Rethinking FID: Towards a Better Evaluation Metric for Image Generation

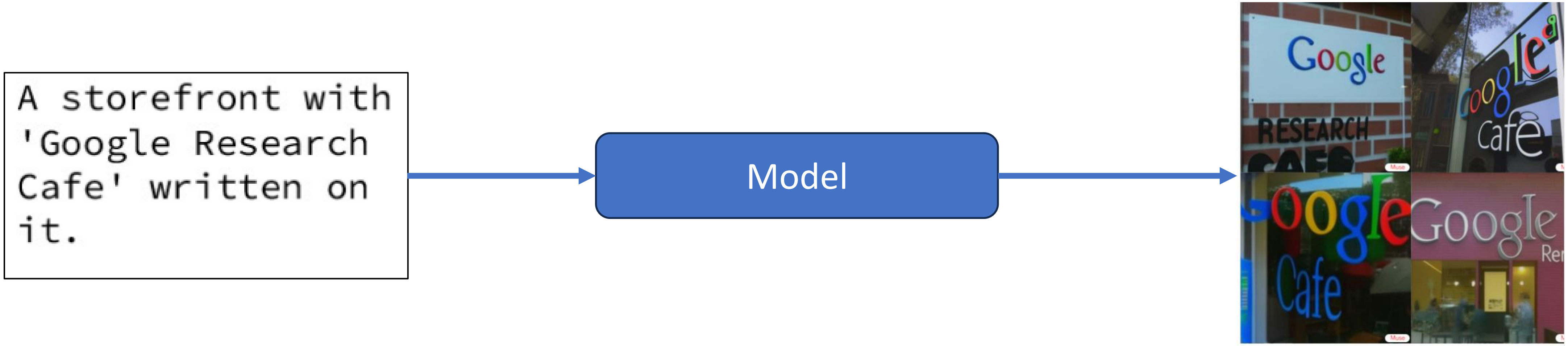
(30 Nov 2023, Google Research)

Denis Shepelev, Jan 2024
https://github.com/denzist/presentations

Plan

- Text-to-image task
- Common evaluation pipeline
- Frechet Inception Distance (FID)
- Solution: CMMD (CLIP embeddings and Maximum Mean Discrepancy distance)
- Metrics comparison
- Conclusion

Text-to-image task

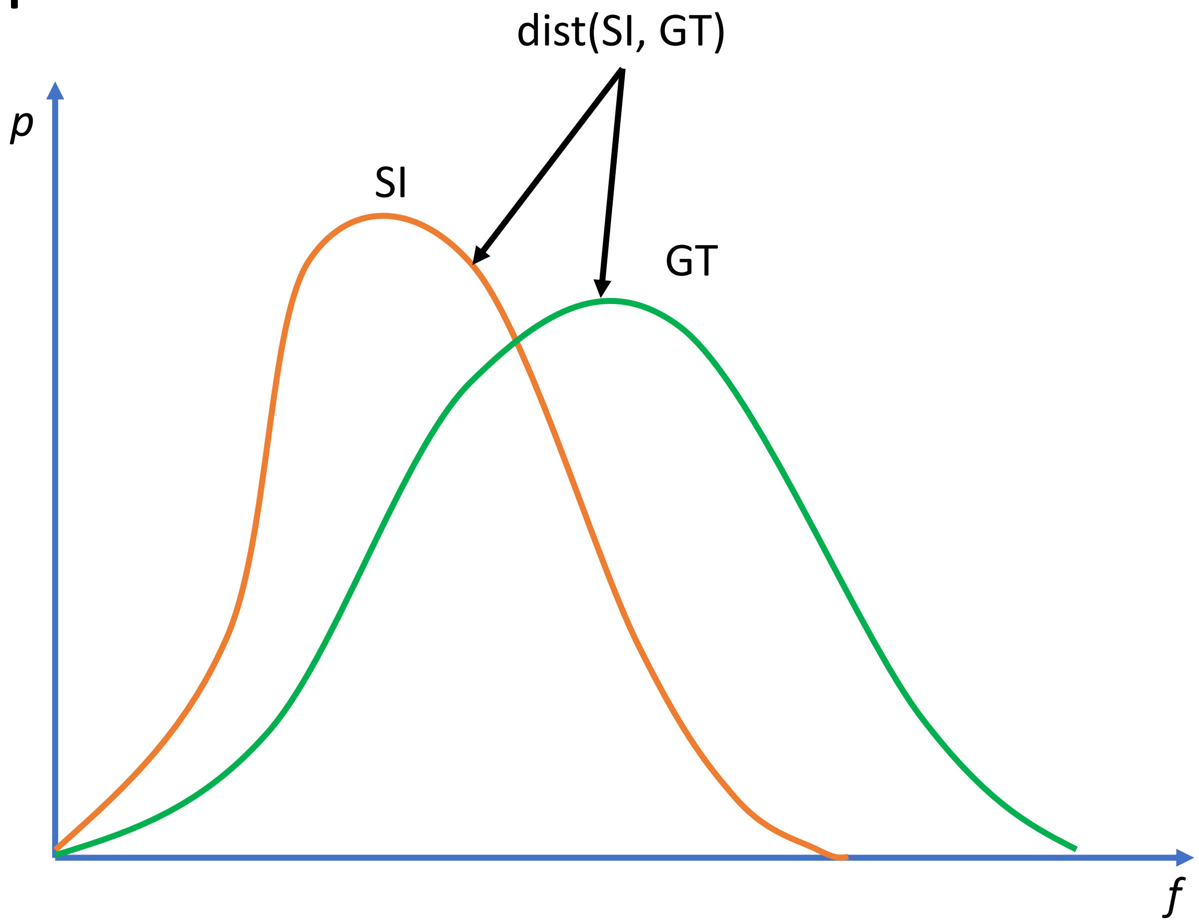


A storefront with 'Google Research Cafe' written on it.

