Rethinking FID Through the Geometry of the Reference Dataset

Yunghee Lee [1] **Byeonghyun Pak** [1]

Abstract

Fréchet Inception Distance (FID) is widely used to evaluate image generators, yet lower FID does not always correspond to better sample quality. We show that this mismatch depends in part on the geometry of the reference dataset. In a controlled study across six datasets, distributional density and effective rank significantly explain how FID changes as sample quality improves. Concentrated datasets tend to yield more favorable FID trends, whereas more dispersed datasets can make FID worsen despite better samples. Attribution to precision and recall and ablations with alternative feature spaces and distances support the same conclusion. These results suggest that distributional metrics should be interpreted together with the geometry of the reference dataset for more reliable benchmarking.

1. Introduction

Recent progress in image generation (Saharia et al., 2022; Ramesh et al., 2022; Rombach et al., 2022) has been accompanied by the widespread adoption of Fréchet Inception Distance (FID) (Heusel et al., 2017) as a standard metric for evaluating generative models. FID estimates the discrepancy between the distribution of Inception-v3 (Szegedy et al., 2016) features extracted from real images and that of generated images, and lower FID has therefore been widely interpreted as evidence of a better generator. In practice, FID has become not only a reporting convention but also a benchmark target that shapes model development, hyperparameter tuning, and claims of progress across the literature (Lu et al., 2025; Kim et al., 2025). This central role implicitly assumes that improvements in FID correspond to improvements in the qualities that matter for downstream or real-world use.

[1]Work done while at Agency for Defense Development, Daejeon, South Korea. Correspondence to: Byeonghyun Pak <bhpak@umd.edu>.

Accepted to the 1st Workshop on Combining Theory and Benchmarks, CTB@ICML 2026, Seoul, South Korea.

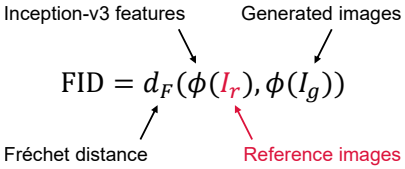

Figure 1. Distributional metrics depend on four components. Prior work has studied the generator, feature extractor, and estimator; we investigate the reference dataset.

However, a growing body of evidence suggests that this assumption is unreliable. Choi et al. (2025) reported that allocating more computation per sample produces visibly sharper images, yet worsens FID on the COCO dataset (Lin et al., 2014). Lee et al. (2025) showed that tuning a generation hyperparameter to minimize FID can yield the worst per-sample quality as measured by ImageReward (Xu et al., 2023). Likewise, Jayasumana et al. (2024) demonstrated cases where stronger image distortions lead to better FID scores. If FID simply measured generator quality, such adjustments should not consistently degrade the metric while improving perceptual quality, or vice versa. These contradictions raise a broader concern: current benchmarking practice may fail to provide reliable or predictive insight into model behavior in realistic settings, even when the reported metric appears rigorous and standardized.

Prior work has mainly sought explanations in the fragility of Inception-v3 (Kynkäänniemi et al., 2022; Jayasumana et al., 2024; Parmar et al., 2022) and in the limitations of the Fréchet distance itself (Chong & Forsyth, 2020; Lee & Lee, 2022). While these analyses are important, they leave one basic aspect of the metric underexplored: **the role of the reference dataset**. Because FID is defined as a distance between the generated distribution and a chosen real-image reference distribution, its behavior necessarily depends on the structure of that reference set. Yet the choice of reference dataset is rarely justified, and is often inherited from prior benchmarks (Dehghani et al., 2021; Otani et al., 2023).

A dataset such as the CelebA-HQ dataset (Karras et al., 2017; Xia et al., 2021) concentrates around a dominant visual mode, whereas the COCO dataset (Lin et al., 2014) spans a far broader and more heterogeneous set of semantic categories; these datasets therefore occupy markedly different regions of feature space. This observation motivates a

fundamental question for theory-informed benchmarking: *which characteristics of the reference dataset determine the properties that FID favors?*

