

# Rethinking FID: Towards a Better Evaluation Metric for Image Generation

Sadeep Jayasumana    Srikumar Ramalingam    Andreas Veit    Daniel Glasner  
Ayan Chakrabarti    Sanjiv Kumar  
Google Research, New York

{sadeep, rsrikumar, aveit, dglasner, ayanchakrab, sanjivk}@google.com

## Abstract

As with many machine learning problems, the progress of image generation methods hinges on good evaluation metrics. One of the most popular is the Fréchet Inception Distance (FID). FID estimates the distance between a distribution of Inception-v3 features of real images, and those of images generated by the algorithm. We highlight important drawbacks of FID: Inception’s poor representation of the rich and varied content generated by modern text-to-image models, incorrect normality assumptions, and poor sample complexity. We call for a reevaluation of FID’s use as the primary quality metric for generated images. We empirically demonstrate that FID contradicts human raters, it does not reflect gradual improvement of iterative text-to-image models, it does not capture distortion levels, and that it produces inconsistent results when varying the sample size. We also propose an alternative new metric, CMMD, based on richer CLIP embeddings and the maximum mean discrepancy distance with the Gaussian RBF kernel. It is an unbiased estimator that does not make any assumptions on the probability distribution of the embeddings and is sample efficient. Through extensive experiments and analysis, we demonstrate that FID-based evaluations of text-to-image models may be unreliable, and that CMMD offers a more robust and reliable assessment of image quality. A reference implementation of CMMD is available at: <https://github.com/google-research/google-research/tree/master/cmmd>.

## 1. Introduction

Text-to-image models are progressing at breakneck speed. Recent models such as [16, 19–21, 26] have been incredibly successful at generating realistic images that remain faithful to text prompts. As with many problems in machine learning, a reliable evaluation metric is key to driving progress. Unfortunately, we find that the most popular metric used in the evaluation of text-to-image models, the

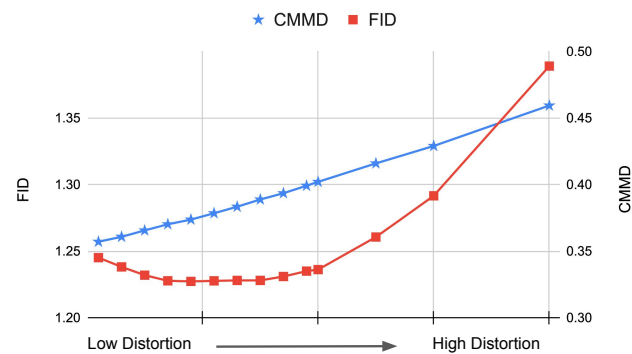


Figure 1. Behaviour of FID and CMMD (lower is better for both) under noise in the VQGAN latent space. CMMD monotonically increases with the distortion level, correctly identifying the degradation in image quality with increasing distortions. FID is wrong; it improves (goes down) for the first few distortion levels, suggesting that quality improves when these more subtle distortions are applied. See Section 6.2 for more details.

Fréchet Inception Distance (FID) [12], may disagree with the gold standard, human raters, in some important cases; and is thus ill-suited for this purpose. We identify some important limitations of the FID through statistical tests and empirical evaluations. To address these shortcomings, we propose an alternative metric: CMMD, which uses CLIP embeddings and Maximum Mean Discrepancy (MMD) distance. Figure 1 shows one of our experiments, the details of which are discussed in Section 6.2, in which FID does not reflect progressive distortion applied to images while CMMD correctly ranks the image sets based on the severity of the distortion.

Evaluating image generation models is a uniquely challenging task. Unlike traditional vision tasks such as classification or detection, we need to evaluate multiple dimensions of performance including quality, aesthetics and faithfulness to the text prompt. Moreover, these are hard-to-quantify concepts which depend on human perception. As a result, human evaluation remains the gold standard for text-to-image research. Since human evaluation is an expensive

	Fréchet distance	MMD distance
Inception embeddings	<ul style="list-style-type: none"> <li>✗ Weak image embeddings</li> <li>✗ Normality assumption</li> <li>✗ Sample inefficient</li> <li>✗ Biased estimator</li> </ul>	<ul style="list-style-type: none"> <li>✗ Weak image embeddings</li> <li>✓ Distribution-free</li> <li>✓ Sample efficient</li> <li>✓ Unbiased estimator</li> </ul>
CLIP embeddings	<ul style="list-style-type: none"> <li>✓ Rich image embeddings</li> <li>✗ Normality assumption</li> <li>✗ Sample inefficient</li> <li>✗ Biased estimator</li> </ul>	<ul style="list-style-type: none"> <li>✓ Rich image embeddings</li> <li>✓ Distribution-free</li> <li>✓ Sample efficient</li> <li>✓ Unbiased estimator</li> </ul>

Table 1. Comparison of options for comparing two image distributions. FID, the current de facto standard for text-to-image evaluation is in the upper-left corner. The proposed metric, CMMD, is in the lower-right corner and has many desirable properties over FID.

solution that does not scale well, researchers often rely on automated evaluation. Specifically, recent works have used FID and CLIP distance to measure image quality and faithfulness to the text prompts, respectively.