Model

Common evaluation pipeline

- GT dataset of (text, images) pairs (COCO 30K)
- Given text prompts model a set of synthesized images (SI)
- Compare distributions of GT and generated sets (in features space)
- Frechet Inception Distance (FID) is usually used



Frechet Inception Distance (FID)

- Features are in space of Inception v3 model, which is trained for the ImageNet classification
- Features distributions *P* and *Q* are multivariate normal distributions

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

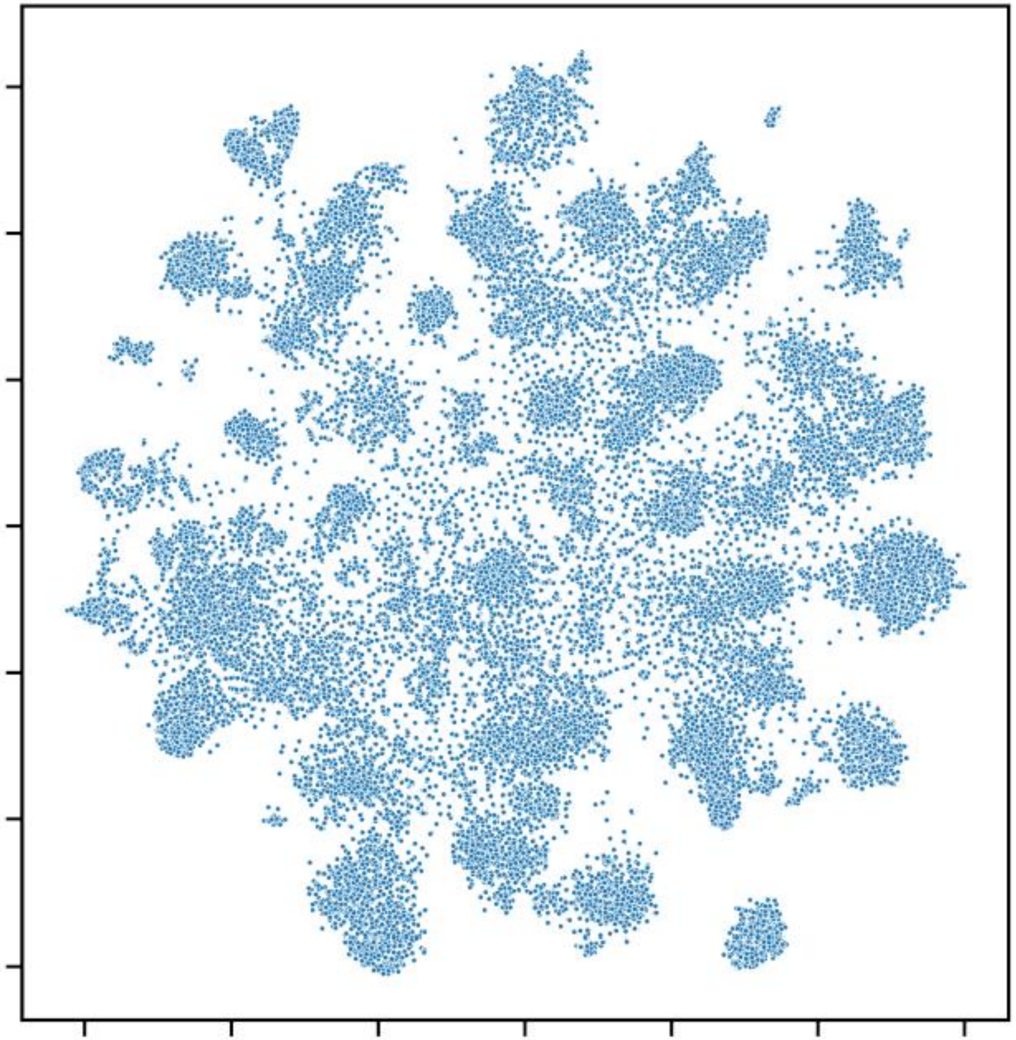
$$\mathrm{dist}_F^2(P,Q) = \|\boldsymbol{\mu}_P-\boldsymbol{\mu}_Q\|_2^2+\mathrm{Tr}(\boldsymbol{\Sigma}_P+\boldsymbol{\Sigma}_Q-2(\boldsymbol{\Sigma}_P\boldsymbol{\Sigma}_Q)^{\frac{1}{2}})$$

Problems:

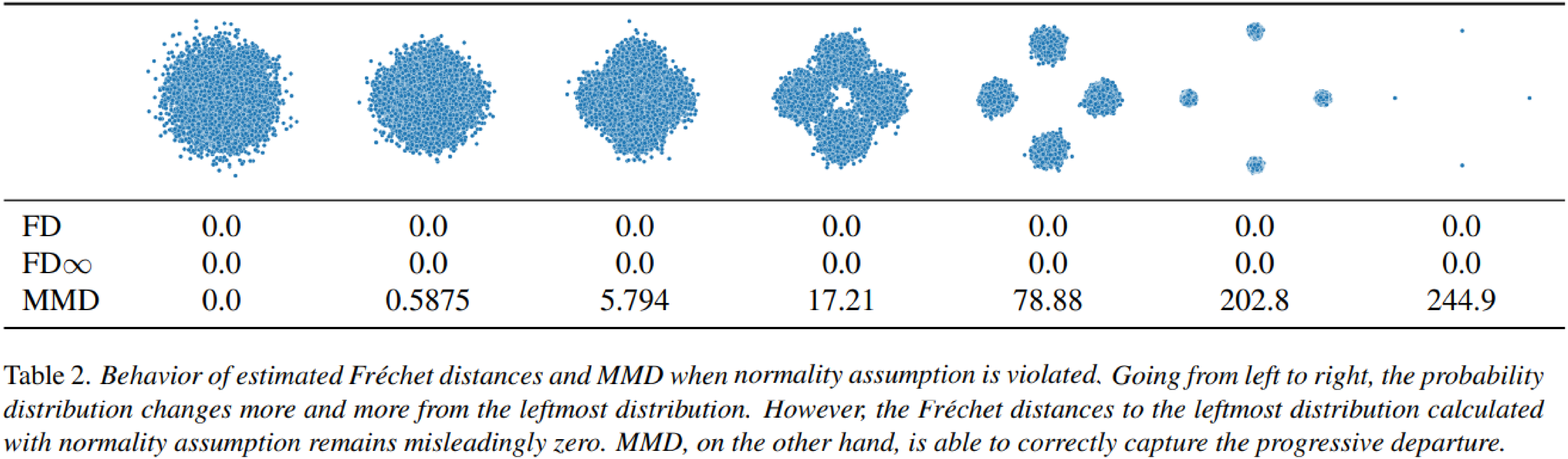
- Inception embeddings for typical image sets are not normally distributed
- FID is biased (see Unbiased FID)
- Need a lot of samples to estimate 2048x2048 covariance matrices

