FIXING THE FRÉCHET INCEPTION DISTANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

The Fréchet Inception Distance (FID) is the staple metric to assess the progress of image generation models. In practice, its evaluation requires a large number of samples ($n \geq 50$k, both real and generated) to be accurate. Hence, the FID is not suitable for scenarios involving a small number of samples, which are common in specialized domains. In this work, we introduce FxD: a strictly compatible replacement to the usual closed-form of the FID. We replace the classical estimator of the Bures-Wasserstein distance between the two population covariance matrices by a variant from the random matrix theory. We show that the classical FID is still biased under the standard evaluation protocol ($n = 50$k) due to its slow convergence. In contrast, FxD converges quickly to the desired quantity, with sample sizes as low as a few thousands images. We demonstrate the interest of FxD for fast and frugal evaluation of generative models both on synthetic datasets and real-world datasets at various scales, including ImageNet-1k, CUB200, Oxford102 and two specialized datasets from medical imaging and remote sensing.

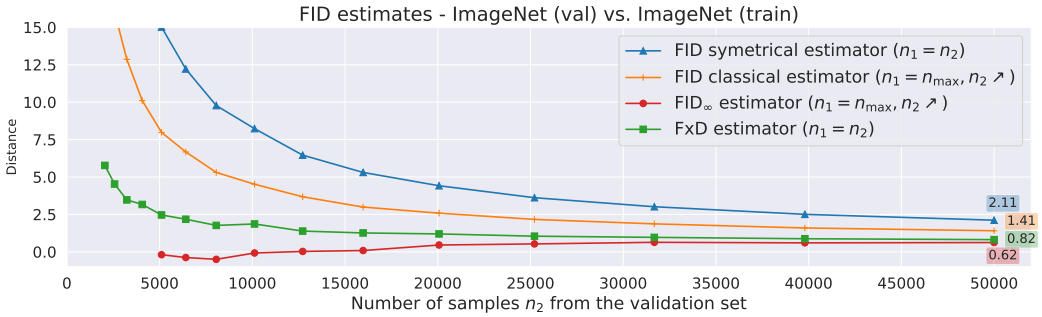


Figure 1: FID estimations using the Inception v3 features Szegedy et al. (2016) for $2048 \leq n \leq 50000$ validation images against the ImageNet-1k train set. Existing FID estimators either fail to converge or provide incorrect estimates (*e.g.* negative distances for FID$_\infty$). Our proposed FID estimator, FxD, quickly converges thanks to its improved estimation of the Wasserstein distance.

1 INTRODUCTION

The Fréchet Inception Distance (FID) was introduced by Heusel et al. (2017) in order to assess the performance of image generation models. It evaluates the perceptual quality of the generated images by comparing them to real ones in the latent space of a deep pre-trained convolutional neural network. Concretely, this metric computes an estimation of the Wasserstein distance between the distributions of real and generated images deep features. Thus, image generators aim to reach the lowest possible FID, indicating that the generated images follow the same (feature) distribution as the real images. This metric has since become mandatory in virtually any work related to deep image generation, and is often brandished as an indicator of the progress of image synthesis (Karras et al., 2020b;a; Ho et al., 2020; Rombach et al., 2022; Lipman et al., 2023; Chefer et al., 2026; Yu et al., 2026). As a sign of its importance, Yang et al. (2026) even stated that "*For nearly a decade, the community has been collectively performing "gradient descent" on this single metric to push the state of the art*".

The FID was initially introduced using the image classifier Inception v3 (Szegedy et al., 2016) as the pre-trained feature extractor. This choice results in some sensitivity to hardware implementations or

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

differences in pre-processing, such as the interpolation algorithm used to resize the images (Parmar et al., 2022). It also induces some intrinsic evaluation bias, *e.g.* invariance to transformations used as data augmentation and low sensitivity to some classes, that have been investigated multiple times(Jung & Keuper, 2021; Kabra & Balakrishnan, 2024). Consequently, Kynkäänniemi et al. (2023) recommended to replace Inception by a self-supervised model such as CLIP (Radford et al., 2021), while Stein et al. (2023) suggested using DINOv2 (Oquab et al., 2024), arguing that the extracted features would be less biased and more closely aligned to human judgment. On paper, one "only" needs to replace the encoder by a suitable pre-trained network to build a new metric, so the FID underwent many declinations in the literature. This has been notably leveraged to evaluate domain-specific generative models: Fréchet audio distance (FAD) (Kilgour et al., 2019), Fréchet video distance (FVD) (Unterthiner et al., 2019), and even Fréchet ChemNet distance (FCD) (Preuer et al., 2018) in molecule generation for drug design. This versatility has participated to keep the FID alive and popular, despite the introduction of alternative metrics.

However, the choice itself of the Wasserstein distance – and its particular choice of implementation – as the underlying metric is rarely questioned. It is worth mentioning that in their original paper, Heusel et al. (2017) propose to compute the FID based on at least 50k samples for both real and synthetic images. Yet, this number is seemingly arbitrary and was never properly motivated. Karras et al. (2020a) (section 4.3) observed later that the FID has a significant biased in the "low-data regime", *i.e.* when it is computed over a few thousands images. Jeevan et al. (2024) observed that FID indeeds requires "tens of thousands" samples to converge. Since image generators can produce an infinite number of synthetic images, the only limiting factor for FID evaluation is the number of real images, excluding computational costs of data generation. Meanwhile, there are many datasets smaller than 50k samples, and even datasets where only $\approx$ 1k real images are available. This is even fewer than the number of dimensions of the feature space of Inception v3, which uses features of dimensionality $p = 2048$. In reality, this means that the FID is often computed between a small number of real samples (a few thousands real images) and a large number of generated samples (around 50k synthetic images). Unfortunately, changing the number of samples does not leave the FID untouched. Chong & Forsyth (2020) observed a bias of the FID when computed with few samples, and introduced FID_∞ as a de-biased estimator. Their solution is to evaluate FID at different number of samples n, fit a linear regression over the $\frac{1}{n}$ values and extrapolate at 0 to obtain an estimate of the FID at "infinite number of samples". While this is effective, FID_∞ thus requires multiple evaluations of the FID, leading to a more computationally intensive evaluation. More recently, Dufour et al. (2026) observed brittleness of the FID estimation, notably by decomposing the sources of variances in its evaluation (seed, sample set, etc.). These challenges led some authors to forego the FID altogether. Various alternatives have been introduced, with different underlying distance, such as KID (Bińkowski et al., 2018; Wang et al., 2025), CMMD (Jayasumana et al., 2024), or MIND (Berthet et al., 2026). While these metrics have their respective merits, they are not compatible replacements and have not yet been as commonly adopted as the FID.

In this work, we challenge some of the common misconceptions held about the FID. In particular, we show that the FID@50k protocol is broken and does not measure what it was pretends to. We propose to repair it by showing it is actually possible to compute the FID on small datasets without sacrificing the accuracy of the estimated value. We leverage recent approaches from the random matrix theory, that introduced new estimators for various classes of distances between covariance matrices (Couillet et al., 2019; Pereira et al., 2024a;b). Notably, Tiomoko & Couillet (2019) introduced an efficient estimator of the Bures-Wasserstein distance that can be used to evaluate the FID. This estimator converges significantly faster than the classical implemented one, especially when the number of samples n is close to the dimensionality p.

Practically, Inception v3 features used in the FID have a dimension $p = 2048$. We will first show that the plug-in estimator of the Wasserstein distance is unreliable in low-data regime, and cannot be reliably used to evaluate image generators on small datasets. We also show that the standard $n = 50k$ protocol is equally unreliable, due to slow convergence of the plug-in estimator, as shown in Fig. 1. Then, we will show that the FID can be robustly estimated with only a few thousands real and synthetic examples using an random matrix theory-based estimator we refer to as FxD. Finally, we will explore the consequences of the improper FID computations and show practical applications of our FxD to accelerate the evaluation of large-scale image generators by computing the FID with less than 10k samples, both real and generated.

