The Fault in Our Metrics: Revisiting Generative Model Evaluation with Chamfer Distance

Chirag Vashist[1,†], **Vaibhav Santhosh**[2], **Tristan Engst**[1], **Yanshu Zhang**[1], **Ke Li**[1,3,4]
[1]Simon Fraser University [2]NIT Karnataka [3]Amii [4]CIFAR

serchirag.github.io/chamfer

Abstract

Evaluation metrics decide which image generative models are judged better, yet widely used metrics are either uninformative or unreliable. Generative performance should be evaluated across two separate axes: the realism of the generated samples, and their coverage of the real distribution. Most popular metrics, including FID and KID, report a single symmetric score that conflates these two axes, which makes diagnosing and comparing models hard. A second family of metrics addresses this limitation by reporting along the two axes separately. We call such metrics disentangled, and they include Precision/Recall and Density/Coverage. However, we show that existing disentangled metrics approximate empirical distributions with oversized volumes, which can cause them to admit off-manifold samples. This causes these metrics to respond unreliably to simple distributional changes and image distortions. We propose Chamfer distance, an intuitive disentangled metric that directly measures cross-set nearest-neighbor distances between real and generated samples. We show that the resulting one-sided distances provide interpretable signals for fidelity and diversity, respond reliably to simple distributional changes, while remaining computationally efficient.

1 Introduction

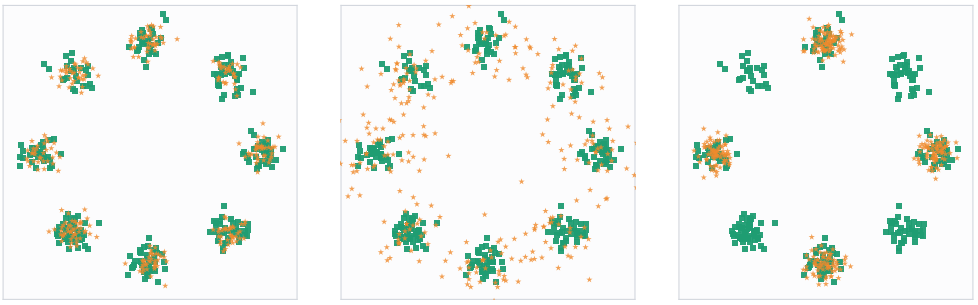
Image-based generative models have advanced rapidly, producing increasingly realistic and diverse samples [12, 15, 18, 20, 30, 32]. As visual quality improves, evaluation metrics increasingly determine which methods are judged better, which trade-offs are visible, and which directions the field prioritizes. Because metrics carry this much weight, they need to be both informative enough to diagnose generative models and reliable enough to respond correctly to changes in the generated distribution.

The most widely used metrics, including FID [14, 28], $\text{FD}_{\text{DINOv2}}$ [38], FDr^{K} [42], KID [3], and CMMD [16], summarize the difference between the real and generated distributions as a *single symmetric score*. These scores are simple and stable to compute, but a single symmetric number has limited diagnostic value: it captures the size of the difference between two distributions, but not its direction. 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 meaningfully analyze and compare the performance of different generative models.

A second line of work addresses this limitation by proposing *disentangled* metrics that separately measure performance along the two axes. The most notable are Precision and Recall [19], which estimate the support of each empirical distribution as a union of isotropic balls centered on its own samples. The metrics are calculated by checking whether a sample from one distribution falls inside the other's estimated support. We show that these balls can have large volumes, which makes the

[†]Corresponding author: chirag_vashist@sfu.ca

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Track on Evaluations and Datasets.



(a) High fidelity and diversity (b) Low fidelity, high diversity (c) Low diversity, high fidelity

Figure 1: **A single symmetric score does not identify which axis a model failed on.** Green squares (■) are the real distribution the models are trained on. Orange stars (★) are samples from three different models with varying generative performance.

estimated support extend well beyond the true manifold. Because membership is a binary test, off-manifold points that are far from any real data can still be counted as valid. This has practical consequences, and we demonstrate that such metrics are unreliable even in simple 2D examples. Moreover, we observe that they are insensitive to image degradation and remain stagnant even as images get visibly corrupted.