In this work, we call for a reevaluation of this approach, in particular, the use of FID as a measure of image quality. We highlight drawbacks of FID, such as incorrectly modeling Inception embeddings of image sets as coming from a multivariate normal distribution and its inconsistent results when varying the sample size (also noted in [5]). We empirically show that, FID can contradict human raters, does not reflect gradual improvement of iterative text-to-image models and does not capture complex image distortions.

Our proposed metric uses CLIP embeddings and the MMD distance. Unlike Inception embeddings, which were trained on about 1 million ImageNet images, restricted to 1000 classes [23], CLIP is trained on 400 million images with corresponding text descriptions [18], making it a much more suitable option for the rich and diverse content generated by modern image generation models and the intricate text prompts given to modern text-to-image models.

MMD, is a distance between probability distributions that offers some notable advantages over the Fréchet distance. When used with an appropriate kernel, MMD is a metric that does not make any assumptions about the distributions, unlike the Fréchet distance which assumes multivariate normal distributions. As shown in [5], FID is a biased estimator, where the bias depends on the model being evaluated. MMD, on the other hand, is an unbiased estimator, and as we empirically demonstrate it does not exhibit a strong dependency on sample size like the Fréchet distance. Finally, it admits a simple parallel implementation. The ability to estimate from a smaller sample size and the fast computation make MMD fast and useful for practical applications. Different options for comparing two image distributions are compared in Table 1. The existing FID metric is in the upper-left corner and has many unfavorable properties. Our proposed metric, CMMD, is in the lower-right corner and avoids the drawbacks of FID.

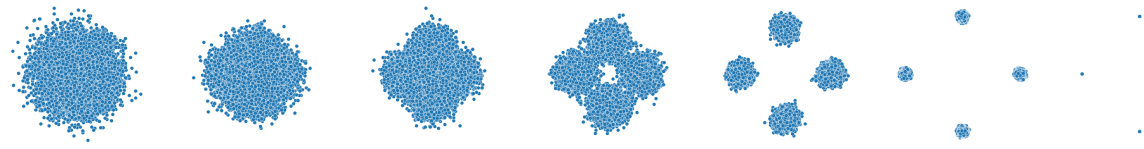
We summarize our contributions below:

- We call for a reevaluation of FID as the evaluation metric for modern image generation and text-to-image models. We show that it does not agree with human raters in some important cases, that it does not reflect gradual improvement of iterative text-to-image models and that it does not capture obvious image distortions.
- We identify and analyze some shortcomings of the Fréchet distance and of Inception features, in the context of evaluation of image generation models.
- We propose CMMD, a distance that uses CLIP features with the MMD distance as a more reliable and robust alternative, and show that it alleviates some of FID's major shortcomings.

## 2. Related Works

Generated image quality has been assessed using a variety of metrics including log-likelihood [9], Inception Score (IS) [1, 22], Kernel Inception Distance (KID) [2, 25], Fréchet Inception Distance (FID) [12], perceptual path length (PPL) [13], Gaussian Parzen window [9], and HYPE [27]. Some of these are not generally applicable. PPL evaluates the consistency of latent vectors with respect to GANs, HYPE requires human input and computing Gaussian Parzen windows can be challenging in high-dimensional spaces.

IS is calculated using the Inception-v3 model [23], which has been trained on ImageNet, to measure the diversity and quality of generated images by leveraging the 1000 class probabilities of the generated images. While IS does not require the original real images, KID and FID are computed by determining the distance between the distributions of real and generated images. KID utilizes the squared MMD distance with a polynomial kernel. FID employs the squared Fréchet distance between two probability distributions, which is also equal to the Wasserstein-2 distance, with the assumption that both distributions are multivariate normal. Both FID and KID suffer from the limitations of the underlying Inception embeddings: they were trained on only 1 million images, limited to 1000 classes. Intuitively,



FD	0.0	0.0	0.0	0.0	0.0	0.0	0.0
FD $_{\infty}$	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MMD	0.0	0.5875	5.794	17.21	78.88	202.8	244.9

Table 2. Behavior of estimated Fréchet distances and MMD when normality assumption is violated. Going from left to right, the probability distribution changes more and more from the leftmost distribution. However, the Fréchet distances to the leftmost distribution calculated with normality assumption remains misleadingly zero. MMD, on the other hand, is able to correctly capture the progressive departure.

we expect this to limit their ability to represent the rich and complex image content seen in modern generated images.