2 BACKGROUND AND DISCUSSION

In the framework of image synthesis, let us assume that $g : z \in \mathbb{R}^d \to I \in \mathbb{R}^{W \times H \times 3}$ is an image generator that takes as an input a random noise vector z of dimension d, and outputs a color image I of spatial dimensions $W \times H$. The FID as proposed in Heusel et al. (2017) evaluates the quality of this generator by computing the a distance between the generated image distribution and a reference image set. This comparison is not done in the image space, but within the latent space of a feature extractor $f : I \in \mathbb{R}^{W \times H \times 3} \to x \in \mathbb{R}^p$, that projects an image into a feature vector x of dimension p. In most cases, f is the Inception v3 model (Szegedy et al., 2016) trained on ImageNet-1k.

Let $\mu = \{x_i^r\}_{i=1}^{n_r} = \{x^r = f(I^r) \mid I^r \in \mathcal{D}_{\text{real}}\}$ denote the empirical distribution of the features extracted from the reference dataset $\mathcal{D}_{\text{real}}$ (*e.g.* N reserved validation images from the dataset used to train the generator g); and let $\nu = \{x_i^g\}_{i=1}^{n_g} = \{x^g = f(g(z)) \mid z \sim \mathcal{N}(0, 1)\}$ denote the empirical distribution of the features extracted on a set of synthetic images. The FID builds upon the Wasserstein-2 distance (Waserstein, 1969) between two distributions μ and ν in $\mathbb{R}^n$, defined as:

$$d_W^2(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|x_1 - x_2\|^2 \, \mathrm{d}\gamma(x_1, x_2) \, . \tag{1}$$

Evaluating the FID of g thus means evaluating the above quantity, which is intractable in the general case (Villani et al., 2009). Resorting to numerical methods (Peyré & Cuturi, 2019; Flamary et al., 2021) is also considered impractical in this scenario, thus the FID relies on the further assumption that μ and ν are drawn from Gaussian distributions, *i.e.*, that $x^r \sim \mathcal{N}(m_r, \Sigma_r)$, and $x^g \sim \mathcal{N}(m_g, \Sigma_g)$. The theoretical expression of the Wasserstein distance between two multivariate Gaussian distributions (Peyré & Cuturi, 2019, Remark 2.31) reduces to the so-called Fréchet distance (Fréchet, 1957):

$$d_W^2(\mathcal{N}(m_r, \Sigma_r), \mathcal{N}(m_g, \Sigma_g)) = \|m_r - m_g\|_2^2 + \operatorname{tr}\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r^{\frac{1}{2}} \Sigma_g \Sigma_r^{\frac{1}{2}}\right)^{\frac{1}{2}}\right), \tag{2}$$

in which $\Sigma^{\frac{1}{2}}$ denotes the only root of Σ in the set of $p \times p$ positive semi-definite matrices $\mathcal{S}_p^+$. This yields a simple closed-form expression that can be used in practice with the empirical distributions μ and ν by replacing the means and covariance matrices by their estimated values. The FID is thus finally defined as:

$$\text{FID}(g) = \|\hat{m}_r - \hat{m}_g\|_2^2 + \operatorname{tr}\left(\hat{\Sigma}_r + \hat{\Sigma}_g - 2\left(\hat{\Sigma}_r^{\frac{1}{2}} \hat{\Sigma}_g \hat{\Sigma}_r^{\frac{1}{2}}\right)^{\frac{1}{2}}\right), \tag{3}$$

in which we use the plug-in sample mean and sample covariance matrix as estimates:

$$\hat{m}_r = \frac{1}{n_r} \sum_{i=1}^{n_r} x_i^r, \quad \text{and} \quad \hat{\Sigma}_r = \frac{1}{n_r} \sum_{i=1}^{n_r} (x_i^r - \hat{m}_r)(x_i^r - \hat{m}_r)^\top, \tag{4}$$

and where $\hat{m}_g$ and $\hat{\Sigma}_g$ are defined accordingly. Last but not least: following from Heusel et al. (2017), the standard procedure is to evaluate the FID with all the real images ($n_r = N$) and $n_g = 50\text{k}$ generated images.

Frow the above derivation, we see that the construction of the FID relies on two major assumptions:

- *Gaussianity*: Reducing (1) to (2) requires the features in the latent space to follow a Gaussian distribution. This is most likely not the case in reality, but the use of $n_r = n_g = 50\text{k}$ samples could allow one to argue that the central limit theorem applies. Whether the expression in (2) can be claimed to be the actual Wasserstein distance or not, it still allows us to evaluate discrepancies between the mean and covariance matrix between the two distributions at hand (even non-Gaussian ones). Considering alternate distances (Mian et al., 2021; Thanwerdas & Pennec, 2023) or plug-in estimators (Maronna, 1976) for these quantities could provide more robust evaluation metrics, but these would eventually be redundant with what the FID is measuring.
- *Asymptotic regime*: It is commonly assumed that n_r and n_g are large enough for (3) to provide an accurate evaluation of the expression in (2). Though this estimator is known to be biased at low sample support (Chong & Forsyth, 2020), the recommended values $n_g = 50\text{k}$ and $n_r \geq 50\text{k}$ are rarely debated, nor investigated. Whether if sufficient or not,

this requirement can be difficult to meet in practice, *e.g.* in application-oriented and/or specialized datasets with limited sample support. The study in Stein et al. (2023) mitigates some concerns by evidencing that, even if not accurately evaluated, most reported FID values do not affect the general trend in model rankings. The FID could thus be used as a relative way to evaluate models against one another, but this underlines a lack of reproducible metric that is not too affected by the sample size.

In the following we propose an alternate evaluation method for the expression in (2) that is suitable to non-Gaussian distributions and low sample support. We rely on the random matrix theory-based estimator of the Bures-Wasserstein distance proposed in Tiomoko & Couillet (2019). On synthetic and real data, we will notably observe that while $n_r = n_g = 50\text{k}$ might not be enough to get an accurate FID with the classical expression in (3), the proposed FxD can achieve such estimation with one order of magnitude less samples.

3 FxD: A LOW-SAMPLE SUPPORT ESTIMATOR OF THE FID

A main issue with the standard implementation of the FID is that the the Bures-Wasserstein matrix distance (Bhatia et al., 2019), defined as

$$d_{\text{BW}}(\Sigma_r, \Sigma_g) = \text{tr}\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r^{\frac{1}{2}}\Sigma_g\Sigma_r^{\frac{1}{2}}\right)^{\frac{1}{2}}\right) \tag{5}$$

is poorly evaluated by the expression in (3). This can be understood as follows: the problem is to estimate the quantity in (5) given two sample sets $\mu = \{x_i^r\}_{i=1}^{n_r}$ and $\nu = \{x_i^g\}_{i=1}^{n_g}$ with respective covariance matrices Σ_r and Σ_g. The classical approach consists in first computing two covariance matrix estimators $\hat{\Sigma}_r$ and $\hat{\Sigma}_g$ (the sample covariance matrices as defined in (4)), and then using them as plug-in in the expression in (5). This two-step approach was shown to be inefficient in several works from the random matrix theory (Couillet et al., 2019; Pereira et al., 2024a;b). Notably, Tiomoko & Couillet (2019) proposes a direct estimator $\hat{d}_{\text{BW}}(\mu, \nu)$ of the desired quantity under the assumption that both the sample size and the observation dimension tend to infinity, while their quotient converges to a positive quantity. The main assumptions and result is reported below matching our notations for convenience[1].