Inceptions embeddings distribution

- t-SNE visualization of Inception embeddings of the COCO 30K dataset. It is easy to identify that embeddings have multiple modes
- Multiple statistical tests strongly refute the hypothesis that Inception embeddings come from a multivariate normal distribution



FID and violation of normality assumption



FD	0.0	0.0	0.0	0.0	0.0	0.0	0.0
FD∞	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.*

CMMD distance

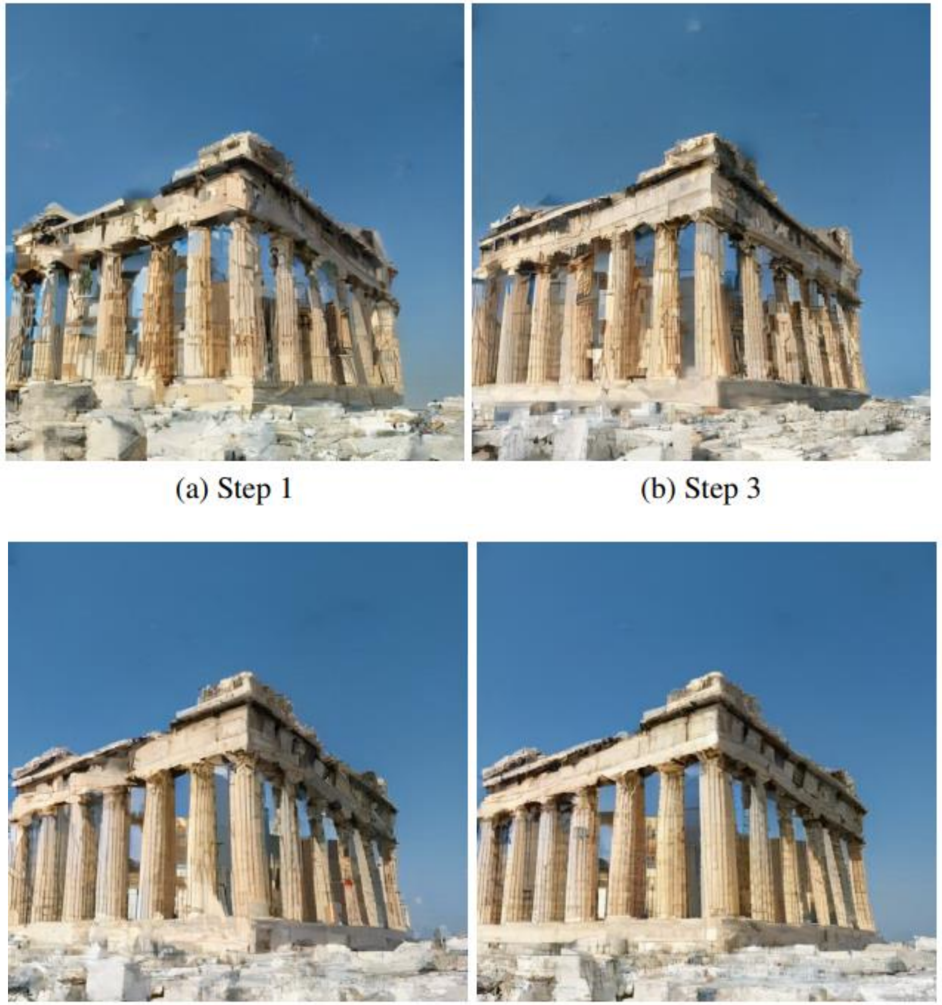
- CLIP embeddings instead of Inception v3
- Maximum Mean Discrepancy (MMD) distance, with a Gaussian RBF kernel – no assumption about distribution
- MMD estimator is unbiased
- When working with high-dimensional vectors such as image embeddings, MMD is sample efficient

$$\begin{aligned}\mathrm{dist}^2_{\mathrm{MMD}}(P, Q) := \mathbb{E}_{\mathbf{x},\mathbf{x}'}[k(\mathbf{x}, \mathbf{x}')] + \mathbb{E}_{\mathbf{y},\mathbf{y}'}[k(\mathbf{y}, \mathbf{y}')] \\ - 2\mathbb{E}_{\mathbf{x},\mathbf{y}}[k(\mathbf{x}, \mathbf{y})],\end{aligned}$$

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

$$k(\mathbf{x}, \mathbf{y}) = \exp(\|\mathbf{x} - \mathbf{y}\|^2 / \sigma^2)$$

Image quality vs number of model iterations (Muse)