In this work, we propose Chamfer distance: a *disentangled* metric that does not rely on inflated approximations of distributions or binary membership tests. Instead, we report fidelity and diversity separately by directly measuring *cross-set* nearest-neighbor distances between real and generated samples. We show that Chamfer distance provides sample-level interpretability, while retaining sensitivity to distributional distortions, reducing dependence on neighborhood parameters, and remaining computationally efficient. Our contributions are as follows:

- We identify a gap in existing evaluation metrics: they either conflate fidelity/diversity or, when they do not separate them, they produce unreliable scores.
- Using a simple two-dimensional example, we show that popular metrics assign perfect scores to two clearly distinct distributions. Through controlled experiments, we show that existing ball-based metrics fail to respond to visible image degradation and do not improve as the candidate set grows denser.
- Finally, we propose Chamfer distance, a disentangled metric that is intuitive, interpretable and computationally efficient. We benchmark modern generative models and show that Chamfer distance remains informative in cases where existing metrics saturate or disagree with each other.

2 Background

In this section, we briefly review existing evaluation metrics. The metrics we discuss in this work compare a *reference set* of real images against a *candidate set*, which in practice contains samples from the model being evaluated. The two distributions are often compared in the embedding space of a pretrained encoder [25, 31, 40].

Distributional Metrics Fréchet Inception Distance (FID) is arguably the most widely used evaluation metric for image generation models. FID embeds real and candidate/generated images into the feature space of a pretrained Inception-v3 network [40], fits a multivariate Gaussian to each set of embeddings, and computes the Fréchet distance between the two fitted Gaussian distributions. Despite its popularity, several works have exposed structural flaws of FID [16, 22, 24] and we discuss them in Appendix A. Other works compute the Fréchet distance with different encoders [38, 42], in part because Inception features align poorly with human judgments of image quality [38]. A different line of work, such as KID and CMMD, avoid the Gaussian assumption of FID by comparing the two sets through Maximum Mean Discrepancy (MMD) [13].

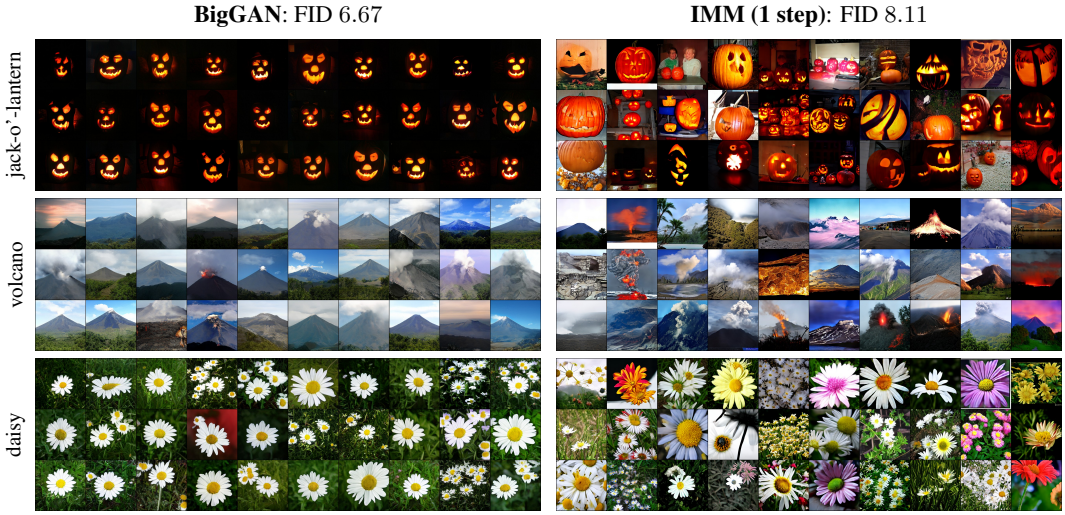


Figure 2: **Random samples for three ImageNet classes**: Samples from IMM look more visually diverse, yet FID ranks BigGAN [5] above IMM [45].