Theorem 1 *(Tiomoko & Couillet, 2019, Theorem 2 & Corollary 1)* Let $\mu = \{x_i^r\}_{i=1}^{n_r}$ and $\nu = \{x_i^g\}_{i=1}^{n_g}$ be two sets of independent and identically distributed random vectors, with respective covariance matrix Σ_r and Σ_g, and with finite fourth order moment entries. Let us assume that $n_r = n_g \triangleq n$. For $a \in (r, g)$, as $n \to \infty$, $p/n \to c \in (0, 1)$, and $\limsup_p \max\{||\Sigma_a^{-1}||, ||\Sigma_a||\} < \infty$ for $||\cdot||$ the operator norm. Let $\lambda \in \mathbb{R}^p$ be the vector of entries $\lambda_1 \leq \ldots \leq \lambda_p$ containing the ordered eigenvalues of $\hat{\Sigma}_r\hat{\Sigma}_g$ (with sample covariance matrices defined in (4)), and $\Lambda = \text{diag}(\lambda)$. Let $\xi_1 \leq \ldots \leq \xi_p$ be the ordered eigenvalues of $\Lambda = \Lambda - \frac{1}{n}\sqrt{\lambda}\sqrt{\lambda}^\top$. Then

$$\frac{1}{p}d_{\text{BW}}(\Sigma_r, \Sigma_g) - \frac{1}{p}\hat{d}_{\text{BW}}(\mu, \nu) \xrightarrow{\text{a.s.}} 0 \tag{6}$$

with $d_{\text{BW}}(\Sigma_r, \Sigma_g)$ defined in (5), and where

$$\hat{d}_{\text{BW}}(\mu, \nu) = \text{tr}(\hat{\Sigma}_r + \hat{\Sigma}_g) - 4n\sum_{j=1}^{p}\left(\sqrt{\lambda_j} - \sqrt{\xi_j}\right). \tag{7}$$

Notice that Theorem 1 only requires the assumption of a finite fourth order moment. This includes the case $x^r \sim \mathcal{N}(\mu_r, \Sigma_r)$ and $x^g \sim \mathcal{N}(\mu_g, \Sigma_g)$, but this assumption is not a requirement. The expression (7) then provides a consistent estimator of the Bures-Wasserstein distance (5) in the asymptotic regime in which n and p tend to infinity with a fixed ratio $p/n = c$ (while $c < 1$ still requires $n > p$). This estimator coincides the usual one in (3) in the standard asymptotic regime

[1] In particular, Tiomoko & Couillet (2019) assumes zero-mean random vectors, but the result directly generalizes to samples covariance matrices constructed from demeaned samples as defined in (4).

(a) FID

```python
def FID(mu1, sigma1, mu2, sigma2):
  diff = mu1 - mu2
  t1 = diff.dot(diff)
  t2 = np.trace(sigma1 + sigma2)
  C2 = sigma1 @ sigma2
  C, _ = linalg.sqrtm(C2)
  t3 = 2 * np.trace(C)
  return t1 + t2 - t3
```

(b) FxD

```python
def FxD(mu1, sigma1, mu2, sigma2, n):
    diff = mu1 - mu2
    t1 = diff.dot(diff)
    t2 = np.trace(sigma1 + sigma2)
    C2 = sigma1 @ sigma2
    ev = np.linalg.svd(C2, compute_uv=False)
    sqrt_ev = np.sqrt(ev).reshape(-1, 1)
    M = np.diag(ev) - (1/n)*sqrt_ev @ sqrt_ev.T
    eta = np.linalg.svd(M, compute_uv=False)
    sqrt_eta = np.sqrt(eta).reshape(-1, 1)
    t3 = 4 * n * np.sum(sqrt_ev - sqrt_eta)
    return t1 + t2 - t3
```

Figure 2: NumPy implementation in Python of the classical FID and FxD. Sample means and sample covariance matrices inputs are defined as in (4).

with fixed p and when $n \to \infty$. Otherwise, it includes additional bias correction terms constructed from the eigenvalues $\{\xi_j\}_{j=1}^{p}$. In practice, (7) can still be evaluated with finite values of n and p, as illustrated by the `NumPy` (Harris et al., 2020) code snippet in Figure 2. We will then evidence that it provides a more accurate estimation than (3) at low sample support, which can then be leveraged to better evaluate equation 2. This leads to the definition of the proposed FxD:

$$\mathrm{FxD}(g) = \|\hat{m}_r - \hat{m}_g\|_2^2 + \hat{d}_{\mathrm{BW}}(\mu, \nu). \tag{8}$$

A last remark is that the result in Tiomoko & Couillet (2019) provides a more general formula for the uneven case $n_r \neq n_g$. This expression however involves an intricate integral that is intractable in closed-form, and thus, requires to be numerically evaluated. This could be leveraged for a more accurate estimation of the FID when n_r is limited, but when n_g is large (since we could in principle generate as much samples as we want from g). Still, we experienced numerical instability for typical feature dimensions ($p \geq 512$) with the standard `quad` function from `SciPy` (Virtanen et al., 2020). We argue here that this is not a major limitation since the results obtained in the case $n_r = n_g$ will already be accurate at low sample support. This also means that the proposed FxD allows for frugality in both real and generated samples.

4 On the FID evaluation at low sample support

4.1 Synthetic experiments

To assess the reliability of different FID estimators, we design synthetic experiments using Gaussians of known covariances, *i.e.* $X_1 \sim \mathcal{N}(\mu_1, \Sigma_1), X_2 \sim \mathcal{N}(\mu_2, \Sigma_2)$. This allows us to know the extract Bures-Wasserstein distance between the underlying distribution and generate as populations as large as needed. For brevity, we focus here synthetic experiments with distributions of equal moments, *i.e.* $\mu_1 = \mu_2$ and $\Sigma_1 = \Sigma_2$, and generate up to 200 000 samples. We consider three types of covariance matrices: identity, low-rank spiked matrices and Toeplitz matrices. To simulate the behavior of the estimators in similar conditions to FID evaluation, we use two setups: $p = 2048$ (same dimensionality as Inception-v3) and $p = 768$ (same dimensionality as CLIP).

We compare four estimators: the standard FID estimator using all "real" data ($n_1 = 200000$) and a variable number of "fake" data n_2, the symmetrical FID estimator using as many "real" and "fake" samples ($n_1 = n_2$), FID_∞ and FxD. We show the resulting curves of FID@n in Fig. 3.[2] Note that, since the underlying distributions are the same, the actual Bures-Wasserstein distance is null.

We observe that using the asymetrical FID by evaluating against all the real data is indeed better than the symmetrical version, but not by much. Both estimators exhibit slow convergence, and produce large bias in the "standard" $n \leq 50$k regime. FID_∞ fares significantly better, with a consistent estimation, robust to sample size n. However, this is at the expense of computation time (FID_∞

[2]We also report runtimes and memory usage in appendix B.