(a) Step 1

(b) Step 3

(c) Step 6

(d) Step 8

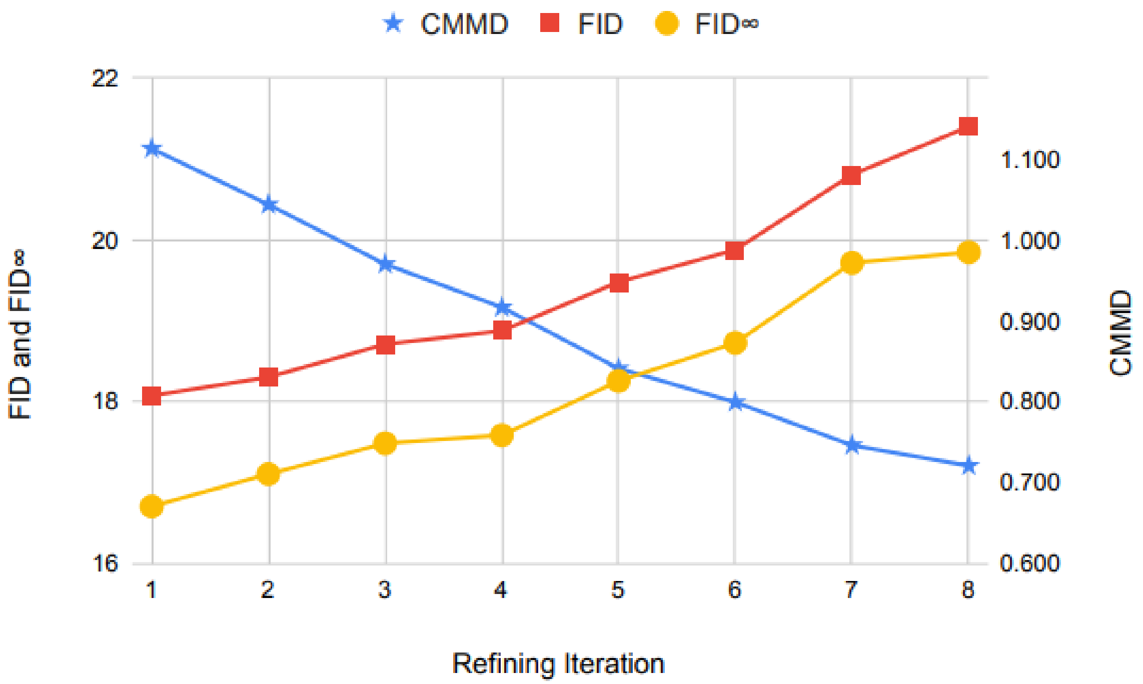


Figure 4. *Behavior of FID and CMMD for Muse steps. CMMD monotonically goes down, correctly identifying the iterative improvements made to the images (see Figure 3). FID is completely wrong suggesting degradation in image quality as iterations progress.* FID_{∞} *has the same behavior as FID.*

Image quality vs number of model iterations (SD1.4)

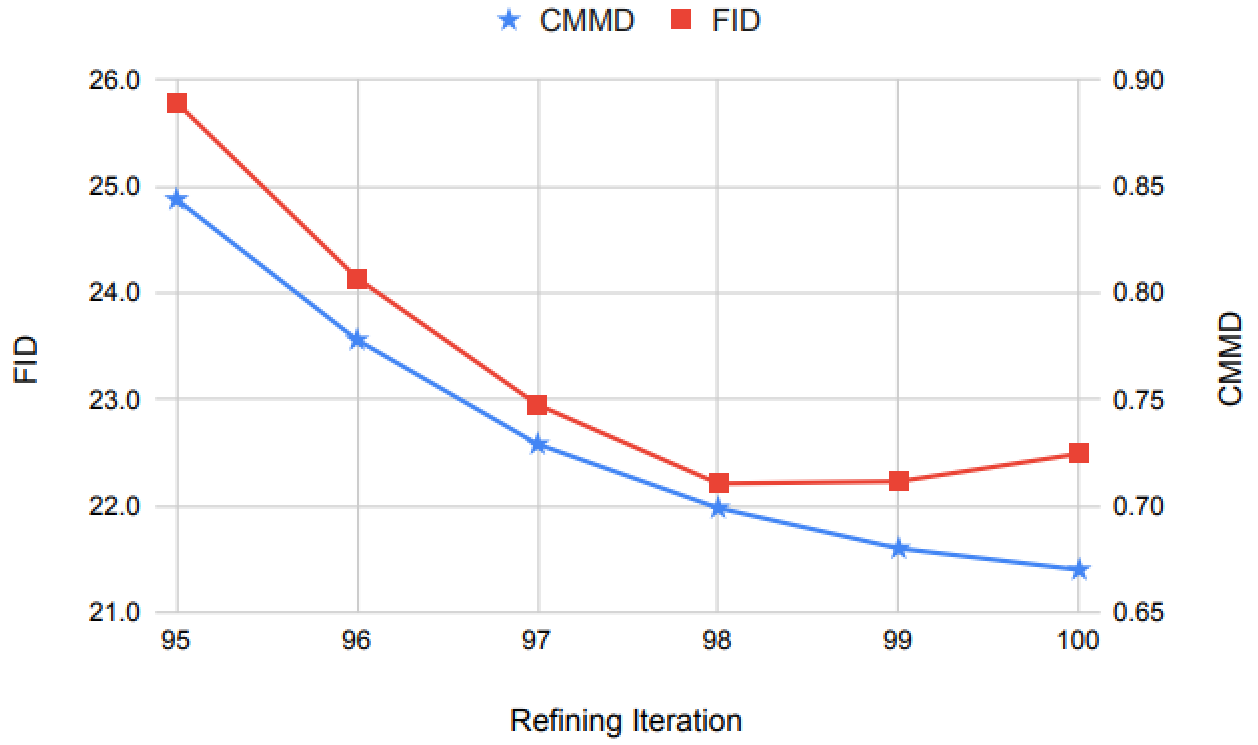
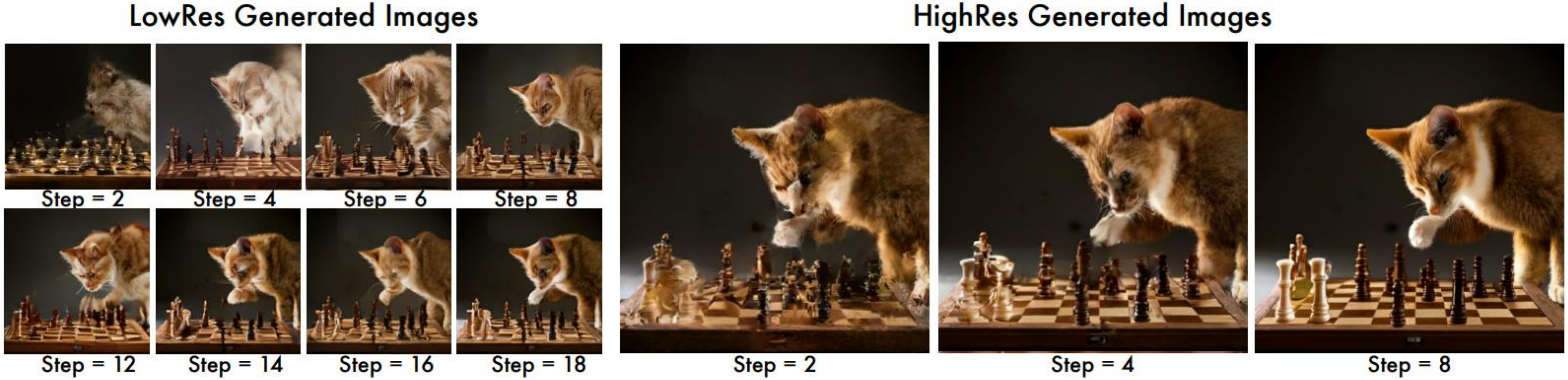