Previous work pointed to unreliability of evaluation metrics in image generation [5, 17]. Chong et al. [5] show that FID is a biased estimator and that the bias depends on the model being evaluated. They propose an extrapolation approach to compute a bias-free estimator: FID $_{\infty}$ . Parmar et al. [17] show that low-level image processing operations such as compression and resizing can lead to significant variations in FID, and advocate the use of anti-aliased resizing operations. In this work, we show that FID’s issues extend well beyond what is discussed in prior works and that FID $_{\infty}$  and/or anti-aliased resizing do not solve them.

### 3. Limitations of FID

In this section we highlight some key limitations of FID. We start with a background discussion of the metric in order to better understand its limitations. Fréchet Inception Distance (FID) is used to measure the discrepancy between two image sets:  $\mathcal{I}$  and  $\mathcal{I}'$ . Usually one set of images are real (for example, from the COCO dataset) and the other set is generated using the image generation model to be evaluated. To calculate FID, Inception-v3<sup>1</sup> embeddings [23] are first extracted for both image sets using the Inception-v3 model trained on the ImageNet classification task. The FID between  $\mathcal{I}$  and  $\mathcal{I}'$  is then defined as the Fréchet distance between these two sets of Inception embeddings.

#### 3.1. The Fréchet Distance

For any two probability distributions  $P$  and  $Q$  over  $\mathbb{R}^d$  having finite first and second moments, the Fréchet distance is defined by [6, 15]:

$$\text{dist}_F^2(P, Q) := \inf_{\gamma \in \Gamma(P, Q)} \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \gamma} \|\mathbf{x} - \mathbf{y}\|^2, \quad (1)$$

<sup>1</sup>Throughout the paper we use the terms Inception and Inception-v3 interchangeably.

where  $\Gamma(P, Q)$  is the set of all couplings of  $P$  and  $Q$ . This is also equivalent to the Wasserstein-2 distance on  $\mathbb{R}^d$ . In general, obtaining a closed-form solution for the Fréchet distance is difficult. However, the authors of [6] showed that a closed-form solution exists for multivariate normal distributions in the form:

$$\text{dist}_F^2(P, Q) = \|\boldsymbol{\mu}_P - \boldsymbol{\mu}_Q\|_2^2 + \text{Tr}(\boldsymbol{\Sigma}_P + \boldsymbol{\Sigma}_Q - 2(\boldsymbol{\Sigma}_P \boldsymbol{\Sigma}_Q)^{\frac{1}{2}}), \quad (2)$$

where  $\boldsymbol{\mu}_P, \boldsymbol{\mu}_Q$  are the means and  $\boldsymbol{\Sigma}_P, \boldsymbol{\Sigma}_Q$  are the covariances of the two multivariate normal distributions  $P$  and  $Q$ . Note that this simplified formula is strictly valid only when both  $P$  and  $Q$  are multivariate normal distributions [6].

For FID, we need to estimate the Fréchet distance between two distributions of Inception embeddings, using two corresponding samples. This is challenging due to the high dimensionality of inception embeddings,  $d = 2048$ . Assuming that the Inception embeddings are drawn from a normal distribution simplifies the problem, allowing us to use Eq. (2) with  $\boldsymbol{\mu}_P, \boldsymbol{\mu}_Q$  and  $\boldsymbol{\Sigma}_P, \boldsymbol{\Sigma}_Q$  estimated from the two samples  $\mathcal{I}$  and  $\mathcal{I}'$ . There are two kinds of error in this procedure:

1. As we show in Section 3.3, Inception embeddings for typical image sets are far from being normally distributed. The implications of this inaccurate assumption when calculating the Fréchet distance are discussed in Section 3.2.
2. Estimating  $(2048 \times 2048)$ -dimensional covariance matrices from a small sample can lead to large errors, as discussed in Section 6.3.

#### 3.2. Implications of Wrong Normality Assumptions

When calculating the Fréchet distance between two distributions, making an incorrect normality assumption can lead to disastrous results. We illustrate this using a 2D isotropic Gaussian distribution at the origin as the reference distribution and by measuring the distance between that and a series of mixture-of-Gaussian distributions generated as described below. The results are summarized in Table 2.

To generate the series of second distributions, we start with a mixture of four Gaussians, each having the same mean and covariance as the reference Gaussian. Since this mixture has the same distribution as the reference distribution, we expect any reasonable distance to measure zero distance between this and the reference distribution (first column of Table 2). We then let the second distribution’s four components get further and further away from each other while keeping the overall mean and the covariance fixed (first row of Table 2). When this happens the second distribution obviously gets further and further away from the reference distribution. However, the Fréchet distance calculated with the normality assumption (note that this is *not* the true Fréchet distance, which cannot be easily calculated) remains misleadingly zero. This happens because the second distribution is normal only at the start, therefore the normality assumption is reasonable only for the first column of the table. Since the second distribution is not normal after that, the Fréchet distance calculated with normality assumption gives completely incorrect results. Note that, as shown in the third row of Table 2,  $FID_{\infty}$ , the unbiased version of FID proposed in [5], also suffers from this shortcoming, since it also relies on the normality assumption. In contrast, the MMD distance described in Section 4 (bottom row of Table 2) is able to capture the progressive departure of the second distribution from the reference distribution. More details of the experiment setup are in Appendix A.

