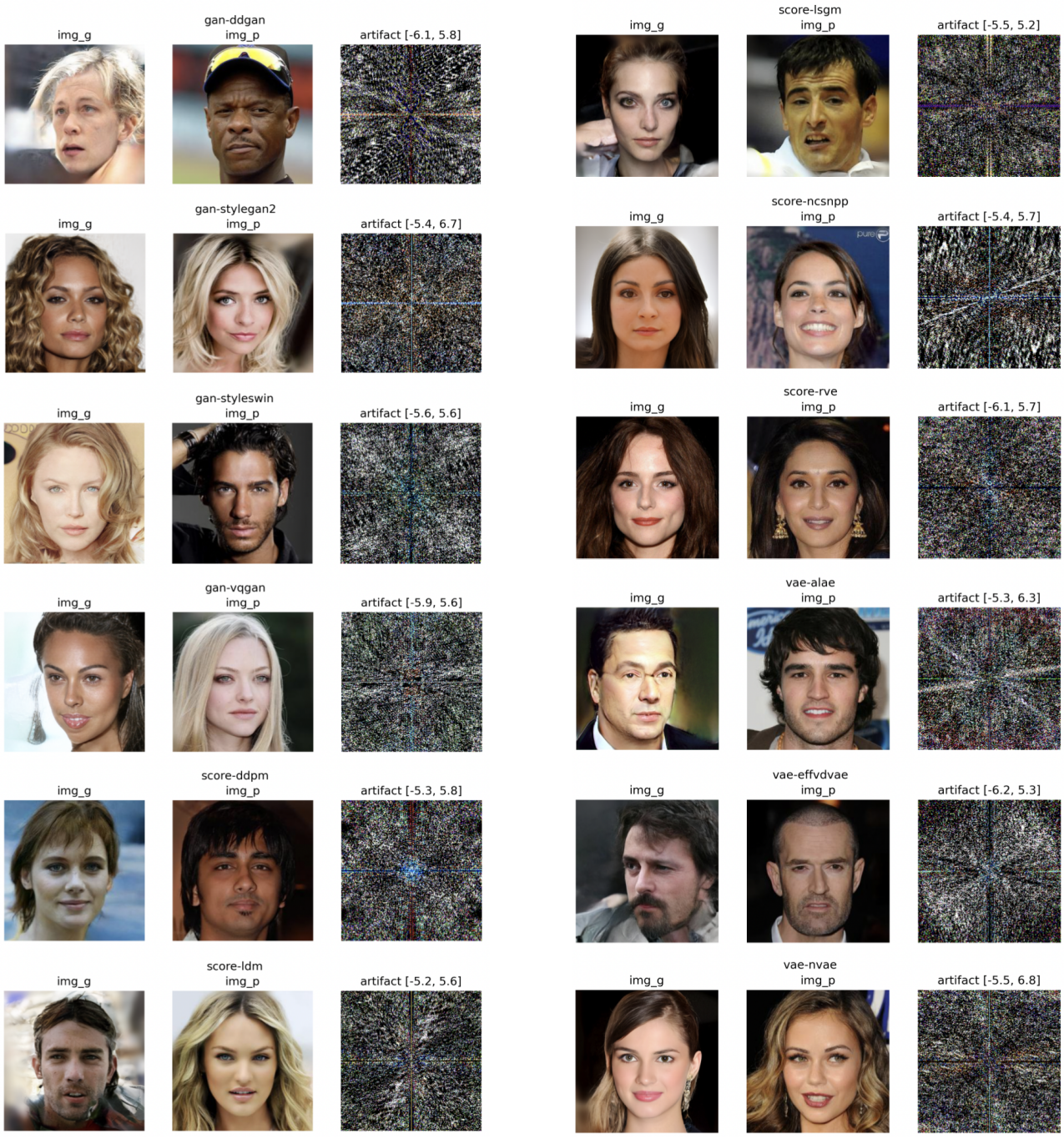


Figure 6. **Visualization of artifacts in the frequency space (GM-CHQ)**. We show some examples of triplets (model-generated image (img_g), closest point on the data manifold (img_p), artifact) from GM-CHQ dataset by computing artifacts (as defined in Sec. 3.1) in frequency domain. img_p is the point on the real data manifold that is closest to img_g in the frequency domain. Artifact is computed as the different between the two points, img_g and img_p, after applying channelwise-FFT.

(a) DDGAN [63]

(b) StyleGAN2 [27]

(c) StyleSwin [66]

(d) VQ-GAN [10]

Figure 7. **Samples from GAN models in GM-CHQ.**

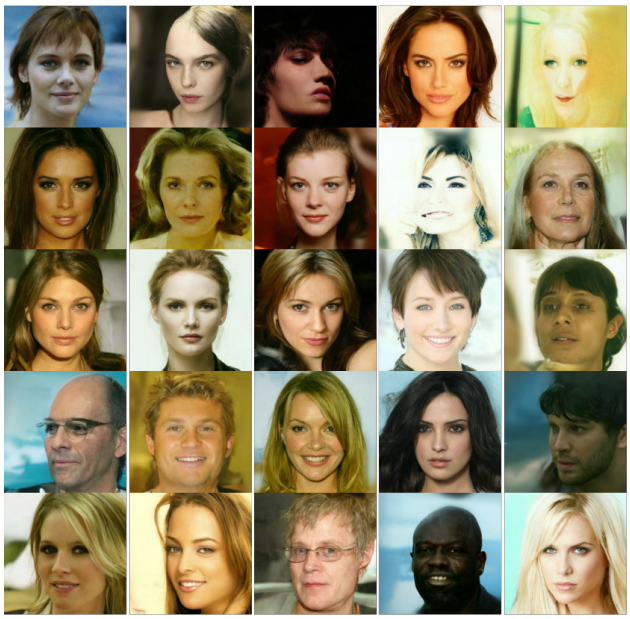
(a) StyleALAE [47]

(b) Efficient VDVAE [19]

(c) NVAE [58]

(d) VAEBM [62]

Figure 8. **Samples from VAE models in GM-CHQ.**



(a) DDPM [22]

(b) LDM [50]

(c) LSGM [59]

(d) NCSN++ [55]

Figure 9. **Samples from score-based (a.k.a. diffusion) models in GM-CHQ.**