Figure 6. *Behavior of FID and CMMD for StableDiffusion steps. CMMD monotonically improves (goes down), reflecting the improvements in the images. FID's behavior is not consistent, it mistakenly suggests a decrease in quality in the last two iterations.*

Human evaluation

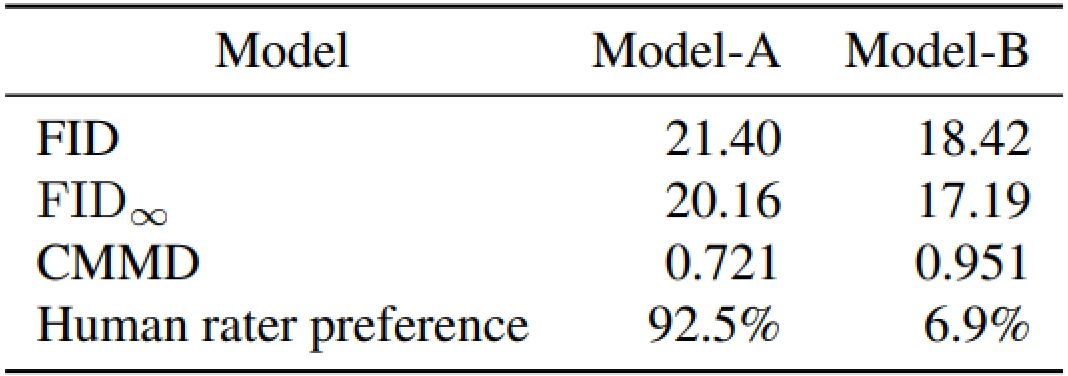


Model-A

- full Muse model
- 24 base-model iterations
- 8 super-resolution iterations

Model-B

- an early stopped Muse model
- 20 base-model iterations
- 3 super-resolution iterations



Model	Model-A	Model-B
FID	21.40	18.42
FID_∞	20.16	17.19
CMMD	0.721	0.951
Human rater preference	92.5%	6.9%

Table 3. *Human evaluation of different models. FID contradicts human evaluation while CMMD agrees.*

Randomized image degradation evaluation (1)

- Obtain VQGAN tokens for each image
- Replace part of the tokens with random ones with some probability *p*
- Reconstruct the image using VQGAN decoder