We note that by formulation, distributional metrics are symmetric: for two distributions P and Q, they satisfy $d(P, Q) = d(Q, P)$. Hence, they reduce the discrepancy between two distributions to a single symmetric score, which can indicate how large that discrepancy is, but not where it comes from. Consider Figure 1, which shows candidate samples from three models evaluated against the same reference distribution. Model A (Figure 1a) is well fitted: its samples lie close to real data (high fidelity) and cover it well (high diversity). Model B (Figure 1b) covers every real sample, but many of its own samples lie far from any real data (low fidelity, high diversity). Model C (Figure 1c) places all of its samples close to real data, but leaves many real samples with no generated sample nearby (high fidelity, low diversity). Any reasonable evaluation metric would rank Model A above both Model B and Model C. A symmetric score for Model B and Model C would be worse compared to Model A, but it would not indicate which of these two failure modes has occurred. Hence, a symmetric score conflates fidelity and diversity, making it hard to analyze and compare different models.

This can be demonstrated by a practical example too. Consider Figure 2 and Figure 9 that shows random samples from BigGAN and IMM (1-step) for different ImageNet classes. Looking at the samples, it is perceptible that IMM has more visually diverse samples. Yet, a distributional metric like FID just reports a single number, which does not capture this obvious visual difference.

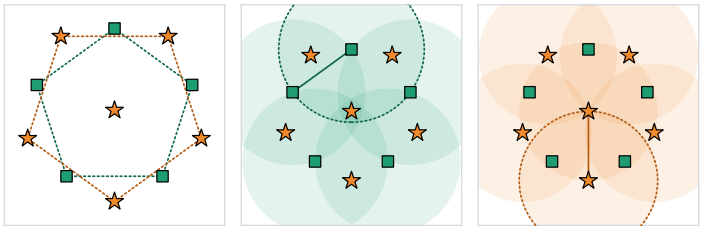
Disentangled Metrics In order to address this drawback, *disentangled metrics* have been proposed that measure generative models along two separate axes. Most existing disentangled metrics follow the *KNN-ball framework*, first proposed by Kynkäänniemi et al. [19]. The support of each empirical distribution is approximated by placing a ball around every sample, with radii set by local nearest-neighbor distances. The union of all such balls gives the estimated support for that empirical distribution. These estimated supports make it possible to answer binary membership queries of the form: "Does a sample x lie in the support of distribution P?". Concretely, let P and Q denote the real and candidate distributions respectively, with samples $\{x_i\}_{i=1}^N \sim P$ and $\{y_j\}_{j=1}^M \sim Q$. The support of P is approximated by the union of balls centered at each real sample: $\widehat{\mathcal{M}}_X := \bigcup_{i=1}^N B(f(x_i), \text{NND}_k(f(x_i)))$ where $B(c, r)$ is the ball of radius r around c in feature space, $f(\cdot)$ is a feature extractor, and $\text{NND}_k(f(x_i))$ is the distance from $f(x_i)$ to its k-th nearest neighbor in $f(x_j)_{j=1,\, j\neq i}^N$. The support of Q is defined analogously. Precision and recall are then:

$$\text{Precision} := \frac{1}{M}\sum_{j=1}^{M} \mathbf{1}\left[f(y_j) \in \widehat{\mathcal{M}}_X\right], \qquad \text{Recall} := \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[f(x_i) \in \widehat{\mathcal{M}}_Y\right].$$

Several works refine this KNN-ball framework by modifying the scoring rule including Density/Coverage [24] and Symmetric Precision/Recall [17]. We cover them briefly in Section 3. Despite these refinements, the underlying principle of all these approaches remains the same: each method

uses *within-set* nearest-neighbor distances to construct local support estimates and evaluates samples by their membership in those estimated supports.

3 Limitations of KNN-Ball based metrics



(a) The two pentagons (b) Reference support (c) Candidate support

Figure 3: **A counterexample for KNN-ball metrics.** Two pentagons: a reference set of green squares (■) and a candidate set of orange stars (★).

Table 1: Metrics for the two pentagons (k=1)

Metric	Value
Precision↑	1.00
Recall↑	1.00
Density↑	2.50
Coverage↑	1.00
Sym. Precision↑	1.00
Sym. Recall↑	1.00
FD↓	0.00

While the motivation behind disentangled metrics is valid, in this paper we argue that *existing* disentangled metrics are unreliable, even in simple cases. These metrics estimate the manifold of each distribution with nearest-neighbour search: they place a ball around every sample, with its radius set by the distance to the k-th nearest neighbour within the same set. In practice, isotropic balls cover not only regions along the true manifold, but also large off-manifold regions far from any real data. Since these metrics rely only on binary membership in this estimate, samples that are far from the true manifold can still be counted as belonging to it, as long as they fall inside some ball.