In this paper, we investigate how the geometry of a reference dataset shapes the behavior of FID. We characterize each dataset using two geometric properties: **distributional density** (Loftsgaarden & Quesenberry, 1965) and **effective rank** (Roy & Vetterli, 2007). We then test whether these properties moderate the effect of generated sample quality on FID. Our results show that distributional density is a significant moderator of the quality-FID relationship. For dispersed datasets, FID can *worsen* even when per-sample quality is *improved*, offering a potential explanation for contradictions noted in earlier studies (Stein et al., 2023). We further decompose FID into precision and recall (Kynkäänniemi et al., 2019) and find that, for such datasets, FID is more strongly correlated with recall, which can improve even as sample quality worsens. This dependence on dataset remains significant when replacing Inception-v3 with DINOv2 (Oquab et al., 2023) and when replacing Fréchet distance with MMD (Gretton et al., 2012) as in KID (Bińkowski et al., 2018). Therefore we argue that **distributional metrics should be interpreted together with reference dataset geometry**, and that understanding the reference dataset is essential for reliable benchmarking of generative models.

In summary, our contributions are as follows:

- We present a controlled empirical study showing that reference dataset geometry, quantified by distributional density and effective rank, significantly moderates FID.
- We perform ablations across alternative feature spaces (DINOv2) and alternative kernel-based metrics (KID), showing that the observed effect is not specific to Inception-v3 features or the Fréchet estimator.
- We provide practical guidance for evaluating distributional metrics: use FID with concentrated datasets, or report it together with the geometric descriptors.

2. Method

2.1. Geometric Properties of Dataset

We summarize each reference dataset with two scalar descriptors of its feature distribution. Both descriptors are computed in the same feature space as the metric.

Distributional Density. Distributional density measures how tightly the reference distribution concentrates around its own points. A high value means the samples are clustered and close to each other; a low value means they are scattered and far away.

Table 1. Distributional density $\langle -\log d_k \rangle$ and effective rank $\mathrm{erank}(A)$ of the six reference datasets in the Inception-v3 feature space.

Dataset	$\langle -\log d_k \rangle$	$\mathrm{erank}(A)$
FFHQ (Karras et al., 2019)	-2.48	1243
CelebA-HQ (Karras et al., 2017)	-2.36	1220
MJHQ-30K (Li et al., 2024)	-2.74	1341
ImageNet (Deng et al., 2009)	-2.68	1431
Flickr30K (Young et al., 2014)	-2.80	1341
COCO (Lin et al., 2014)	-2.67	1337

A standard estimator of concentration is the kNN density (Loftsgaarden & Quesenberry, 1965), $\hat{p}(x) \propto d_k(x)^{-D}$, where $d_k(x)$ denotes the Euclidean distance from x to its k-th nearest neighbor and D is the feature dimension. However, in the Inception-v3 feature space with $D = 2048$, the exponent makes $\hat{p}$ impractical as a dataset-level summary, as its values span tens of orders of magnitude across datasets. To obtain a more stable descriptor, we take the logarithm and then average over the reference set, reducing the pointwise estimate to a single scalar. The resulting descriptor, which we call the *mean kNN log-density*, is defined as

$$\langle -\log d_k \rangle = \frac{1}{n} \sum_{i=1}^{n} -\log d_k(x_i), \tag{1}$$

where n is the number of reference points. We use $k = 80$ throughout this paper.

Effective Rank. Effective rank measures how many linear directions the feature distribution spreads over. A high value means the support occupies many principal components of the feature space; a low value means it collapses onto a few. The effective rank of a matrix A is (Roy & Vetterli, 2007)

$$\mathrm{erank}(A) = \exp\left(H(\boldsymbol{\sigma}/\|\boldsymbol{\sigma}\|_1)\right), \tag{2}$$

where $\boldsymbol{\sigma}$ is the vector of singular values of A, $\|\boldsymbol{\sigma}\|_1$ is the L1 norm of $\boldsymbol{\sigma}$, and $H(\boldsymbol{p}) = -\sum_k p_k \log p_k$ is the Shannon entropy (Shannon, 1948). $\mathrm{erank}(A)$ equals $\mathrm{rank}(A)$ when nonzero singular values are all equal, which gives it the interpretation of weighted dimensionality. We use $\mathrm{erank}(A)$ as the descriptor where A is the centered feature matrix.