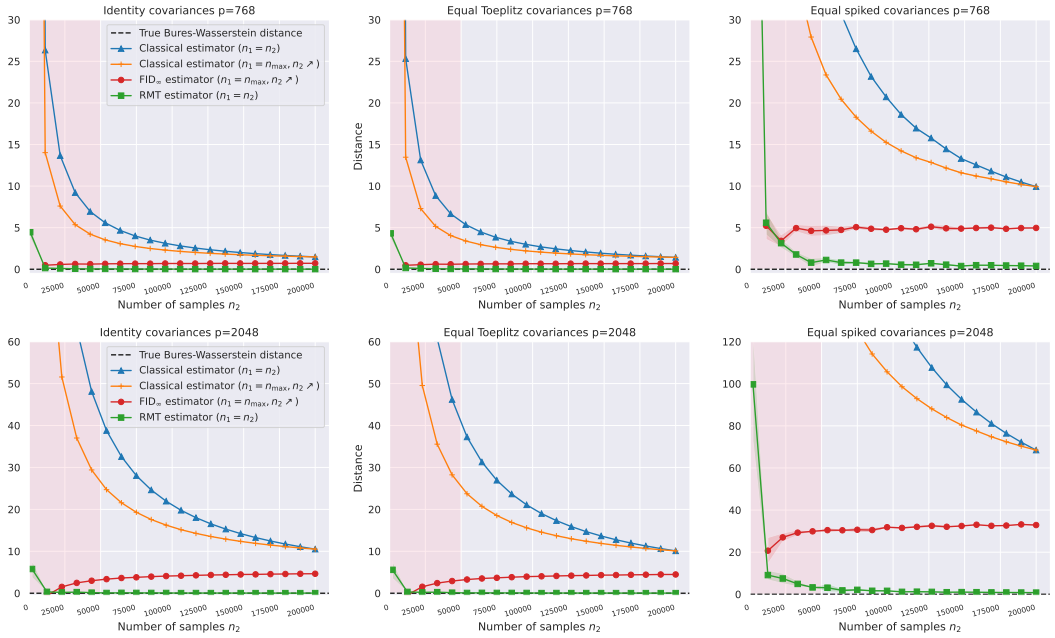


Figure 3: Synthetic experiments on FID estimation over centered Gaussians of equal covariances with variable number of observations (average over 5 random samplings) for dimension $p = 768$ (as typically used for *e.g.* CLIP or DINO) and $p = 2048$ (*e.g.* Inception v3). The classical estimator of the Wasserstein distance converges orders of magnitude slower than FID$_\infty$ or the RMT estimator in all considered cases. Highlighted parts on the left correspond to $n \leq 50000$ samples, which is the commonly considered support used for FID in the literature.

requires a recommended $k = 15$ FID evaluations to extrapolate). Moreover, it tends to underestimate at very low sample support and still overestimates under large data-regimes.[3]

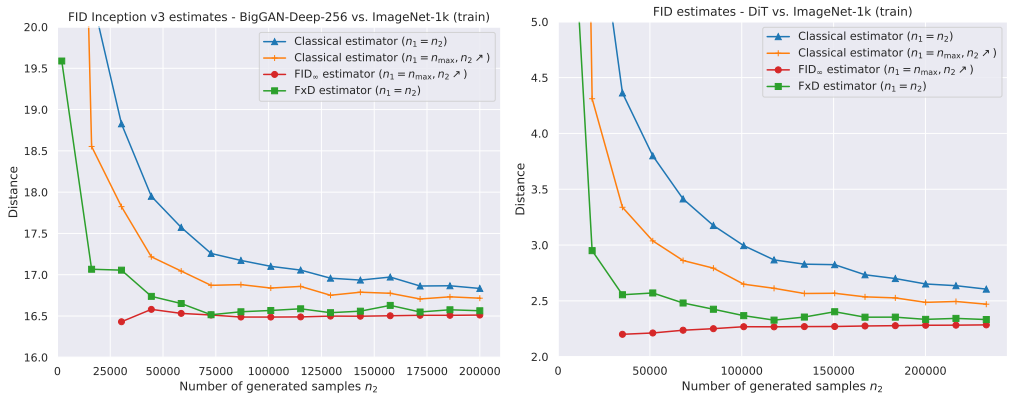
In comparison, FxD converges an order of magnitude quicker than the standard estimator. This is consistent with the results of Tiomoko & Couillet (2019), although we show here they generalize at even higher dimensions. As expected, the gap between the classical and the RMT-based estimators widens when p increases, as the n/p ratio becomes less favorable. This makes the RMT estimator particularly well-suited to FID evaluation, which uses high-dimensionality features.

4.2 Case study on ImageNet

The ILSVRC 2012 "ImageNet" dataset is a staple of generative modeling for image synthesis. It is the main evaluation dataset on which FID is reported. The most common protocol consists in computing the FID of 50k synthetic images against its train set. To alleviate computational burden, feature statistics (mean and covariance) are pre-computed on the train set, leaving "only" 50 000 to generate. We show below that generating that many images is not enough to robustly estimate the FID, unless FxD is used instead of the classical estimator.

To demonstrate this, we generate respectively 200 000 and 220 000 images using two heavily-cited generators from the image synthesis literature: BigGAN (Brock et al., 2018) and DiT (Peebles & Xie, 2023). We consider both models pretrained on ImageNet 256×256, based on their respective PyTorch implementations. We then compute the FID using the classical estimator – both asymetrical ($n_r = N$ the entire training set) and symetrical ($n_r = n_g$) and FxD. We use an approximate version of the FID$_\infty$ by extrapolating the FID curve instead of resampling 15 batches to alleviate the computational

[3]While our implementation of FID$_\infty$ is identical to Stein et al. (2023), it slightly differs from the original paper (Chong & Forsyth, 2020). We acknowledge that the residual bias might be a quirk *e.g.* of the choice of batches sizes used to fit the interpolation and not an intrinsic property of FID$_\infty$. See Appendix C.3 for details.



(a) BigGAN Brock et al. (2018)

(b) DiT Peebles & Xie (2023)

Figure 4: Comparison of FID estimations at different number of generated samples from BigGAN and DiT, trained on ImageNet-1k Deng et al. (2009) (256×256). FID remains overestimated using the classical estimators even at n =200 000 samples, and the FID@50k remains far from the asymptotic regime. FxD is more accurate than FID@50k with only a few thousand samples, and has mostly converged for $n \leq$50 000.

burden. Results are reported in Fig. 1 as a sanity check, evaluating the FID of ImageNet-1k train vs. val splits, and in Fig. 4 evaluating the generated images against the ImageNet-1k train split.

First, at $n_g = 50$k, the classical FID estimator (orange curve) is still far from converging, which is consistent with our synthetic experiments. Second, both FID$_\infty$ and FxD converge to a very close estimate. This demonstrates that the actual Fréchet distance is ≈20% lower than what is measured by FID@50k. In addition, we observe that FxD is already as accurate as FID@50k with one order of magnitude less samples, *i.e.* ≈5000. While we have no reason to believe that the biased estimates impact relative rankings of generative models, this calls into question the theoretical validity of the FID@50k evaluation protocol. Yet, the classical FID estimator would need a wasteful amount of samples to properly estimate the underlying Wasserstein distance. Instead, FxD can produce representative estimates in low sample support, and is a lot more trustworthy than FID in the tens of thousands samples regime, allowing for reliable comparison even at different sample sizes.

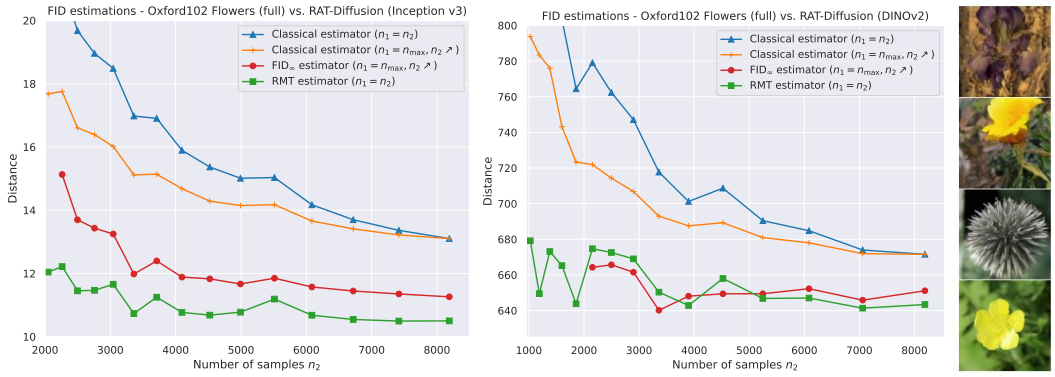
4.3 Case study on small datasets

While less common, smaller datasets (e.g. $n \leq 50000$ images) are still a fruitful playground for generative modeling. In particular, methods that improve the data efficiency of generative models, *e.g.* using pretrained models (Grigoryev et al., 2022) or efficient architectures (Karras et al., 2020a; Ye et al., 2024), often benchmark on small datasets. Specialized domains are of particular interest here due limited data availability, *e.g.* in remote sensing (Sebaq & ElHelw, 2024) or medical imaging (Yazdani et al., 2026). Unfortunately, this had led many authors to evaluate the FID of generated images against very small datasets of real images, resulting in heavily biased scores that misrepresent the actual performance of the models.