Consider a simple conceptual example that illustrates this. Figure 3a shows two distributions: one of green squares (■) and another of orange stars (★). Clearly, the two distributions are distinct and any bona fide metric should be able to tell them apart. Yet, as shown in Table 1, existing ball-based metrics obtain perfect scores, and wrongly conclude that the distributions are identical.

Precision and Recall. Precision measures the fraction of candidate samples that lie within the estimated real support, while Recall measures the fraction of real samples that lie within the estimated candidate support. As Figures 3b and 3c show, both metrics achieve a perfect score of 1.

Density and Coverage [24]: Density counts how many real balls contain each candidate, while Coverage measures the fraction of real balls that contain at least one candidate. As Figure 3b shows, Coverage achieves a perfect score of 1, and Density is even higher at 2.5.

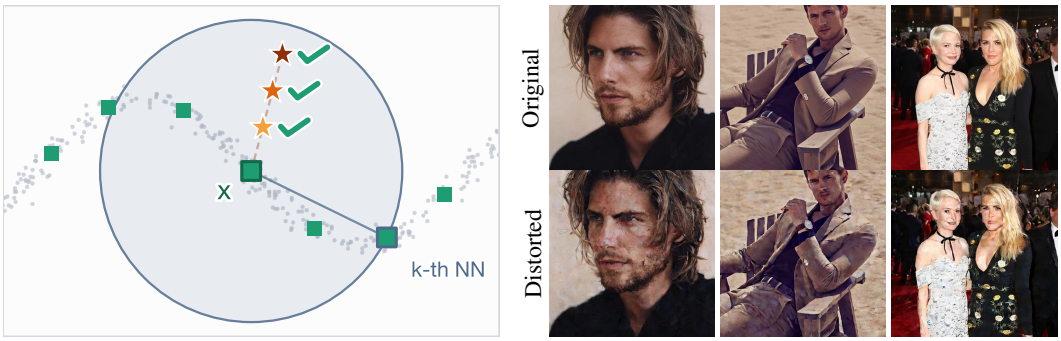
Symmetric Precision and Recall. Khayatkhoei and AbdAlmageed [17] introduce complements of Precision and Recall. For example, cPrecision measures the fraction of candidate samples whose ball contains at least one real sample. Symmetric Precision is defined by taking the minimum of Precision and cPrecision. Similarly, Symmetric Recall is the minimum of Recall and Coverage. In the case of Figure 3a both symmetric metrics achieve a perfect score of 1.

Fréchet Distance (FD) under a Gaussian Assumption. Even though it is not a disentangled ball-based metric, it is worthwhile to note that FD, which underlies FID, $\text{FD}_{\text{DINOv2}}$ [38] and FDr^{K} [42], also fails here. Both sets have the same mean and covariance, so their fitted Gaussians coincide, leading to a perfect FD score of 0.

This example illustrates how existing metrics can be unreliable even in a simple 2D setting. In the following subsections, we explore how the same mechanism affects ball-based metrics on real image data using simple experiments.

3.1 Insensitivity to Image Distortions

In general, any useful evaluation metric should respond when samples are degraded by perceptible distortions. We test this property using 5K randomly sampled LAION-5B images [37]. For each image, we create a paired distorted sample by first encoding the image with a pretrained Stable Diffusion VAE [33]. Then, Gaussian noise with mean zero and standard deviation σ is added to the



(a) The inflated estimated support phenomenon. (b) Real images and their distortions.

Figure 4: **The binary membership test and insensitivity to distortions. (a)** KNN-ball metrics estimate the support around each point x with a ball whose radius is equal to the k-th nearest neighbour distance. Gray dots (●) show the data manifold and green squares (■) the reference samples. Orange stars (★) are increasingly distorted versions of x, obtained by adding an increasing degree of noise. These distorted points can lie far from the manifold, but as long as they lie within the ball, they are counted as valid. **(b)** The original and distorted images from the LAION-5B dataset. Here, the distorted images correspond to a noise level of $\sigma = 0.225$.