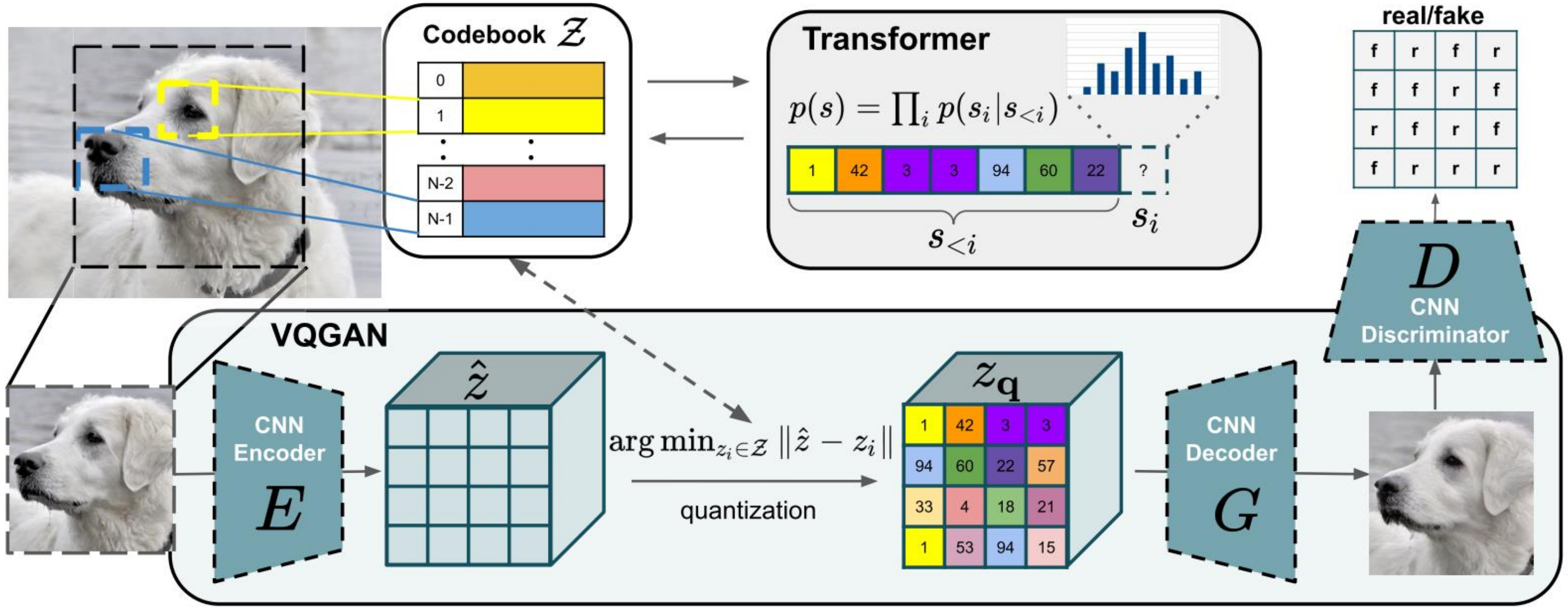


Figure 2. Our approach uses a convolutional *VQGAN* to learn a codebook of context-rich visual parts, whose composition is subsequently modeled with an autoregressive transformer architecture. A discrete codebook provides the interface between these architectures and a patch-based discriminator enables strong compression while retaining high perceptual quality. This method introduces the efficiency of convolutional approaches to transformer based high resolution image synthesis.

Randomized image degradation evaluation (2)

- The images get more and more distorted with increasing *p*
- We expect that quality loss should increase with *p*

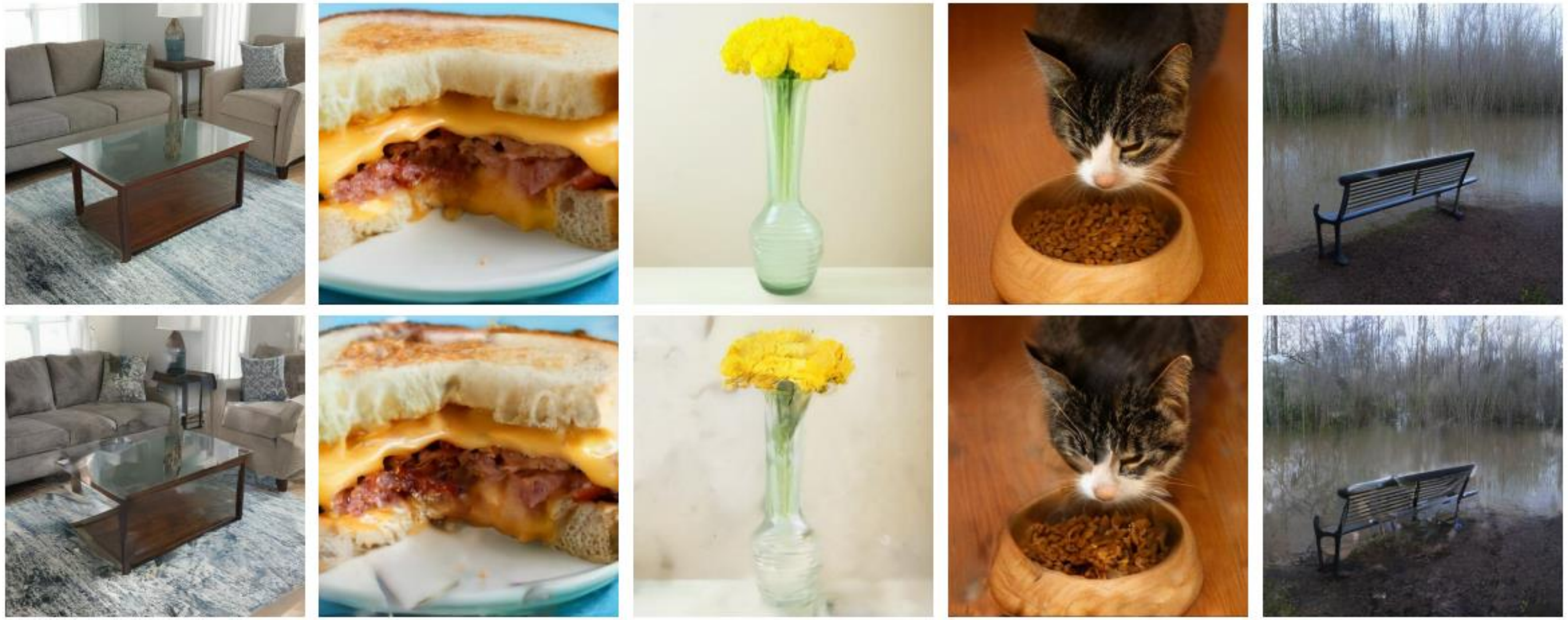


Figure 5. *Behavior of FID and CMMD under distortions. Images in the first row (FID: 21.40, CMMD: 0.721) are undistorted. Images in the second (FID: 18.02, CMMD: 1.190) are distorted by randomly replacing each VQGAN token with probability* $p = 0.2$. *The image quality clearly degrades as a result of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

Randomized image degradation evaluation (3)