Table 1 reports both descriptors on the six reference datasets used in this study. The face datasets, FFHQ and CelebA-HQ, score highest on distributional density and lowest on effective rank, consistent with their single-domain structure. The open-domain datasets, ImageNet, Flickr30K, and COCO, score in the opposite direction.

2.2. Statistical Analysis

We investigate how FID behavior depends on the reference dataset through statistical analysis. In particular, we focus

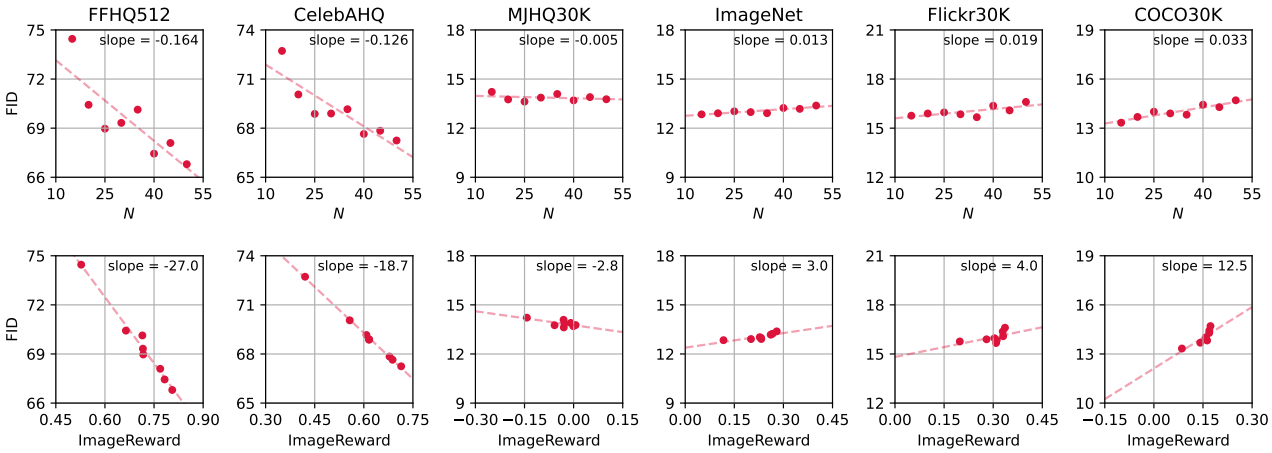

Figure 2. FID across six reference datasets under a fixed generator and a sweep over the number of denoising steps N. Top row: FID as a function of N. Bottom row: FID as a function of ImageReward over the same sweep. Although image quality generally improves with more denoising steps, FID exhibits dataset-dependent trends and may either increase or decrease depending on the reference dataset.

on how FID fragility varies across reference datasets. We ask: *for a given dataset, how does FID change as per-sample quality improves?* To answer this, we introduce an independent variable X for sample quality and a dependent variable Y for FID, and use the slope of a linear regression model to quantify their relationship.

Omnibus Test. We first test whether the slopes vary significantly across datasets. The null hypothesis H_0 is that *the slopes do not vary by dataset*, whereas the alternative hypothesis H_1 is that *the slopes vary significantly across datasets*. To test H_0, we fit a two-level hierarchical linear model (Raudenbush & Bryk, 2002). Using a likelihood ratio test (Stram & Lee, 1994), we compute a test statistic D, which follows a mixed χ^2 distribution under H_0. We report D and its corresponding p-value for testing H_0. Additional details are provided in Appendix B.

Moderation Test. We next test whether the geometric descriptors moderate the slope. To do so, we introduce a covariate Z, representing a geometric descriptor, into the hierarchical linear model and test for a cross-level interaction such that the slope varies with Z. The null hypothesis H_0 is that *the slope does not depend on* Z, whereas the alternative hypothesis H_1 is that *the slope depends on* Z. We report the cross-level interaction coefficient γ_{11}, its corresponding p-value from the Wald test (Wald, 1943), and R^2_{slope}, the proportion of slope variance explained by Z. Additional details are provided in Appendix B.

