The Fault in Our Metrics

Revisiting Generative Model Evaluation with Chamfer Distance

Chirag Vashist, Vaibhav Santhosh, Tristan Engst, Yanshu Zhang, Ke Li

NEURAL INFORMATION PROCESSING SYSTEMS

Why Evaluation Metrics Matter?

Models are trained on millions of images; we cannot check each image visually by eye. Hence, we use an automatic score.
As visual quality improves, evaluation metrics increasingly determine which methods are judged better, which trade-offs are visible, and which directions the field prioritizes.



Existing evaluation metrics are neither informative nor reliable

Aggregate Metrics

These metrics summarize the difference between the real and generated distributions as a **single symmetric score.**

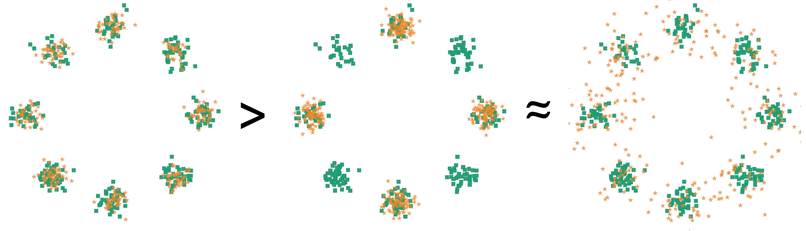
1. Fréchet Metrics with Gaussian assumption: FID and FD-DINOv2

$$\mathrm{FD}(P,Q) \;=\; \|\mu_P - \mu_Q\|_2^2 \;+\; \mathrm{Tr}\Big(\Sigma_P + \Sigma_Q - 2\,(\Sigma_P\Sigma_Q)^{1/2}\Big)$$

2. MMD: KID, CMMD

$$\mathrm{MMD}^2(P,Q) \;=\; \mathbb{E}_{x,x'\sim P}\big[k(x,x')\big] \;+\; \mathbb{E}_{y,y'\sim Q}\big[k(y,y')\big] \;-\; 2\,\mathbb{E}_{x\sim P,\,y\sim Q}\big[k(x,y)\big]$$

Issue: A generative model can be deficient along two distinct axes: fidelity (whether its samples look realistic), and diversity (whether it covers the full range of the real data). A single symmetric number conflates these two axes, making it hard to analyze and compare the performance of different generative models.



High Fidelity
High Diversity

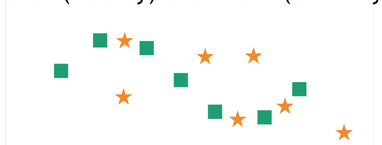
High Fidelity
Low Diversity

Low Fidelity
High Diversity

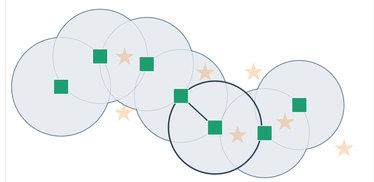
Disentangled Metrics

Disentangled metrics measure fidelity and diversity separately using two scores. Existing disentangled metrics rely on the KNN-ball isotropic method. The most famous are Precision (fidelity) and Recall (diversity).

Given: Two distributions P and Q



Step 1: Estimate the manifold using a union of isotropic balls. The radii of the balls are decided by within-set NN search.



Step 2: Use a binary membership test to calculate the metric. Here:

Precision = $\frac{4}{7} \approx 57\%$



Recall is computed by repeating the procedure in the opposite direction.

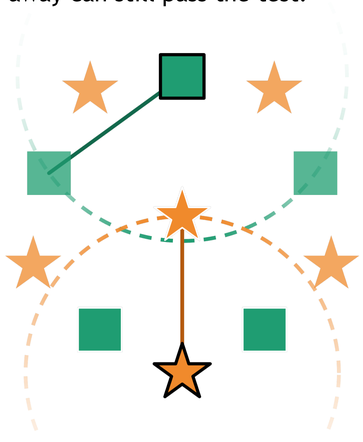
Issue: The estimated manifold can be large. Because these metrics rely on a binary test, samples that lie far away can still pass the test.