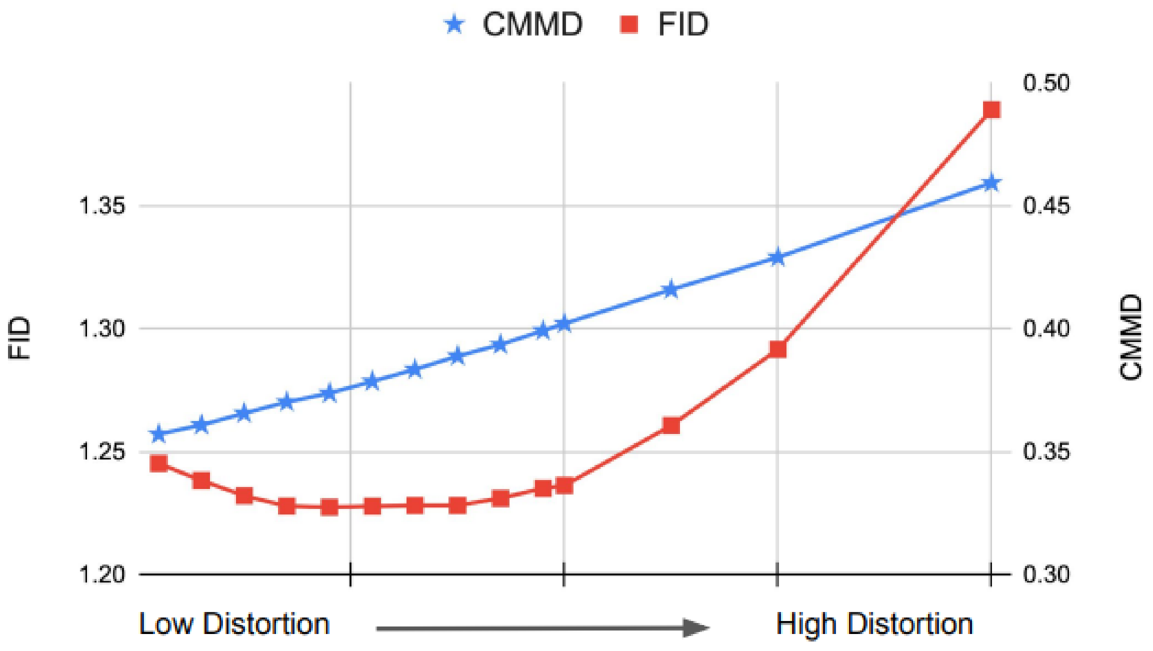


Figure 1. *Behaviour of FID and CMMD under distortions. CMMD monotonically increases with the distortion level, correctly identifying the degradation in image quality with increasing distortions. FID is wrong. It improves (goes down) for the first few distortion levels, suggesting that quality improves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

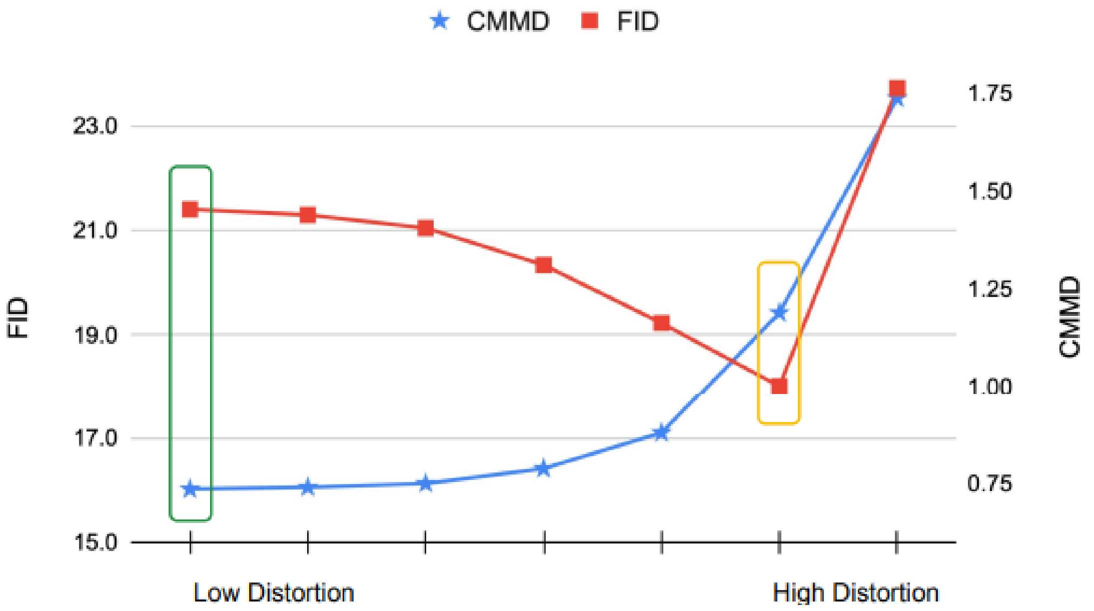
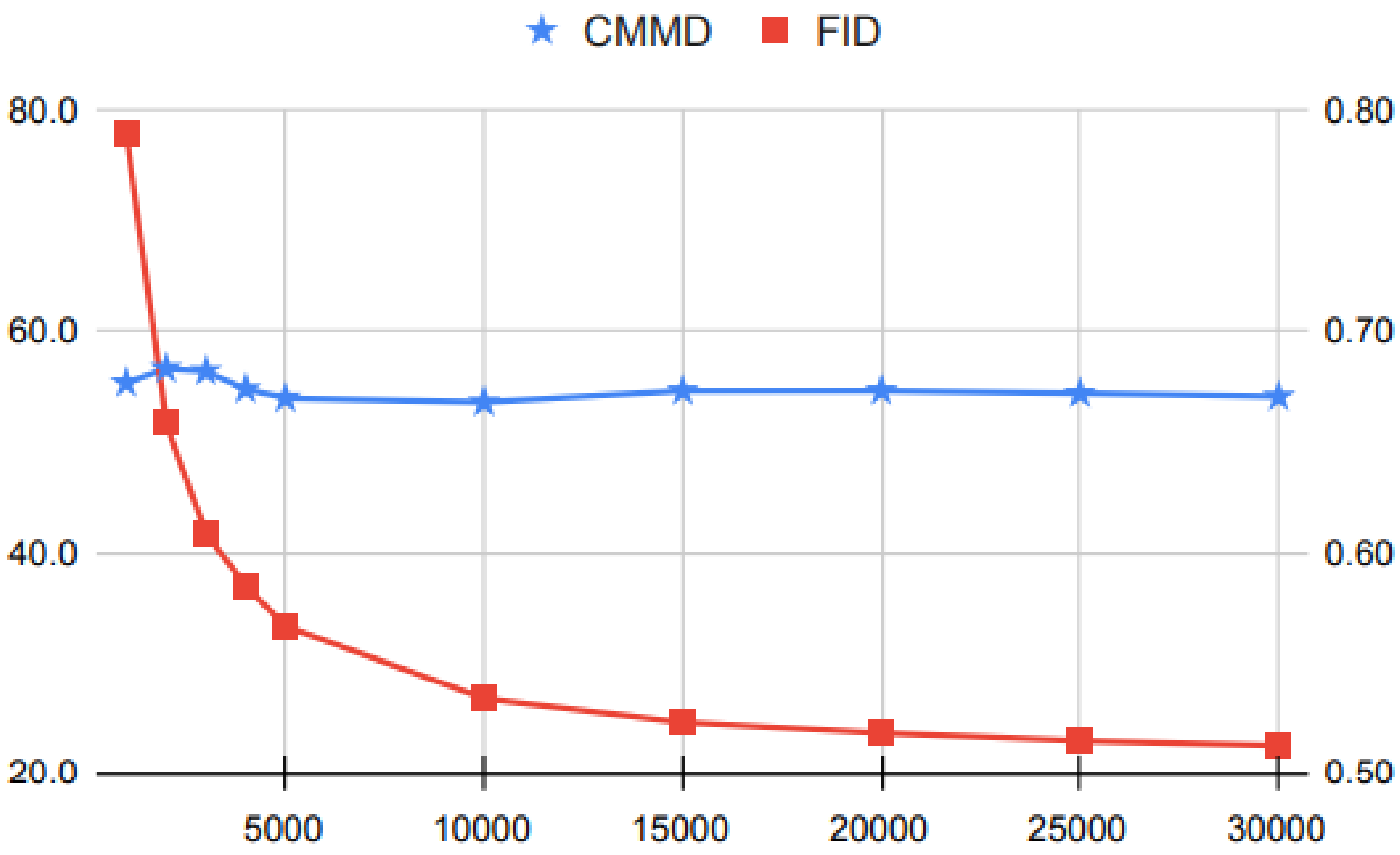