2.3. Experimental Design

We analyze six reference datasets spanning a spectrum from concentrated to dispersed feature distributions: FFHQ and CelebA-HQ on the concentrated side, MJHQ-30K, ImageNet, Flickr30K, and COCO on the dispersed side. For each dataset, we generate samples with Stable Diffusion 1.5 (Rombach et al., 2022) using DDIM (Song et al., 2020) at 512×512, classifier-free guidance scale 7.5, and a fixed random seed per prompt, while sweeping the number of denoising steps $N \in \{15, 20, 25, 30, 35, 40, 45, 50\}$. Prompt sources per dataset are listed in Appendix A. For each pair of dataset and N we generate one image per reference example, so the generated set matches the reference set in both size and conditioning. We evaluate each set using FID, with KID and $\text{FD}_{\text{DINOv2}}$ (Stein et al., 2023) as ablations, ImageReward (Xu et al., 2023) as a proxy for perceptual quality, and precision and recall (Kynkäänniemi et al., 2019) as diagnostic measures.

3. Results

Figure 2 compares FID with per-sample quality across reference datasets. The relationship between FID and sample quality exhibits distinct trends across datasets. On CelebA-HQ and FFHQ, the slope is negative: FID decreases as the number of denoising steps increases and ImageReward improves. By contrast, on COCO, Flickr30K, and ImageNet, the slope is positive. MJHQ-30K lies between these two regimes. These results show that although per-sample quality generally improves with more inference steps, FID may either increase or decrease depending on the reference dataset. This suggests that FID does not reflect a fixed notion of generator quality, but rather favors different properties depending on the geometry of the reference dataset.

3.1. Statistical Analysis

Omnibus Test. We first test whether the slope of the relationship between sample quality and FID varies across

Table 2. Statistical test results. IR stands for ImageReward. For each choice of X, the omnibus test result (D and p) is reported first, followed by the moderation test results (γ_{11},p, and R^2_{slope}).

X	Y	Z	D or γ_{11}	p	R^2_{slope}
N	FID	-	44.3	$< .001$	-
N	FID	$\langle -\log d_k \rangle$	-0.0323	$< .001$	0.707
N	FID	$\text{erank}(A)$	0.0314	.002	0.661
IR	FID	-	90.9	$< .001$	-
IR	FID	$\langle -\log d_k \rangle$	-0.120	.007	0.548
IR	FID	$\text{erank}(A)$	0.119	.010	0.530

Table 3. Coefficient of determination (R^2) of FID with precision and recall for each dataset. The larger value is highlighted in **bold**.

Dataset	R^2(Precision, FID)	R^2(Recall, FID)
FFHQ	**0.989**	0.672
CelebA-HQ	**0.951**	0.001
MJHQ-30K	**0.734**	0.025
ImageNet	0.690	**0.949**
Flickr30K	0.314	**0.850**
COCO	0.676	**0.833**

datasets. Using $X = N$ and $Y =$ FID, the omnibus test rejects the null hypothesis that all datasets share a common slope ($D = 44.3$, $p < .001$). We obtain the same conclusion when using $X =$ ImageReward and $Y =$ FID ($D = 90.9$, $p < .001$). These results indicate that the slope differences in Figure 2 reflect systematic differences in how FID responds to sample quality across datasets.

Moderation Test on Distributional Density. We next test whether distributional density explains why the slope differs across datasets; the results are reported in Table 2. Using $X = N$, $Y =$ FID, and $Z = \langle -\log d_k \rangle$, we find a significant negative cross-level interaction with $R^2_{\text{slope}} = 0.707$. This indicates that distributional density explains 70.7% of the between-dataset variation in slopes. Moreover, the negative sign of γ_{11} indicates that denser reference datasets tend to have smaller, or more negative, N-to-FID slopes. These results suggest that FID is more likely to improve with additional inference steps when the reference distribution is dense. The same pattern holds with $X =$ ImageReward, indicating that distributional density moderates the relationship between sample quality and FID.

Moderation Test on Effective Rank. We then test whether effective rank similarly explains the slope differences across datasets, with results also reported in Table 2. Using the same model with Z defined as effective rank, we find a significant positive cross-level interaction ($\gamma_{11} = 0.0314$, $p = .002$) with $R^2_{\text{slope}} = 0.661$. This indicates that effective rank explains 66.1% of the between-dataset variation in slopes. The positive sign of γ_{11} indicates that datasets with higher effective rank tend to have larger N-to-FID slopes. The same pattern holds with $X =$ ImageReward. In turn, this suggests that reference datasets spanning many directions in feature space are more likely to make FID worsen as sample quality improves.