For this case study, we compute estimations using Inception v3 ($p = 2048$) and DINOv2 (ViT-L-14 backbone, $p = 1024$). Additional results with CLIP embeddings are available in appendix A.1.

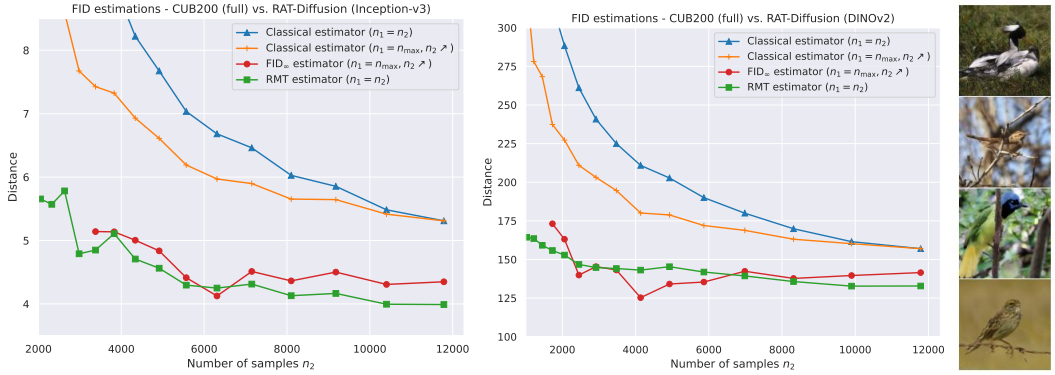
Fine-grained datasets As an example, we consider two datasets of modest sizes: CUB200-2011 (Wah et al., 2011) ($n_r = 11788$) and Oxford Flowers 102 (Nilsback & Zisserman, 2008) ($n_r = 8189$). Both share a similar motivation: classification over similar entities (birds and flowers) into respectively 200 and 102 fine-grained categories. These types of datasets often contain a few hundreds samples per class, leading to overall size in the 1000-10 000 images range. They are a common test bench to evaluate generative modeling performance of small details, necessary for fine-grained recognition.

We use pretrained checkpoints from RAT-Diffusion (Ye & Liu, 2024) to generate $n_g = n_r$ synthetic images for both datasets. Note that, as is often the case in these scenarios, the entire dataset is used,



(a) Inception v3 features ($p = 2048$) (b) DINO v2 features ($p = 1024$) Samples

Figure 5: Comparison of FID estimations at different number of generated samples from RAT-Diffusion (Ye & Liu, 2024) trained on Oxford Flowers 102 dataset (Nilsback & Zisserman, 2008).



(a) Inception v3 features ($p = 2048$) (b) DINO v2 features ($p = 1024$) Samples

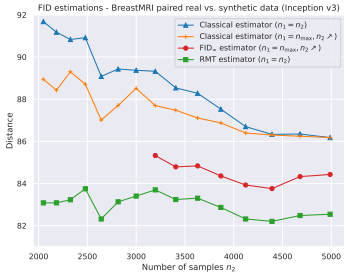
Figure 6: Comparison of FID estimations at different number of generated samples from RAT-Diffusion (Ye & Liu, 2024) trained on CUB200-2011 dataset (Wah et al., 2011).

including train, test – and validation, if it exists – splits, to mitigate the small number of samples available. Estimations at various n_g are reported in Fig. 5 for Flowers 102 and Fig. 6 for CUB200.

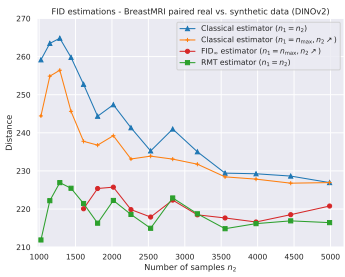
We observe similar trends on both datasets. First, the classical FID consistently overestimates, as shown by the behavior of FID$_\infty$ and FxD which both converge to approximately the same values over the full dataset. Second, consistent with experiments on synthetic data, FID$_\infty$ gives slightly larger values than FxD, confirming that it is not entirely debiased, even on real data. Third, for $n_g \approx 2p$, FxD gives an approximation that is within 20% of the final value. This demonstrates that FID can be accurately evaluated in the low-data regime, provided that an appropriate estimator is used.

Specialized domains To drive our point further, we also consider two heavily specialized domains: medical imaging and remote sensing. We compute FID estimations on the paired real-synthetic breast MRI from Konz et al. (2024) ($n_r = n_g = 4988$) (Figs. 7a and 7b) based on the Duke MRI dataset (Saha et al., 2021) and on synthetic remote sensing RGB images from RSDiff (Sebaq & ElHelw, 2024) evaluated against the test set of RSCID (Lu et al., 2018) ($n_r = n_g = 1093$) (Fig. 7c). Note that the latter pushes FID estimation to its limits, considering than the number of samples is inferior of the features dimensionality, thus making the covariance matrix estimation particularly brittle.

Once again, we observe consistent behavior of our estimators, with FID over-estimating and FxD stabilizing quickly. On very low sample support over RSCID, we argue that it is likely that the classical FID values are nowhere near their actual values. This motivates us to recommend avoiding the classical estimator in favor of FxD in all scenarios.



(a) BreastMRI – Inception v3



(b) BreastMRI – DINOv2



Estimator	Inception	DINO
FID	68.87	412.4
FID$_\infty$	63.01	392.9
FxD	33.60	297.6

(c) 1093 synthetic RSdiff images vs. 1093 real images (RSCID test set)

Figure 7: Comparison of FID estimations on two specialized datasets: paired real-synthetic Breast MRI (Konz et al., 2024; Saha et al., 2021) and satellite images from RSCID (Lu et al., 2018) generated by RSDiff (Sebaq & ElHelw, 2024).

5 Findings and Discussion

Considering our synthetic and real-word experiments, we can draw three main conclusions from our results on the evaluation of generative models:

1. **The classical FID implementation is unreliable**: the usual FID estimator from Eq. (3) does not measure what it is supposed to. Our results confirm previous observations that the FID@n is unreliable and cannot be compared for different values of n. This has led in the past to some incorrect comparisons, such as using the CIFAR10 train vs. test FID as a "lower bound" for the FID@50k, despite CIFAR10 test set having only 10 000 images.
2. **The FID$_\infty$ is not enough**: while improving the accuracy of the FID estimation, the common implementation of FID$_\infty$ still often present a biased estimate of the actual Wasserstein distance. In addition, it was not designed for low sample support, and it is computationally expensive, requiring computing FID@n up to 15 times, for values of $n \geq 5000$.
3. **The FxD is robust, cheap, and frugal**: FxD provide consistent estimations over the entire $n \geq 10$k range for usual features dimensionalities ($p = 768$ or $p = 2048$), using a single evaluation of Eq. (8). For very small sample support ($n < 10$k), it is significantly more accurate than competing approaches, and remains reasonable even in the range $n = 2p$. The FxD often requires only thousands of samples to converge, not tens of thousands.

Overall, we argue that the FID should not be evaluated using the classical estimator, as its estimates are heavily biased and too dependent on sample size. We advocate for the use of FxD as a strictly compatible plug-and-play replacement in its stead. This is facilitated by its simple closed-form expression, making it compatible with virtually any implementation of the FID – or its derivatives.