### 3.3. Incorrectness of the Normality Assumption

When estimating the Fréchet distance, it is assumed that the Inception embeddings for each image set (real and generated), come from a multivariate normal distribution. In this section, we show that this assumption is wrong. As discussed in Section 3.2, making a wrong normality assumption about the underlying distribution can lead to completely wrong results.

It should not be surprising that Inception embeddings for a typical image set do not have a multivariate normal distribution with a single mode. Inception embeddings are activations extracted from the penultimate layer of the Inception-v3 network. During training, these activations are classified into one of 1000 classes using a *linear* classifier (the last fully-connected layer of the Inception-v3 network). Therefore, since the Inception-v3 network obtains good classification results on the ImageNet classification task, one would expect Inception embeddings to have at least 1,000 clusters or modes. If this is the case, they cannot be normally distributed.

Figure 2 shows a 2-dimensional t-SNE [24] visualization of Inception embeddings of the COCO 30K dataset, commonly used as the reference (real) image set in text-to-image FID benchmarks. It is clear that the low dimensional visualization has multiple modes, and therefore, it is also

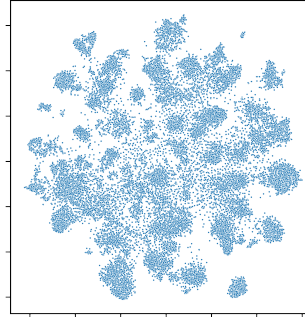


Figure 2. *t-SNE visualization of Inception embeddings of the COCO 30K dataset. Even in the reduced-dimensional 2D representation, it is easy to identify that embeddings have multiple modes and do not follow a multivariate normal distribution.*

clear that the original, 2048-dimensional distribution is not close to a multivariate normal distribution.

Finally, we applied three different widely-accepted statistical tests: Mardia’s skewness test, Mardia’s kurtosis test, and Henze-Zirkler test to test normality of Inception embeddings of the COCO 30K dataset. All of them *strongly* refute the hypothesis that Inception embeddings come from a multivariate normal distribution, with  $p$ -values of virtually zero (indicating an overwhelming confidence in rejecting the null hypothesis of normality). The details of these tests can be found in Appendix B.

To be clear, we do not expect CLIP embeddings to be normally distributed either. It is FID’s application of Fréchet distance with its normality assumption to non-normal Inception features, that we object to. In fact, CLIP embeddings of COCO 30K also fail the normality tests with virtually zero  $p$ -values, indicating that it is not reasonable to assume normality on CLIP embeddings either.

## 4. The CMMD Metric

In this section, we propose a new metric to evaluate image generation models, using CLIP embeddings and the Maximum Mean Discrepancy (MMD) distance, with a Gaussian RBF kernel. The CMMD (stands for CLIP-MMD) metric is the squared MMD distance between CLIP embeddings of the reference (real) image set and the generated image set.

CLIP embeddings [18] have changed the way we think about image and text representations by learning them in a joint space. CLIP trains an image encoder and a text encoder jointly using 400 million image-text pairs containing complex scenes. In contrast, Inception-v3 is trained on ImageNet, which has on the order of 1 million images which are limited to 1000-classes and only one prominent object per image. As a result, CLIP embeddings are better suited for representing the diverse and complex content we see in images generated by modern image generation algorithms and the virtually infinite variety of prompts given to text-to-image models.

To compute the distance between two distributions we use the MMD distance [10, 11]. MMD was originally de-

veloped as a part of a two-sample statistical test to determine whether two samples come from the same distribution. The MMD statistic calculated in this test can also be used to measure the discrepancy between two distributions. For two probability distributions  $P$  and  $Q$  over  $\mathbb{R}^d$ , the MMD distance with respect to a positive definite kernel  $k$  is defined by:

$$\text{dist}_{\text{MMD}}^2(P, Q) := \mathbb{E}_{\mathbf{x}, \mathbf{x}'}[k(\mathbf{x}, \mathbf{x}')] + \mathbb{E}_{\mathbf{y}, \mathbf{y}'}[k(\mathbf{y}, \mathbf{y}')] - 2 \mathbb{E}_{\mathbf{x}, \mathbf{y}}[k(\mathbf{x}, \mathbf{y})], \quad (3)$$

where  $\mathbf{x}$  and  $\mathbf{x}'$  are independently distributed by  $P$  and  $\mathbf{y}$  and  $\mathbf{y}'$  are independently distributed by  $Q$ . It is known that the MMD is a metric for characteristic kernels  $k$  [8, 11].

