FLD+: Data-efficient Evaluation Metric for Generative Models

Supplementary Material

1. Details of Distortions

Gaussian Noise: We construct a noise matrix N with values drawn from a $\mathcal{N}(0, 1)$ Gaussian distribution and scaled to the range $[0, 255]$. The noisy image is then computed by combining the original image matrix X with the noise matrix N as $(1 - \alpha) \cdot X + \alpha \cdot N$, where $\alpha \in \{0, 0.001, 0.005, 0.01, etc\}$ determines the amount of noise added to the image. A larger value of α introduces more noise and the noisy image is clipped to ensure that all pixel values remain within the valid range $[0, 255]$.

Gaussian Blur: The image is convolved with a Gaussian kernel with kernel size k. The kernel size controls the intensity of the blur, with larger kernels producing greater smoothing effects. The standard deviation (σ is set to 0, allowing it to be automatically computed based on the kernel size. This approach simulates different levels of image degradation, ranging from minimal to significant blurring, by convolving each image with a Gaussian kernel. In this case, $k \in \{3, 5, 7, 9, 11\}$, and the Gaussian blur is applied uniformly across the image.

Salt and Pepper Noise Addition: To add salt and pepper noise to an image, random values are generated for each pixel. The probability p controls the proportion of pixels that will be altered. For salt noise, pixels with random value less than $\frac{p}{2}$ are set to the maximum intensity value (255). Similarly, for pepper noise, pixels with random value greater than $1-\frac{p}{2}$ are set to the minimum intensity value (0). The amount of noise is directly proportional to the value of p, here $p \in \{0, 0.001, 0.005, 0.01, etc\}$ with larger values of p resulting in a higher number of noisy pixels in the image.

For experiments involving sampling efficiency, we used a CelebA-HQ pre-trained diffusion model for generating images [1]. For experiments on checking the performance of our metric on progressive image generation, we used the original DDPM model [2].

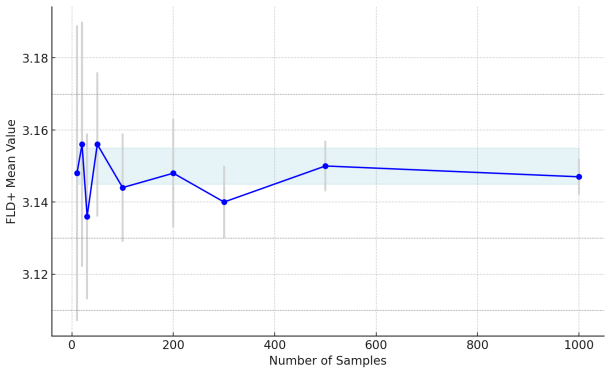


Figure 1. The behavior of FLD+ across different sample sizes clearly demonstrates that FLD+ achieves reliable results with fewer than 500 samples. The variance in FLD+ computed clearly goes down when sample size is more than 500. The mean FLD+ (blue) is reported after 10 runs for each sample size.

References

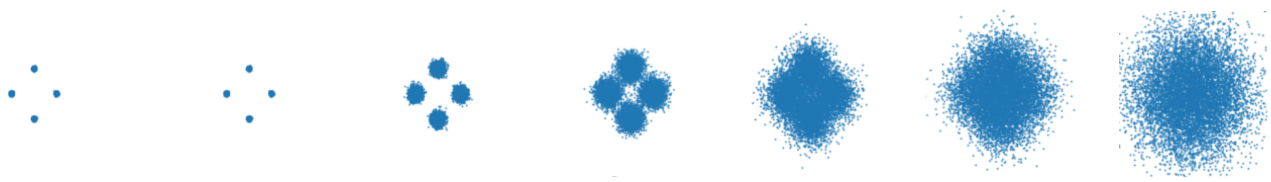
[1] Tianqi Chen. Ddpm: Pytorch implementation of denoising diffusion probabilistic models. https://github.com/tqch/ddpm-torch, 2023. Accessed: 2024-09-30. 1

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2020. Curran Associates Inc. 1

[3] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9307–9315, 2023. 2

[4] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan, 2020. 2

Table 1. FD, unbiased FD, and FLD+ values when the normality assumption is violated show the advantage of using FLD+. The leftmost image represents a reference mixture-of-Gaussian distribution, progressively from left to right the subsequent distributions deviate further from the true distribution keeping mean and standard deviation same. The FD of all the subsequent mixture distributions to the reference mixture distribution, calculated under the normality assumption, remain misleadingly zero [3]. In contrast, FLD+ accurately captures the increasing deviation from the reference distribution, demonstrating its robustness in scenarios where the normality assumption is violated.



FD	0	0	0	0	0	0
Unbiased FD	0	0	0	0	0	0
FLD+	2.72	12.68	21.11	31.19	47.94	80.64

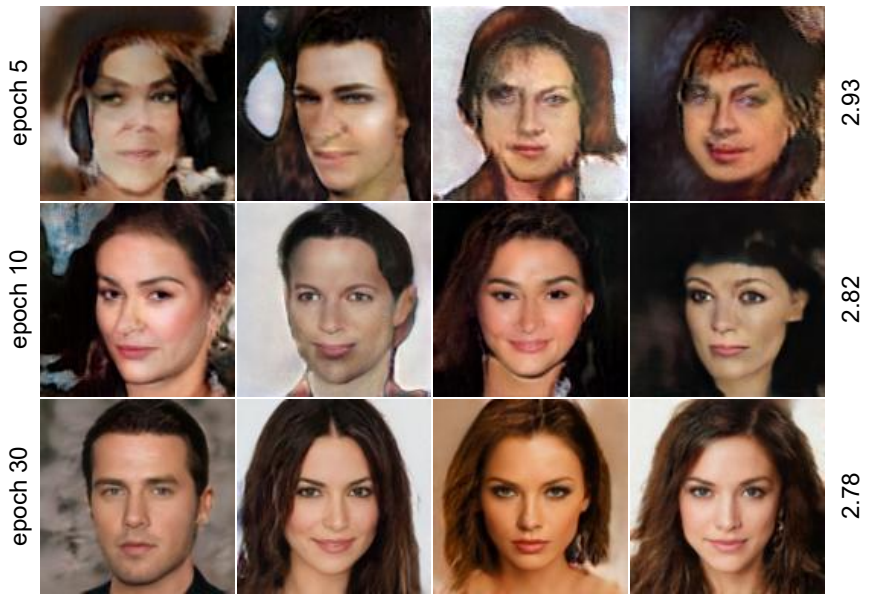


Figure 2. The figure illustrates the monotonicity of FLD+ when evaluating generated images across different training epochs for StyleGAN2 [4]. As the number of training epochs increases, the distortion in the images decreases, and FLD+ accurately reflects this reduction, demonstrating its ability to track changes in image quality across the adversarial training epochs.