Figure 7. *Behavior of FID and CMMD under latent space noise added to generated images. CMMD monotonically goes up, reflecting the quality degradation of the images. FID's behavior is inconsistent, it mistakenly suggests an increase of quality. Image sets highlighted in green and yellow are visualized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

Sample efficiency

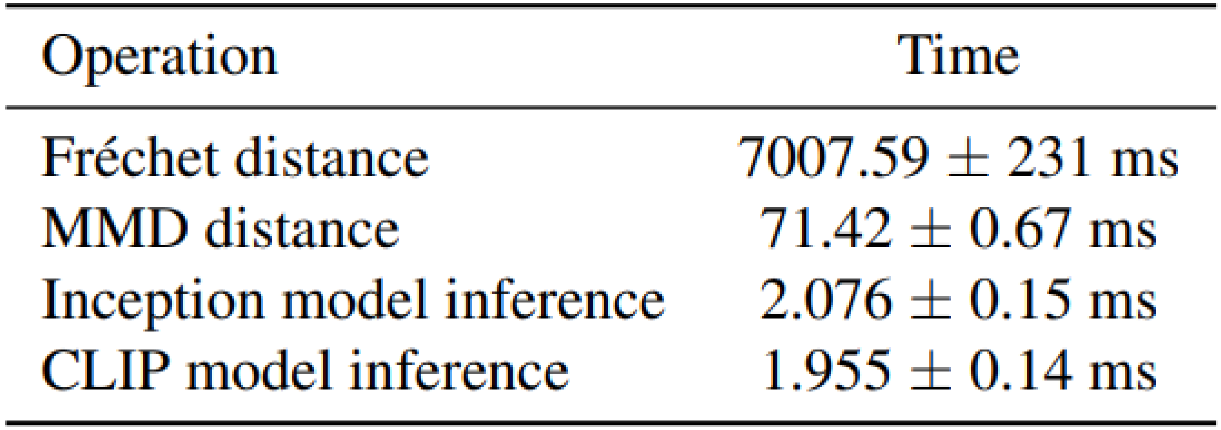
- FID evaluation is costly and time consuming: reliable estimation of FID requires more than 20,000 generated images
- In contrast, CMMD requires only a small number of images to be generated
- Practical implications: fast estimating of metric tracked during training or comparing a large number of models



Stable Diffusion evaluation at different sample sizes sampled randomly from the COCO 30K

Computational cost

- The cost of computing the Frechet distance (FD) is dominated by the matrix square root operation on a $d \times d$ matrix, which is expensive and not easily parallelizable
- Computing MMD is $O(n^2\ d)$, however, in practice, MMD can be computed very efficiently, since it only involves matrix multiplications which are trivially parallelizable and highly optimized
- 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
- In the same table, we also report the runtime for Inception and CLIP feature extraction for a batch of 32 images



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

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

Conclusions

CMMD is a new evaluation metric for text-to-image, build upon CLIP embeddings and Maximum Mean Discrepancy distance

- **Pros:**
  - Better aligned with human perception
  - For reliable computations CMMD requires a smaller number of images to be generated
  - Faster computations
  - As consequence, can be tracked during training or can be used to compare a large number of models
- **Coins:**
  - All experiments on Muse and SD1.4
  - No leaderboard of text-to-image models according to FID and CMMD
  - Limited image distortions
  - No ablations on CLIP+FID

Literature

- Paper
- Muse
- VQGAN
- Unbiased FID