Consider this example on the right

- Precision = Recall = 1 (perfect)
- FID = 0 (perfect)
- Coverage, Density, Symmetric P/R: also perfect

The two distributions are clearly distinct, yet the standard metrics say they are identical.

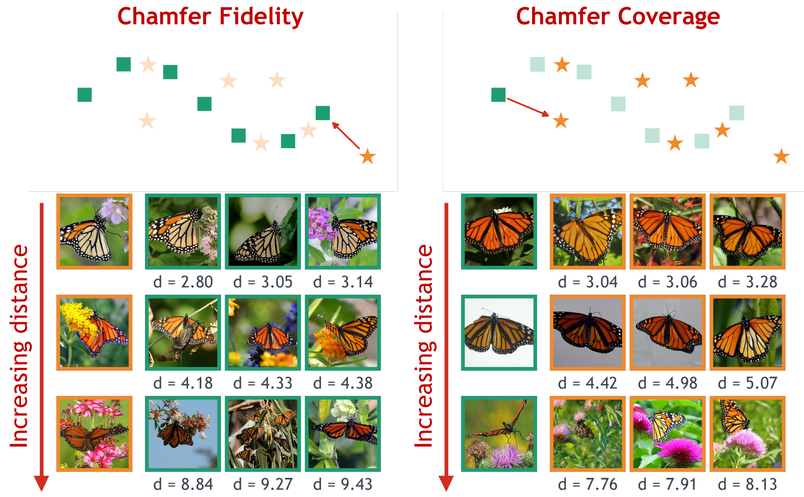
We need better metrics.



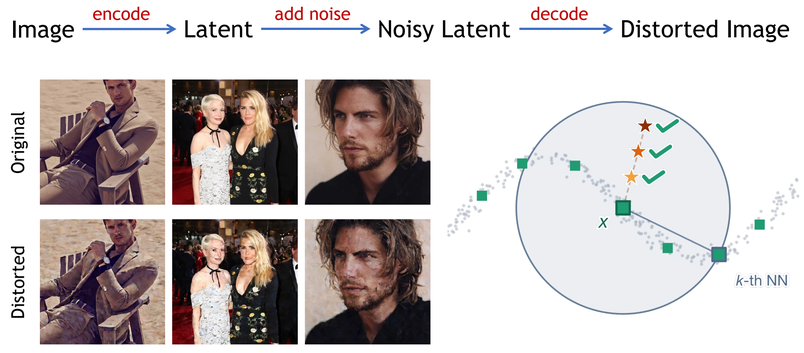
Disentanglement is right. Balls are the wrong tool.

Our Solution

Idea: For each sample, find its nearest neighbour distance in the other set. Metric is computed by averaging over all samples in the distribution.



Results: Image Distortion



	Precision / Recall		Density / Coverage		Symmetric P / R		Chamfer (ours)	
Noise σ	Prec. ↑	Recall ↑	Dens. ↑	Cov. ↑	Prec. ↑	Recall ↑	Fid. ↓	Cov. ↓
0.10	1.00	1.00	1.21	1.00	1.00	1.00	**31.38**	**31.40**
0.125	1.00	1.00	1.22	1.00	1.00	1.00	**31.94**	**31.98**
0.15	1.00	1.00	1.22	1.00	1.00	1.00	**32.60**	**32.65**
0.175	1.00	1.00	1.21	0.99	1.00	0.99	**33.35**	**33.40**
0.20	1.00	1.00	1.21	0.99	1.00	0.99	**34.11**	**34.17**
0.225	1.00	1.00	1.18	0.99	0.99	0.99	**34.96**	**35.03**
% change	−0.2%	−0.2%	−2.3%	−0.3%	−0.2%	−0.3%	**+11.4%**	**+11.6%**