Table 4. Ablation study results. For each alternative metric, the first row reports the omnibus test results (D and p), followed by moderation test results (γ_{11}, p, and R^2_{slope}).

X	Y	Z	D or γ_{11}	p	R^2_{slope}
N	KID	-	46.4	$< .001$	-
N	KID	$\langle -\log d_k \rangle$	-0.0343	$< .001$	0.763
N	KID	$\text{erank}(A)$	0.0315	.005	0.596
N	$\text{FD}_{\text{DINOv2}}$	-	26.6	$< .001$	-
N	$\text{FD}_{\text{DINOv2}}$	$\langle -\log d_k \rangle$	-0.0108	$< .001$	0.827
N	$\text{FD}_{\text{DINOv2}}$	$\text{erank}(A)$	0.0110	$< .001$	0.837

3.2. Attribution to Precision and Recall

We further decompose FID into precision and recall to examine which component more strongly explains the observed variation in FID. Precision measures the fidelity of generated images to the real data distribution, whereas recall measures their diversity, that is, how well the generated samples cover that distribution. To assess which component is more strongly associated with FID, we compute the R^2 of separate ordinary least squares fits for each dataset. The results are summarized in Table 3. For concentrated datasets such as FFHQ and CelebA-HQ, FID is more strongly correlated with precision. By contrast, for dispersed datasets such as ImageNet, Flickr30K, and COCO, FID is more strongly correlated with recall. This suggests that when the reference dataset is dispersed, FID tends to reward coverage of the real distribution more than sample-level fidelity. In other words, FID can *worsen* even when precision *improves*, if recall *declines* as per-sample quality increases. Viewed from this diagnostic perspective, the result supports our main finding that the behavior of FID depends systematically on reference dataset geometry.

3.3. Ablation Study

We perform an ablation study on other components of FID. Specifically, we replace the Fréchet distance with kernel MMD to obtain KID, and replace Inception-v3 with DINOv2 to obtain $\text{FD}_{\text{DINOv2}}$. Table 4 summarizes the results of the omnibus and moderation tests using these alternative distributional metrics. In all cases, the omnibus test rejects H_0, indicating that the slope remains dataset-dependent regardless of the choice of distance or feature space. This suggests that prior explanations focusing only on the distance measure or feature extractor do not fully explain the

fragility of FID. Moreover, in the moderation tests, the geometric descriptors explain even more slope variance with DINOv2 features than with Inception-v3 features, indicating that the observed effect is not driven by known limitations of Inception-v3. These results show that our geometric descriptors generalize beyond FID and remain informative across alternative distributional metrics.

4. Conclusion

In this paper, we showed that the behavior of FID systematically depends on the geometry of the reference dataset. Across 6 datasets, distributional density and effective rank significantly explained how FID changes as sample quality improves. Concentrated datasets made FID more aligned with precision, whereas dispersed datasets made it more aligned with recall. These findings remained robust across alternative feature spaces and distance measures. Therefore, distributional distance metrics should be interpreted together with the geometry of the reference dataset. When using FID as an evaluation criterion for generative models, we suggest using concentrated reference datasets such as FFHQ. Conversely, for open-domain settings, we recommend reporting FID together with the geometric descriptors of the reference dataset.

Acknowledgements

We sincerely thank Byeongju Woo and Hoseong Kim for their constructive discussions and support. We also appreciate Sol Park and Minkyu Song for providing insightful feedback. This work was supported by the Agency For Defense Development Grant Funded by the Korean Government (912A45701).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

Choi, J., Kang, J., and Han, B. Enhanced diffusion sampling via extrapolation with multiple ode solutions. *arXiv preprint arXiv:2504.01855*, 2025.

Chong, M. J. and Forsyth, D. Effectively unbiased fid and inception score and where to find them. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6070–6079, 2020.

Dehghani, M., Tay, Y., Gritsenko, A. A., Zhao, Z., Houlsby, N., Diaz, F., Metzler, D., and Vinyals, O. The benchmark lottery. *arXiv preprint arXiv:2107.07002*, 2021.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., and Kumar, S. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9307–9315, 2024.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Kim, D., Hwang, J., Oh, C., and Park, J. Mixdit: Accelerating image diffusion transformer inference with mixed-precision mx quantization. *IEEE Computer Architecture Letters*, 2025.

Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., and Aila, T. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019.

Kynkäänniemi, T., Karras, T., Aittala, M., Aila, T., and Lehtinen, J. The role of imagenet classes in fréchet inception distance. *arXiv preprint arXiv:2203.06026*, 2022.

Lee, J. and Lee, J.-S. Trend: Truncated generalized normal density estimation of inception embeddings for gan evaluation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2022.

Lee, Y., Pak, B., Hong, J., and Kim, H. Tortoise and hare guidance: Accelerating diffusion model inference with multirate integration. *arXiv preprint arXiv:2511.04117*, 2025.

Lhoest, Q., Del Moral, A. V., Jernite, Y., Thakur, A., Von Platen, P., Patil, S., Chaumond, J., Drame, M., Plu, J., Tunstall, L., et al. Datasets: A community library for natural language processing. In *Proceedings of the 2021 conference on empirical methods in natural language processing: system demonstrations*, pp. 175–184, 2021.

Li, D., Kamko, A., Akhgari, E., Sabet, A., Xu, L., and Doshi, S. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Loftsgaarden, D. O. and Quesenberry, C. P. A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, 36(3):1049–1051, 1965.

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., and Zhu, J. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, 22 (4):730–751, 2025.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Otani, M., Togashi, R., Sawai, Y., Ishigami, R., Nakashima, Y., Rahtu, E., Heikkilä, J., and Satoh, S. Toward verifiable and reproducible human evaluation for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14277–14286, 2023.

Parmar, G., Zhang, R., and Zhu, J.-Y. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11410–11420, 2022.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Raudenbush, S. W. and Bryk, A. S. *Hierarchical linear models: Applications and data analysis methods*, volume 1. sage, 2002.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Roy, O. and Vetterli, M. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pp. 606–610. IEEE, 2007.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494, 2022.

Shannon, C. E. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Stein, G., Cresswell, J., Hosseinzadeh, R., Sui, Y., Ross, B., Villecroze, V., Liu, Z., Caterini, A. L., Taylor, E., and Loaiza-Ganem, G. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in Neural Information Processing Systems*, 36:3732–3784, 2023.

Stram, D. O. and Lee, J. W. Variance components testing in the longitudinal mixed effects model. *Biometrics*, pp. 1171–1177, 1994.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Wald, A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society*, 54 (3):426–482, 1943.

Xia, W., Yang, Y., Xue, J.-H., and Wu, B. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2256–2265, 2021.

Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., and Dong, Y. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 15903–15935, 2023.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2:67–78, 2014.

Table 5. Details of datasets used for experiments.

Dataset	Image count	Caption source	HuggingFace identifier
FFHQ	70,000	Auto-generated	`Ryan-sjtu/ffhq512-caption`
CelebA-HQ	29,987	Auto-generated	`oftverse/control-celeba-hq`
MJHQ-30K	30,000	Midjourney prompts	`xingjianleng/mjhq30k`
ImageNet	50,000	Class names	`BenSchneider/imagenet-val`
Flickr30K	31,014	Human-annotated	`nlphuji/flickr30k`
COCO	30,000	Human-annotated	`sayakpaul/coco-30-val-2014`

A. Additional Details on Datasets

Table 5 shows additional details of the datasets used for experiments. All datasets were downloaded using the HuggingFace `datasets` library (Lhoest et al., 2021).

B. Additional Details on Statistical Tests

In this section, we cover the details of statistical tests performed.