Given two sets of vectors,  $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$  and  $Y = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n\}$ , sampled from  $P$  and  $Q$ , respectively, an unbiased estimator for  $d_{\text{MMD}}^2(P, Q)$  is given by,

$$\begin{aligned} \hat{\text{dist}}_{\text{MMD}}^2(X, Y) &= \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j \neq i}^m k(\mathbf{x}_i, \mathbf{x}_j) \\ &+ \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i}^n k(\mathbf{y}_i, \mathbf{y}_j) - \frac{2}{mn} \sum_{i=1}^m \sum_{j=1}^n k(\mathbf{x}_i, \mathbf{y}_j). \end{aligned} \quad (4)$$

Some advantages of MMD over the Fréchet distance are:

1. MMD metric, when used with a characteristic kernel [8], is *distribution-free*. That is, it does not make any assumptions about the distributions  $P$  and  $Q$ . In contrast, the Fréchet distance in Eq. (2) assumes normality and is liable to give erroneous results when this assumption is violated.
2. As shown in [5], the FID estimated from a finite sample has a bias that depends on the model being evaluated, to the extent that the sample size can lead to different rankings of the models being evaluated. Removing this bias requires a computationally expensive procedure involving computation of multiple FID estimates [5]. In contrast, the MMD estimator in Eq. (4), is *unbiased*.
3. When working with high-dimensional vectors such as image embeddings, MMD is *sample efficient*. Fréchet distance, on the other hand, requires a large sample to reliably estimate the  $d \times d$  covariance matrix. This will be further elaborated on in Section 6.3.

As the kernel in the MMD calculation, we use the Gaussian RBF kernel  $k(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2/2\sigma^2)$ , which is a characteristic kernel. Details about this choice are in Appendix C. We set the bandwidth parameter set to  $\sigma = 10$ . Empirically, we observed that the bandwidth does not significantly affect the overall trends of the metric. We therefore fixed it at 10 to obtain consistent values for the metric. Since the MMD metric with the Gaussian kernel is bounded above at 2 (when the two distributions are maximally different), it gives small values for general distributions. We

therefore scale up the value in Eq. (4) by 1000 to obtain more readable values. For the CLIP embedding model, we use the publicly-available ViT-L/14@336px model, which is the largest and the best performing CLIP model [18]. Also note that we have  $m = n$  in Eq. (4) for text-to-image evaluation since we evaluate generated images against real images sharing the same captions/prompts. Our code for computing CMMD has been made publicly available.

## 5. Human Evaluation

We now present a human evaluation to show that FID (and KID) do not agree with human perception of image quality. To this end, we picked two models, Model-A: the full Muse model as described in [3] with 24 base-model iterations and 8 super-resolution model iterations. Model-B: an early-stopped Muse model with only 20 base-model iterations and 3 super-resolution model iterations. This was done intentionally to reduce the quality of produced images. We use a Muse model trained on the WebLI dataset [4], generously made available to us by the Muse authors. The choice of early-stopping iterations is arbitrary: as shown in Figure 4, FID is consistently better (lower) for all early-stopped models when compared with the full model (Model-A).

We performed a side-by-side evaluation where human raters were presented with two images, one generated from Model-A and the other generated from Model-B. We used the same random seeds to ensure that image content and degree of alignment to the prompt are the same. This allowed the raters to focus on image quality. The raters were asked to evaluate which image looked better. Raters had the option of choosing either image or that they are indifferent. All image pairs were rated by 3 independent raters, hired through a high-quality crowd computing platform. The raters were not privy to the details of the image sets and rated images purely based on the visual quality. The authors and the raters were anonymous to each other.

We used all PartiPrompts [26], which is a collection of 1633 prompts designed for text-to-image model evaluation. These prompts cover a wide range of categories (abstract, vehicles, illustrations, art, world knowledge, animals, outdoor scenes, etc.) and challenge levels (basic, complex fine-

Model	Model-A	Model-B
FID	21.40	18.42
FID <sub>∞</sub>	20.16	17.19
KID	0.0105	0.0080
CMMD	0.721	0.951
Human rater preference	92.5%	6.9%

Table 3. Human evaluation. FID and KID contradict human evaluation while CMMD agrees. Lower is better for all metrics.



Figure 3. The quality of the generated image monotonically improves as we progress through Muse’s refinement iterations. CMMD correctly identifies the improvements. FID, however, incorrectly indicates a quality degradation (see Figure 4). Prompt: “The Parthenon”.