This opens some questions and several opportunities for research on generative models. First, the bias of the FID might impact the perceived model performance. Stein et al. (2023) showed that relative rankings based on the FID$_\infty$ are similar to those obtained using the FID, and we do not expect FxD to challenge this. However, the impression of progress might have been overstated due to biased estimators. Moreover, recent works, *e.g.* FID lottery (Dufour et al., 2026), have shown that the FID evaluation can be brittle due to multiple variance sources (sample set, seed, etc.). FxD might provide a more robust estimate that is less sensitive to these types of noise. Finally, training generative models by directly optimizing the FID has been a nearly decade-long wish of the community. On paper, Eq. (3) is indeed differentiable with respect to the features. Some preliminary works have even implemented sophisticated algorithms to accelerate backpropagation through FID (Doan et al., 2020; Mathiasen & Hvilshøj, 2021). Due to the assumption that FID requires large sample sizes, more recent works have looked into computing it on "virtually large" batches, *e.g.* with a queue system (Yang et al., 2026). FxD might open up new options, considering that batch sizes of 2048 (*e.g.* $n = 2p$ for DINOv2) are not that implausible on large-scale compute clusters. While we concentrated on image synthesis, similar experiments could be conducted on other modalities, such as audio (Kilgour et al., 2019) or molecules (Preuer et al., 2018). Although we do not expect strongly different results, feature dimensionality and typical sample size might slightly alter our findings on different domains.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic data sets: RAT-Diffusion (Ye & Liu, 2024), RSDiff (Sebaq & ElHelw, 2024), DiT (Peebles & Xie, 2023), and BigGAN (Brock et al., 2018).

We have not used generative AI tools to formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results.

We have used generative AI tools to create and edit software code in the following ways: i) use of GPT-4.1 to compare our implementation of the RMT estimator against the baseline MATLAB implementation from Tiomoko & Couillet (2019); ii) use of GPT-5.6 Luna to produce scripts to generate the images based on the repositories of the aforementioned generative models; iii) use of Copilot (GPT-5.6 Luna and Claude Haiku 4.5) for code auto-completion and debugging. All code produced by these tools was vetted, tested and rewritten by the first author.

We have used generative AI tools to complete our literature review regarding generative models on small datasets by searching for relevant papers with public codes and checkpoint using GPT-5.6 Luna. RAT-Diffusion (Ye & Liu, 2024) is the sole reference identified during this process that had not been identified by the authors beforehand.

We have not used generative AI tools to formulate questions for surveys or interviews, create or modify scientific figures or images, suggest experimental parameters, create or edit software code, creation of artifacts, draft parts of a research paper, transcribe recordings of research material, summarize or analyse existing literature, discover research topics or identify gaps, brainstorming, sourcing/searching for information, edit a research paper to improve readability, format references, suggest a structure for a research paper, propose a title or keywords for a research paper.

REPRODUCIBILITY STATEMENT

A canonical implementation of FxD, alongside the scripts to reproduce our main experiments is available at: `https://anonymous.4open.science/r/fxd-1ED1/`.

We commit to making all necessary data and code available to reproduce our experiments, including generated images, and extracted features, to facilitate ulterior comparisons.

REFERENCES

Quentin Berthet, Yu-Han Wu, Clement Crepy, Romuald Elie, Klaus Greff, and Michael Eli Sander. MIND: Monge inception distance for generative models evaluation. *arXiv preprint arXiv:2605.06797*, 2026.

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the Bures–Wasserstein distance between positive definite matrices. *Expositiones mathematicae*, 37(2):165–191, 2019.

Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=r1lUOzWCW`.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis. September 2018. URL `https://openreview.net/forum?id=B1xsqj09Fm&noteId=HklmZ1xqhm`.

Hila Chefer, Patrick Esser, Dominik Lorenz, Dustin Podell, Vikash Raja, Vinh Tong, Antonio Torralba, and Robin Rombach. Self-supervised flow matching for scalable multi-modal synthesis. In *Forty-third International Conference on Machine Learning*, 2026. URL `https://openreview.net/forum?id=HoThWhfxiK`.

Min Jin Chong and David Forsyth. Effectively unbiased FID and Inception Score and where to find them. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6069–6078, 2020. doi: 10.1109/CVPR42600.2020.00611.

Romain Couillet, Malik Tiomoko, Steeve Zozor, and Eric Moisan. Random matrix-improved estimation of covariance matrix distances. *Journal of Multivariate Analysis*, 174:104531, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Khoa D. Doan, Saurav Manchanda, Fengjiao Wang, Sathiya Keerthi, Avradeep Bhowmik, and Chandan K. Reddy. Image Generation Via Minimizing Fréchet Distance in Discriminator Feature Space, March 2020. URL http://arxiv.org/abs/2003.11774. arXiv:2003.11774 [cs.CV].

Nicolas Dufour, Alexei A Efros, and Patrick Perez. The FID lottery: Quantifying hidden randomness in generative model evaluation. In *ECCV 2026 Workshop on Empirical Theory in Representation Learning*, 2026. URL https://openreview.net/forum?id=O1uayf5dX2.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, et al. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

Maurice Fréchet. Sur la distance de deux lois de probabilité. *Annales de l'ISUP*, 6(3):183–198, 1957.

Timofey Grigoryev, Andrey Voynov, and Artem Babenko. When, why, and which pretrained GANs are useful? In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=4Ycr8oeCoIh.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

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. *Advances in Neural Information Processing Systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

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

Pranav Jeevan, Neeraj Nixon, and Amit Sethi. Normalizing Flow-Based Metric for Image Generation, October 2024. URL http://arxiv.org/abs/2410.02004. arXiv:2410.02004 [cs.CV].

Steffen Jung and Margret Keuper. Internalized biases in Fréchet inception distance. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021.

Krish Kabra and Guha Balakrishnan. F?D: On Understanding the Role of Deep Feature Spaces on Face Generation Evaluation. pp. 8327–8332, 2024. URL https://openaccess.thecvf.com/content/CVPR2024W/ReGenAI/html/Kabra_FD_On_Understanding_the_Role_of_Deep_Feature_Spaces_on_CVPRW_2024_paper.html.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114, 2020a.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8107–8116. IEEE, 2020b.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet Audio Distance: A Reference-Free Metric for Evaluating Music Enhancement Algorithms. In *Interspeech 2019*, pp. 2350–2354, 2019. doi: 10.21437/Interspeech.2019-2219.

Nicholas Konz, Yuwen Chen, Haoyu Dong, and Maciej A. Mazurowski. Anatomically-Controllable Medical Image Generation with Segmentation-Guided Diffusion Models. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pp. 88–98, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-72104-5. doi: 10.1007/978-3-031-72104-5_9.

Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of imagenet classes in fréchet inception distance. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=4oXTQ6m_ws8.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. Exploring Models and Data for Remote Sensing Image Caption Generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2183–2195, April 2018. ISSN 1558-0644. doi: 10.1109/TGRS.2017.2776321. URL https://ieeexplore.ieee.org/document/8240966.

Ricardo Antonio Maronna. Robust M-Estimators of Multivariate Location and Scatter. *The Annals of Statistics*, 4(1):51 – 67, 1976. doi: 10.1214/aos/1176343347. URL https://doi.org/10.1214/aos/1176343347.

Alexander Mathiasen and Frederik Hvilshøj. Backpropagating through Fréchet Inception Distance, April 2021. URL http://arxiv.org/abs/2009.14075. arXiv:2009.14075 [cs.LG].

Ammar Mian, Guillaume Ginolhac, Jean-Philippe Ovarlez, Arnaud Breloy, and Frédéric Pascal. An overview of covariance-based change detection methodologies in multivariate sar image time series. *Change Detection and Image Time Series Analysis 1: Unsupervised Methods*, pp. 73–108, 2021.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

Anton Obukhov, Maximilian Seitzer, Po-Wei Wu, Semen Zhydenko, Jonathan Kyl, and Elvis Yu-Jing Lin. High-fidelity performance metrics for generative models in pytorch, 2020. URL https://github.com/toshas/torch-fidelity. Version: 0.4.0, DOI: 10.5281/zenodo.3786539.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in GAN evaluation. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11400–11410. IEEE, 2022.