Omnibus Test. We use a 2-level hierarchical linear model (Raudenbush & Bryk, 2002) to find out if there is a significant difference between slopes across datasets. Let X_{ij} be the X value where i is the index for individual observation for each N and j is the index for datasets. Similarly, Y_{ij} is the Y value for observation i in dataset j. Corresponding to the null hypothesis H_0 such that there is no difference of slopes, we fit a random intercepts only model

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + \epsilon_{ij}, \qquad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \tag{3}$$
$$\beta_{0j} = \gamma_{00} + u_{0j}, \qquad u_{0j} \sim \mathcal{N}(0, \tau_{00}), \tag{4}$$
$$\beta_{1j} = \gamma_{10} \tag{5}$$

for the parameters γ_{00}, γ_{10}, and τ_{00}. For the alternative hypothesis H_1, we fit a random intercepts and slopes model

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + \epsilon_{ij}, \qquad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \tag{6}$$
$$\beta_{0j} = \gamma_{00} + u_{0j}, \tag{7}$$
$$\beta_{1j} = \gamma_{10} + u_{1j}, \qquad \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{01} & \tau_{11} \end{bmatrix} \right) \tag{8}$$

for the parameters γ_{00}, γ_{10}, τ_{00}, τ_{01}, and τ_{11}. Then we perform a likelihood ratio test (LRT) between these two models. Let the likelihood of each model be L_0 and L_1. The test statistic $D = -2(\log L_0 - \log L_1)$ is known to follow a 50:50 mixture of χ^2_1 and χ^2_2 distributions under H_0 (Stram & Lee, 1994). We report the test statistic D and the corresponding p-value.

Moderation Test. We use a similar model to find out if a geometric descriptor covariate Z moderates the relationship between X and Y. Let Z_j be the value of Z for dataset j. We fit the model

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + \epsilon_{ij}, \qquad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \tag{9}$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01}Z_j + u_{0j}, \tag{10}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}Z_j + u_{1j}, \qquad \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{01} & \tau_{11} \end{bmatrix} \right). \tag{11}$$

The cross-level interaction coefficient γ_{11} shows the level of interaction, or how much Z moderates the slope β_{1j}. We perform a Wald test (Wald, 1943) on the null hypothesis $\gamma_{11} = 0$ and report the corresponding p-value. Also we summarize the magnitude of moderation by the proportion of variance τ_{11} explained by introducing Z (Raudenbush & Bryk, 2002),

$$R^2_{\text{slope}} = 1 - \tau_{11}^{(\text{mod})} / \tau_{11}^{(\text{omn})}. \tag{12}$$

C. Theoretical Analysis on a Toy Model

In this section, we provide a theoretical analysis on why the slope of quality to FID should depend on the geometric descriptors using a simple toy model. FID is defined as the 2-Wasserstein distance between the Inception-v3 features of reference and generated image sets, where the distribution of features are assumed to be normal. For two normal distributions $\mathcal{N}(\mu_r, \Sigma_r)$ and $\mathcal{N}(\mu_g, \Sigma_g)$, 2-Wasserstein distance has a closed form

$$W_2^2 = \|\mu_r - \mu_g\|_2^2 + \operatorname{tr}\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r^{1/2}\Sigma_g\Sigma_r^{1/2}\right)^{1/2}\right). \tag{13}$$

Toy Model. We define reference and generated feature distributions for our toy model. Let the reference features $\mathbf{x} \in \mathbb{R}^D$ follow a normal distribution

$$\mathbf{x} \sim \mathcal{N}(0, P) \tag{14}$$

where $P = \operatorname{diag}(1, \ldots, 1, 0, \ldots, 0)$ is a projection matrix with r ones and $D-r$ zeros. The features span r linear dimensions on the feature space. Then we obtain generated feature samples $\hat{\mathbf{x}} \in \mathbb{R}^D$ by adding noise to the reference features as

$$\hat{\mathbf{x}} = \mathbf{x} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \lambda^2 I) \tag{15}$$

where $\lambda > 0$ is the noise level and I is the identity matrix. Then the generated feature samples follow another normal distribution

$$\hat{\mathbf{x}} \sim \mathcal{N}(0, P + \lambda^2 I). \tag{16}$$

Using the definition of 2-Wasserstein distance we have

$$W_2^2 = \operatorname{tr}\left(P + \left(P + \lambda^2 I\right) - 2\left(P^{1/2}\left(P + \lambda^2 I\right)P^{1/2}\right)^{1/2}\right) \tag{17}$$

$$= \operatorname{tr}\left(2P + \lambda^2 I - 2\sqrt{1+\lambda^2}P\right) \tag{18}$$

$$= D\lambda^2 + 2r\left(1 - \sqrt{1+\lambda^2}\right) \tag{19}$$

$$= (D-r)\lambda^2 + \mathcal{O}\left(\lambda^4\right). \tag{20}$$

Since $0 < r \leq D$, a large r value reduces the magnitude of FID's response to a change in quality, and a small r value amplifies it.