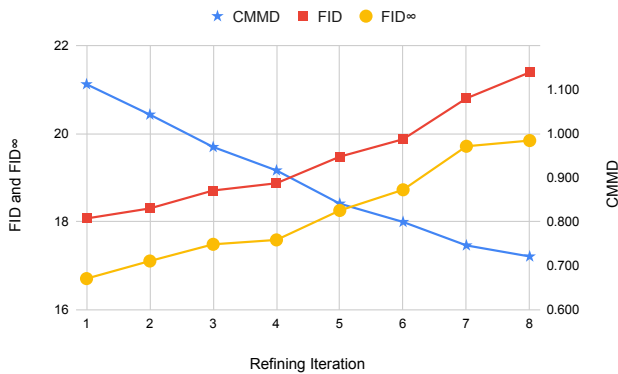


Figure 4. Behavior of FID and CMMD for Muse steps. CMMD monotonically goes down, correctly identifying the iterative improvements made to the images (see Figure 3). FID is completely wrong suggesting degradation in image quality as iterations progress. FID $\infty$  has the same behavior as FID.

grained detail, imagination, etc.). Evaluation results are summarized in Table 3. For each comparison, we consider a model as the winner if 2 or more raters have preferred the image produced by that model. If there is no consensus among the raters or if the majority of the raters selected are indifferent, no model wins. We observed that Model-A was preferred in 92.5% of the comparisons, while Model-B was preferred only 6.9% of the time. The raters were indifferent 0.6% of the time. The Cohen’s kappa coefficient for rater agreement was 0.9937, indicating a very high agreement rate. It is therefore clear that human raters overwhelmingly prefer Model-A to Model-B. However, COCO 30K FID, its unbiased variant FID $\infty$ , and KID unfortunately say otherwise. On the other hand, the proposed CMMD metric correctly aligns with the human preference.

## 6. Performance Comparison

We now compare FID with the proposed CMMD metric under various settings to point out the limitations of FID

while highlighting the benefits of CMMD. In all our experiments, we use the COCO 30K dataset [14] as the reference (real) image dataset. Zero-shot evaluation on this dataset is currently the de facto evaluation standard for text-to-image generation models [3, 20, 21]. Throughout our experiments, where applicable, we use high-quality bicubic resizing with anti-aliasing as suggested in [17]. This prevents any adverse effects of improperly-implemented low level image processing operations on FID as those reported in [17].

For Stable Diffusion [20], we use the publicly available Stable Diffusion 1.4 model. We evaluate all models without any additional bells and whistles such as CLIP sorting.

### 6.1. Progressive Image Generation Models

Most modern text-to-image generation models are iterative. For example, diffusion models [20, 21] require multiple denoising steps to generate the final image, the Parti model [26] auto-regressively generates image tokens one at a time. While the Muse model [3] generates multiple tokens at a time, it still requires iterative sampling steps to generate the final image, as shown in Figure 3. Gradually improving the quality of the generated images in each step, these methods go from poor quality images or pure noise images to unprecedented photo-realism. This progression in quality is obvious to a human observer and we would expect any reasonable metric to monotonically improve as we progress through iterations of image generation.

Figure 4 shows FID, FID $\infty$ , and CMMD values for progressive Muse iterations. FID and FID $\infty$  incorrectly suggest that the image quality degrades, when the quality improvements are obvious as illustrated in Figure 3. In contrast, CMMD correctly identifies the quality improvements made during Muse’s iterative refinements. As seen in Figure 4, we consistently observe in our experiments that FID and FID $\infty$  have the same behavior although absolute values are different. This is not surprising since FID $\infty$  is derived from FID and inherits many of its shortcomings.

Figure 6 shows an evaluation of the last 5 iterations



Figure 5. Behavior of FID and CMMD under distortions. Images in the first row (FID: 21.40, CMMD: 0.721) are undistorted. Images in the second (FID: 18.02, CMMD: 1.190) are distorted by randomly replacing each VQGAN token with probability  $p = 0.2$ . The image quality clearly degrades as a result of the distortion, but FID suggests otherwise, while CMMD correctly identifies the degradation.

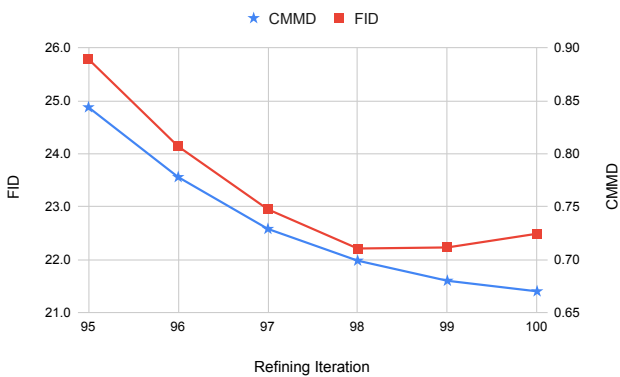


Figure 6. Behavior of FID and CMMD for StableDiffusion steps. CMMD monotonically improves (goes down), reflecting the improvements in the images. FID’s behavior is not consistent, it mistakenly suggests a decrease in quality in the last two iterations.

of a 100-iteration Stable Diffusion model. Our proposed CMMD metric monotonically improves (decreases) with the progression of the iterations, whereas FID has unexpected behavior. We focus on the more subtle differences in the final iterations of Stable Diffusion, since both FID and CMMD showed monotonicity at the easily-detectable high noise levels in the initial iterations.