William Peebles and Saining Xie. Scalable Diffusion Models with Transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182, October 2023. doi: 10.1109/ICCV51070.2023.00387. URL https://ieeexplore.ieee.org/abstract/document/10377858. ISSN: 2380-7504.

Roberto Pereira, Xavier Mestre, and David Gregoratti. Asymptotics of distances between sample covariance matrices. *IEEE Transactions on Signal Processing*, 72:1460–1474, 2024a.

Roberto Pereira, Xavier Mestre, and David Gregoratti. Consistent estimation of a class of distances between covariance matrices. *IEEE Transactions on Information Theory*, 70(11):8107–8132, 2024b.

Gabriel Peyré and Marco Cuturi. Computational optimal transport with applications to data sciences. *Foundations and trends® in machine learning*, 11(5-6):355–607, 2019.

Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Gunter Klambauer. Fréchet ChemNet distance: a metric for generative models for molecules in drug discovery. *Journal of chemical information and modeling*, 58(9):1736–1741, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, New Orleans, LA, USA, June 2022. IEEE. ISBN 978-1-6654-6946-3. doi: 10.1109/CVPR52688.2022.01042. URL https://ieeexplore.ieee.org/document/9878449/.

Ashirbani Saha, Michael R Harowicz, Lars J Grimm, Jingxi Weng, Elizabeth H Cain, Connie E Kim, Sujata V Ghate, Ryan Walsh, and Maciej A Mazurowski. Dynamic contrast-enhanced magnetic resonance images of breast cancer patients with tumor locations [data set]. *The Cancer Imaging Archive*, 10:5, 2021.

Ahmad Sebaq and Mohamed ElHelw. RSDiff: Remote sensing image generation from text using diffusion model. *Neural Computing and Applications*, 36(36):23103–23111, 2024. ISSN 1433-3058. doi: 10.1007/s00521-024-10363-3. URL https://doi.org/10.1007/s00521-024-10363-3.

George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in Neural Information Processing Systems*, 36:3732–3784, 2023.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016.

Yann Thanwerdas and Xavier Pennec. O(n)-invariant Riemannian metrics on spd matrices. *Linear Algebra and its Applications*, 661:163–201, 2023. ISSN 0024-3795. doi: https://doi.org/10.1016/j.laa.2022.12.009. URL https://www.sciencedirect.com/science/article/pii/S0024379522004360.

Malik Tiomoko and Romain Couillet. Random matrix-improved estimation of the Wasserstein distance between two centered Gaussian distributions. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pp. 1–5. IEEE, 2019.

/u/donshell. [D] A better way to compute the Fréchet Inception Distance (FID), April 2023. URL https://www.reddit.com/r/MachineLearning/comments/12hv2u6/d_a_better_way_to_compute_the_frchet_inception/.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. FVD: A new metric for video generation, 2019. URL https://openreview.net/forum?id=rylgEULtdN.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Zixiao Wang, Farnia Farzan, Zhenghao Lin, Yunheng Shen, and Bei Yu. On the distributed evaluation of generative models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 7703–7712, October 2025.

Leonid Nisonovich Waserstein. Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*, 5(3):64–72, 1969.

Jiawei Yang, Zhengyang Geng, Xuan Ju, Yonglong Tian, and Yue Wang. Representation Fréchet Loss for Visual Generation. *arXiv preprint arXiv:2604.28190*, 2026.

Milad Yazdani, Yasamin Medghalchi, Pooria Ashrafian, Ilker Hacihaliloglu, and Dena Shahriari. Flow Matching for Medical Image Synthesis: Bridging the Gap Between Speed and Quality. In James C. Gee, Daniel C. Alexander, Jaesung Hong, Juan Eugenio Iglesias, Carole H. Sudre, Archana Venkataraman, Polina Golland, Jong Hyo Kim, and Jinah Park (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pp. 216–226. Springer Nature Switzerland, 2026. doi: 10.1007/978-3-032-05325-1_21.

Senmao Ye and Fei Liu. Data extrapolation for text-to-image generation on small datasets, October 2024. URL https://arxiv.org/abs/2410.01638.

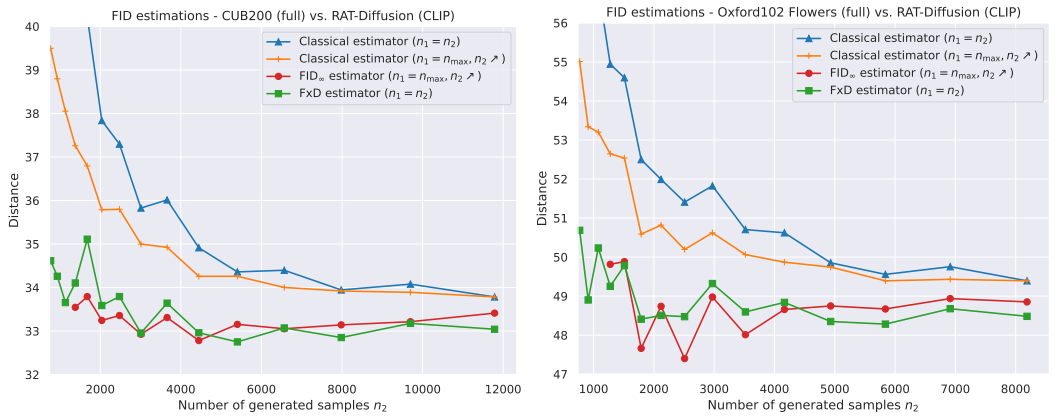
Senmao Ye, Huan Wang, Mingkui Tan, and Fei Liu. Recurrent Affine Transformation for Text-to-Image Synthesis. *IEEE Transactions on Multimedia*, 26:462–473, 2024. ISSN 1941-0077. doi: 10.1109/TMM.2023.3266607. URL https://ieeexplore.ieee.org/abstract/document/10100901.

Yongsheng Yu, Wei Xiong, Weili Nie, Yichen Sheng, Shiqiu Liu, and Jiebo Luo. PixelDiT: Pixel Diffusion Transformers for Image Generation. In *2026 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14273–14282. IEEE, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Yu_PixelDiT_Pixel_Diffusion_Transformers_for_Image_Generation_CVPR_2026_paper.html.

A Additional results

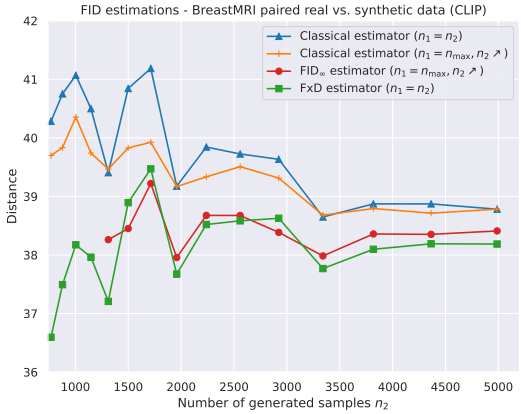
A.1 Results with CLIP features

CLIP is an alternative feature extractor that is occasionally used in replacement of Inception v3 when evaluating image generators, especially in the text-to-image setup. For the sake of completeness, we present in this section additional FID estimations using CLIP as the feature extractor. We use OpenAI's CLIP with the ViT-L 14 backbone, producing features fo dimensionality $p = 768$. This allows us to reduce even further the minimal number of samples to estimate the Fréchet distance – as a reminder, we need $n > p$ samples to obtain a full rank sample covariance matrix. This is one of the main interest of switching backbone: reducing feature dimensionality also allow for more robust estimation with less samples. As shown in Fig. 8, the behavior of the estimators on CLIP features is consistent with the one observed in the main results.