Geometric Descriptors. When the number of samples n goes to infinity, we have

$$\operatorname{erank}(A) = \operatorname{rank}(A) = r. \tag{21}$$

Also, when r is larger, the distance to the k-th nearest neighbor should increase since there are more linear dimensions in which the samples are spread, resulting in lower distributional density. Therefore the geometric descriptors govern the magnitude of FID's response to quality changes in this toy model. Reproducing the empirical sign reversal on dispersed datasets, however, would require a richer noise model that captures anisotropic mismatch between the generated and reference covariances. We leave this to future work.

D. Future Work

An important direction for future work is to test whether our findings generalize across different generative models. In this paper, we use Stable Diffusion 1.5 (Rombach et al., 2022) with the DDIM sampler (Song et al., 2020) to isolate the effect of the reference dataset. However, the six reference datasets span markedly different domains, from single-domain face images to open-domain natural scenes. As a result, the generator may fit these domains to different degrees, and such variation could in principle act as an unmeasured covariate alongside dataset geometry. Because we are not aware of a single openly available generator that is competitive on both face synthesis and open-domain text-to-image generation, we adopt a widely benchmarked diffusion model as a representative testbed. A natural extension is therefore to replicate our moderation analysis with other generator families and samplers, in order to test whether the dependence on reference geometry is specific to Stable Diffusion 1.5 with DDIM.

Another important direction for future work is to expand the set of reference datasets. The literature on generative model evaluation relies on a relatively small pool of reference datasets, often inherited from prior benchmarks (Dehghani et al., 2021; Otani et al., 2023). We selected six datasets to span the concentrated-to-dispersed axis described in Section 2.3, which was sufficient to reject the omnibus null hypothesis and to explain a large fraction of the between-dataset slope variance using geometric descriptors. With a larger and more diverse set of reference datasets, between-dataset variance components could be estimated more precisely, and additional candidate moderators could be tested with less risk of overfitting at the dataset level.

E. Licenses

- **Stable Diffusion 1.5** - weights released under the CreativeML Open RAIL-M license (v1.0; https://github.com/CompVis/stable-diffusion/blob/main/LICENSE)
- **FID** - clean-FID implementation by Parmar et al., released under the MIT License (v1.0; https://github.com/GaParmar/clean-fid/blob/main/LICENSE)
- **FFHQ**:
  - Dataset materials (metadata, scripts, documentation) released under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license (CC BY-NC-SA 4.0; https://creativecommons.org/licenses/by-nc-sa/4.0/) by NVIDIA Corporation.
  - Individual images carry their original Flickr-author licenses (a mix of CC BY 2.0, CC BY-NC 2.0, Public Domain Mark 1.0, CC0 1.0, and U.S. Government Works); per-image license is recorded in the dataset metadata (https://github.com/NVlabs/ffhq-dataset/blob/master/LICENSE.txt).
- **CelebA-HQ** - released for non-commercial research purposes only under the CelebA dataset terms by MMLab, CUHK (https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html).
- **MJHQ-30K** - released by Playground AI on HuggingFace without an explicit dataset-level license (https://huggingface.co/datasets/playgroundai/MJHQ-30K); underlying images are Midjourney generations governed by the Midjourney Terms of Service (https://docs.midjourney.com/hc/en-us/articles/32083055291277-Terms-of-Service).
- **ImageNet** - released for non-commercial research and educational use under the ImageNet Terms of Access by Princeton University and Stanford University (https://image-net.org/accessagreement).
- **Flickr30K**:
  - Annotations released for non-commercial research and educational use by Hockenmaier et al. (https://shannon.cs.illinois.edu/DenotationGraph/).
  - Underlying images governed by Flickr Terms of Use; users must comply with Flickr's rules when reusing or redistributing any Flickr30K images.
- **MS COCO 2014**:
  - Annotations released under the Creative Commons Attribution 4.0 International license (CC BY 4.0; https://creativecommons.org/licenses/by/4.0/)
  - Underlying images governed by Flickr Terms of Use; users must comply with Flickr's rules when reusing or redistributing any COCO images.