## 6.2. Image Distortions

Here, we provide additional evidence that FID does not accurately reflect image quality under complex image distortions. It was shown in [12] that FID accurately captures image distortions under low-level image processing distortions such as Gaussian noise and Gaussian blur. Since Inception embeddings are trained on ImageNet images without extreme data augmentation, it is not surprising that FID

is able to identify these distortion. However, in this section, we show that FID is unable to identify more complex noise added in the latent space.

To this end, we take a set of images generated by Muse and progressively distort them by adding noise in the VQGAN latent space [7]. For each image, we obtain VQGAN tokens, replace them with random tokens with probability  $p$ , and reconstruct the image with the VQGAN detokenizer. Example distortions are shown in Figure 5. The images get more and more distorted with increasing  $p$  and the quality loss with increasing  $p$  is visibly obvious. However, as shown in Figure 7, FID fails to reflect the degradation in image quality for increasing values of  $p$ . Our CMMD metric, on the other hand, monotonically worsens (increases) with the distortion level  $p$ , correctly identifying the quality regression. We replicated the same results with images generated by a Stable Diffusion as well. Furthermore, Figure 1 shows that FID behaves poorly also when we measure the distances between progressively distorted versions (using the same procedure) of the COCO 30K dataset and the reference clean version of that dataset.

Figure A in the appendix shows how CMMD increases with increasing levels of image distortions such as blur, Gaussian noise and JPEG compression artifacts.

## 6.3. Sample Efficiency

As stated in Section 4, calculating FID requires estimating a  $2048 \times 2048$  covariance matrix with 4 million entries. This requires a large number of images causing FID to have poor sample efficiency. This has also been noted by the authors of [5]. The proposed CMMD metric does not suffer from this problem thanks to its usage of MMD distance instead of the Fréchet distance.

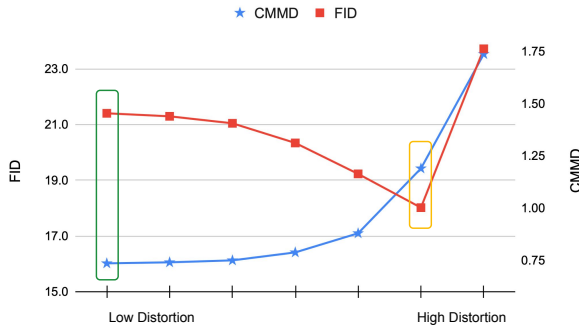


Figure 7. Behavior of FID and CMMD under latent space noise added to generated images. CMMD monotonically goes up, reflecting the quality degradation of the images. FID’s behavior is inconsistent, it mistakenly suggests an increase of quality. Image sets highlighted in green and yellow are visualized in Figure 5’s top and bottom rows, respectively.

In Figure 8 we illustrate this by evaluating a Stable Diffusion model at different sample sizes (number of images) sampled randomly from the COCO 30K dataset. Note that we need more than 20,000 images to reliably estimate FID, whereas CMMD provides consistent estimates even with small image sets. This has important practical implications: development of image generation models requires fast on-line evaluation, e.g. as a metric tracked during training. Another relevant scenario is comparing a large number of models. Since reliable estimation of FID requires generating a large number of images, FID evaluation is costly and time consuming. In contrast, CMMD can be evaluated fast by generating only a small number of images. CMMD evaluation is faster than FID evaluation for two reasons: 1) it requires only a small number of images to be generated. 2) once the images are generated the computation of CMMD is faster than the FID computation as discussed in the next section.

#### 6.4. Computational Cost

Let  $n$  be the number of images, and let  $d$  be the embedding length. The cost of computing the Fréchet distance (FD) is dominated by the matrix square root operation on a  $d \times d$  matrix, which is expensive and not easily parallelizable. The cost of computing the unbiased version  $FD_\infty$  is even higher, since it requires computing FD multiple times with different sample sizes. The asymptotic complexity of computing MMD is  $O(n^2d)$ . However, in practice, MMD can be computed very efficiently, since it only involves matrix multiplications which are trivially parallelizable and highly optimized in any deep learning library such as Tensorflow, PyTorch, and JAX.