(a) CUB200 (RAT-Diff)

(b) Oxford Flowers 102 (RAT-Diff)



(c) Synthetic paired BreastMRI dataset

Figure 8: Additional results of FID estimations with CLIP ViT-L 14 as the feature extractor.

```python
def EfficientFID(mu1, sigma1, mu2, sigma2):
    diff = mu1 - mu2
    t1 = diff.dot(diff)
    t2 = np.trace(sigma1 + sigma2)
    t3 = 2 * np.real(linalg.eigvals(sigma1 @ sigma2) ** 0.5).sum()
return t1 + t2 - t3
```

Figure 9: Efficient FID computation adapted from the Torch implementation by /u/donshell (2023).

B BENCHMARKING FID COMPUTATIONS

A legitimate question in introducing FxD is computation time. Indeed, FxD involves two more matrix decompositions compared to the standard FID. Before comparing methods, it seems important to acknowledge that traditional implementation of FID as described in Fig. 2 can be improved. User /u/donshell (2023) on the /r/MachineLearning forum from Reddit proposed an alternative efficient implementation as described in Fig. 9 that is both faster and more memory efficient.

We provide in Table 1 a report of compute times and peak memory usage of the various FID estimators presented in our work, including both versions of the FID. For the sake of comparison, all methods are ran 3 times on the same CPU (AMD EPYC 7413) for $p = 64, 128, 512, 1024, 2048$ and $n_r = n_g = 50000$ samples. We *exclude* the statistics computation (sample means and sample covariances), except for FID$_\infty$ for which it is mandatory to recompute them on the smaller batch sizes. Note that in most cases, estimating the covariance (especially for large number of samples)

Table 1: Runtime and peak memory of FID estimators.

p	Estimator	Mean (ms)	Median (ms)	Min (ms)	Max (ms)	Peak (MiB)
64	FID efficient	0.61	0.61	0.60	0.63	0.10
64	FID standard	1.17	1.15	1.14	1.23	0.25
64	FID_∞	205.92	203.53	183.74	230.49	48.93
64	FxD	1.67	1.67	1.65	1.70	0.16
128	FID efficient	10.71	10.95	6.08	15.10	0.29
128	FID standard	87.33	83.97	81.02	97.00	1.00
128	FID_∞	956.95	835.14	608.24	1427.47	97.91
128	FxD	26.78	28.62	22.12	29.59	0.63
512	FID efficient	410.29	406.02	395.36	429.50	4.15
512	FID standard	534.18	541.49	504.35	556.69	14.27
512	FID_∞	7758.59	7738.62	7128.62	8408.55	394.64
512	FxD	549.51	630.41	383.91	634.21	8.01
1024	FID efficient	1193.74	1193.15	1132.38	1255.70	16.30
1024	FID standard	1698.82	1658.02	1617.31	1821.15	57.02
1024	FID_∞	22741.21	22761.04	22390.34	23072.25	797.28
1024	FxD	2243.03	2233.78	2072.11	2423.19	32.02
2048	FID efficient	4781.43	4604.88	4412.39	5327.02	64.59
2048	FID standard	7650.73	8014.10	6496.96	8441.15	228.03
2048	FID_∞	72535.00	72506.20	71919.05	73179.74	1626.55
2048	FxD	9445.39	9754.19	8822.18	9759.79	128.05

largely dominates both the execution time and the memory usage by at least one order of magnitude. However, this step is also common to all methods and we do not see any trivial way to skip it.

Overall, results are in line as expectations. In terms of compute time, EfficientFID $<$ FID $<$ FxD $<<$ FID_∞. This is unsurprising: FID_∞ requires k times more evaluations of the FID and is therefore the slowest. It is interesting to observe that FxD is only slightly slower than FID and therefore adds a only small overhead that is dwarfed in several orders of magnitude by the possibility to evaluate the FID with a smaller n, thus gaining in overall efficiency. On the memory side, we observe that our implementation of FxD is quite efficient and uses *less* RAM than the standard FID. While we did not optimize for this, it is an appreciable bonus, but the even lower footprint of EfficientFID shows that additional improvements are likely reachable.

C Implementation details

C.1 Pre-processing

Feature extraction and pre-processing (*e.g.* resizing) is based on torch-fidelity (`https://github.com/toshas/torch-fidelity`) (Obukhov et al., 2020), using the provided CLIP, DINOv2 and Inception-v3 models. All images are converted to RGB, by duplicating channels if needed (*e.g.* for grayscale images). Our implementation of FID_∞ is re-implemented based on the original paper but was tested and confirmed to give identical results as Stein et al. (2023) from the `https://github.com/layer6ai-labs/dgm-eval/` code base. We do not resize the real images outside of this preprocessing, which explains why our results differ from those reported by the original authors for ImageNet-1k comparisons.

For all models, we use the checkpoints and hyperparameters from the authors, without tuning. For Oxford Flowers 102, we compare the full dataset by concatenating train/test splits against as many generated images. We do the same for CUB200-2011 and concatenate the train/val/test splits. For BreastMRI, we simply use the paired data as is, as the authors already released a synthetic/real paired dataset. For RSDiff, the authors initially chose to evaluate only against the test set of RSCID, thus restricting the evaluation to $n_r = 1093$. While unconventional, we have chosen to stick to this protocol to show that the measured FID values are likely to be very wrong in this setup. Finally, for

BigGAN, we use a fixed truncation value of $\psi = 0.4$ without any tuning, and therefore is not directly comparable to the best FID reported in the paper.

C.2 SYNTHETIC COVARIANCES

Regarding the synthetic experiments, we build the covariance matrices as follow:

1. the identity model sets $\Sigma = \mathbf{I}$ without particular scaling;
2. the Toeplitz models sets $[\Sigma]_{ij} = \rho^{|i-j|}$ with $\rho = 0.4$, following the construction proposed by the Tiomoko & Couillet (2019) article;
3. the spiked model sets $\Sigma = \Sigma_r + \mathbf{I}$, where Σ_r is a positive semi-definite matrix of rank $r = \frac{p}{2}$ constructed as a diagonal matrix $[\Sigma_r]_{ii} = 100\frac{p-i}{p}$ if $i < \frac{p}{2}$ else 0.

The latter tends to better represent the low rank nature of real covariance matrices of features. However, we stress that our synthetic experiments use actual Gaussians to draw the samples, whereas real features are expected to be non-gaussians. This leads to a residual gap between the simulations and the real-world experiments. The final behavior of the estimators is therefore heavily dependent both on the feature extractor, but also on the feature distribution, which is dictated by the dataset.

C.3 SOME NOTES ON FID_∞

Chong & Forsyth (2020) observed that the FID exhibits a bias term that scales in $1/n$ and suggested debiasing it by extrapolation. The core idea is to evaluate k values of FID@n for different sample size $n_1 < n_2 < \cdots < n_k$. These values are used to fit a linear regression of FID@n against $1/n$, and the FID_∞ is then constructed by setting $n \to +\infty$. Their original implementation[4] and the subesquent re-implementation from Stein et al. (2023)[5] share the same default parameters:

- never use less than $n = 5000$ samples;
- evaluate the FID at $k = 15$ different number of samples;
- always evaluate the FID by varying the number of synthetic samples n_g, and use $n_r = n_{\text{max}}$;

As far as we know, these assumptions have never really been challenged. The issues observed at low sample support might be due to non-negligible additional bias terms in $O(n^{-2})$, that would render the linear regression inaccurate. It is also likely that using smaller values of k could lead to worse estimates. In practice, we observed that the FID_∞ occasionally results in negative estimates of the Fréchet distance when used in the low data regime. As stated in Tiomoko & Couillet (2019), the estimator in Theorem 1 has also non zero probability to be negative by construction, but is was not observed in our real-world experiments. It is possible that using the Sobol sequences to sample the images, as proposed in Chong & Forsyth (2020), would alleviate some of the aforementioned of the FID_∞. It is however impractical as it would require sampling the image generators differently depending on the chosen evaluation metric. We have thus decided against exploring this avenue.

[4] https://github.com/mchong6/FID_IS_infinity

[5] https://github.com/layer6ai-labs/dgm-eval