Table 4 shows an empirical runtime comparison of computing FD and MMD on a set of size  $n = 30,000$  with  $d = 2048$  dimensional features on a TPUv4 platform with a JAX implementation. For FD calculations, we use our JAX

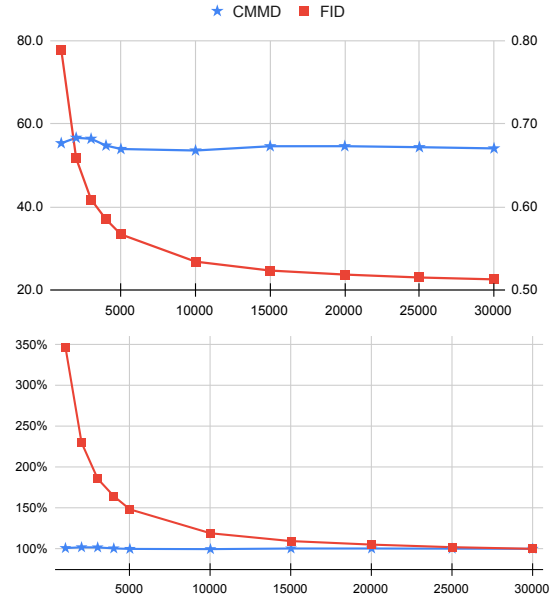


Figure 8. Behavior of FID and CMMD under different sample sizes. Top: absolute values of the metrics. Bottom: Values relative to the value at 30k sample size.

Operation	Time
Fréchet distance	$7007.59 \pm 231$ ms
MMD distance	$71.42 \pm 0.67$ ms
Inception model inference	$2.076 \pm 0.15$ ms
CLIP model inference	$1.955 \pm 0.14$ ms

Table 4. Comparing runtime for computing Fréchet/MMD distances and Inception/CLIP feature extractions.

implementation and publicly available PyTorch/numpy implementations from [17] and [5] and report the best runtime. In the same table, we also report the runtime for Inception and CLIP feature extraction for a batch of 32 images.

## 7. Discussion

We encourage image generation researchers to rethink the use of FID as a primary evaluation metric for image quality. Our findings that FID correlates poorly with human raters, that it does not reflect gradual improvement of iterative text-to-image models and that it does not capture obvious distortions add to a growing body of criticism [5, 17]. We are concerned that reliance on FID could lead to flawed rankings among the image generation methods, and that good ideas could be rejected prematurely. To address these concerns we propose CMMD as a more robust metric, suitable for evaluation of modern text-to-image models.

## Acknowledgment

We would like to thank Wittawat Jitkrittum and the anonymous reviewers for the valuable feedback.



## References

- [1] Shane Barratt and Rishi Sharma. A note on the inception score, 2018. [2](#)
- [2] Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs, 2021. [2](#)
- [3] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. Muse: Text-to-image generation via masked generative transformers. *ICML*, 2023. [5](#), [6](#)
- [4] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model, 2022. [5](#)
- [5] Min Jin Chong and David A. Forsyth. Effectively unbiased FID and inception score and where to find them. *CoRR*, abs/1911.07023, 2019. [2](#), [3](#), [4](#), [5](#), [7](#), [8](#)
- [6] D.C Dowson and B.V Landau. The Fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455, 1982. [3](#)
- [7] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. [7](#)
- [8] Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Bharath K. Sriperumbudur. Characteristic kernels on groups and semigroups. In *NeurIPS*. Curran Associates, Inc., 2008. [5](#)
- [9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *NeurIPS*, 2014. [2](#)
- [10] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. In *NeurIPS*. MIT Press, 2006. [4](#)
- [11] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13(1):723–773, 2012. [4](#), [5](#)
- [12] 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, 2018. [1](#), [2](#), [7](#)
- [13] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *CoRR*, abs/1812.04948, 2018. [2](#)
- [14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, pages 740–755. Springer, 2014. [6](#)
- [15] Maurice Fréchet. Sur la distance de deux lois de probabilité. *Annales de l’ISUP*, 1957. [3](#)
- [16] Midjourney, 2022. <https://www.midjourney.com>. [1](#)
- [17] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On Aliased Resizing and Surprising Subtleties in GAN Evaluation. In *CVPR*, 2022. [3](#), [6](#), [8](#)
- [18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. [2](#), [4](#), [5](#)
- [19] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *preprint*, 2022. [1](#)
- [20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. [6](#)
- [21] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *preprint*, 2022. [arXiv:2205.11487]. [1](#), [6](#)
- [22] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *CoRR*, abs/1606.03498, 2016. [2](#)
- [23] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015. [2](#), [3](#)
- [24] Laurens van der Maaten and Geoffrey E. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008. [4](#)
- [25] Qiantong Xu, Gao Huang, Yang Yuan, Chuan Guo, Yu Sun, Felix Wu, and Kilian Q. Weinberger. An empirical study on evaluation metrics of generative adversarial networks. *CoRR*, abs/1806.07755, 2018. [2](#)
- [26] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjian Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. In *ICML*, 2022. [1](#), [5](#), [6](#)
- [27] Sharon Zhou, Mitchell L. Gordon, Ranjay Krishna, Austin Narcomey, Durim Morina, and Michael S. Bernstein. HYPE: human eye perceptual evaluation of generative models. *CoRR*, abs/1904.01121, 2019. [2](#)