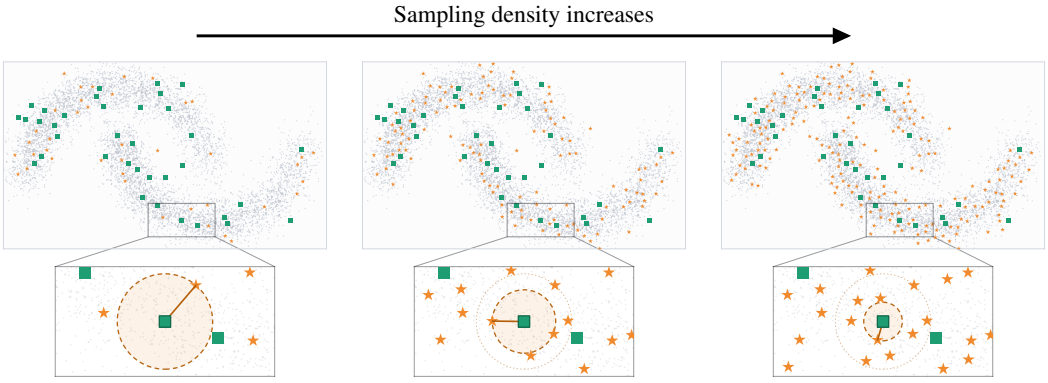
encoded latents. Varying σ controls the degree of corruption. Finally, the corrupted encodings are decoded back to pixel space. For this experiment the metrics are evaluated on features extracted using DINOv2 (ViT-g/14) [25] with neighborhood size fixed at $k = 3$. Clean images are used as reference, while distorted images form the candidate set. Figure 4b shows representative original-distorted pairs at one noise level, illustrating the *perceptual degradation* induced by the corruption process. As shown in Table 2, KNN-ball based metrics remain nearly unchanged as the noise level increases. This happens because ball-based metrics only test for binary membership inside inflated supports. Figure 4a illustrates this in two dimensions: as long as the distorted images remain within the ball, the associated metrics do not register any corruption to the images. Additional results for the same experiment on ImageNet dataset can be found in Appendix H

Table 2: **Sensitivity to Image Distortions.** Metrics are computed using DINOv2 (ViT-g/14) [25] features for 5K LAION images and their distorted counterparts across increasing latent noise levels. Rows correspond to the standard deviation σ of the Gaussian noise added.

	PR [19]		**DC** [24]		**Symmetric** [17]		**Chamfer** *(Ours)*	
σ	Prec.↑	Recall↑	Density↑	Coverage↑	Prec.↑	Recall↑	Fidelity↓	Coverage↓
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

3.2 Insensitivity to Sampling Density

Another limitation of KNN-ball based metrics is that once these balls are large enough, the associated metrics can saturate even when the candidate set becomes substantially denser. In this paragraph, we describe the experimental setup. We consider an *oracle setting*, where the candidate/generated set consists of real images rather than samples from a trained generative model. Equivalently, this can be viewed as sampling from a perfect oracle generator that only produces real images. This removes image quality as a confounding factor from our analysis, since every candidate sample is an actual real image. For the following analysis, the candidate set is formed directly from real ImageNet [34] training images, with randomly sampled subsets of sizes 10K, 30K, and 50K.



(a) 40 candidate samples. (b) 120 candidate samples. (c) 200 candidate samples.

Figure 5: **Sensitivity to sampling density.** A denser candidate set should be recognized as covering the reference set better. Gray dots (●) show the underlying two-dimensional distribution, green squares (■) a fixed reference set, and orange stars (★) the candidate set. Both sets are drawn from the same distribution.

Expected behaviour Before turning to the results, it is useful to reason through what a diversity metric is intended to measure and how it should behave. Diversity asks whether the candidate set represents the different regions of the real data distribution. In the oracle setting, increasing the size of the candidate set gives a denser empirical sample from that same real distribution. Therefore, for any fixed reference image, as the candidate set size increases, it becomes more likely that some candidate image lies nearby. Figure 5 illustrates it in two dimensions: as sampling density increases, each reference point (■) is better covered by a candidate point (★). Aggregated over the reference set, a well-calibrated diversity metric should therefore indicate increased coverage of the reference distribution as the candidate set size increases and hence should get better. We call this desideratum *sensitivity to sampling density*.

Table 3: **Oracle real-subset evaluation on ImageNet-1K** The standard Precision/Recall evaluation protocol uses a 10K-image real subset as the reference set [8]. We evaluate real ImageNet training subsets of size 10K, 30K, and 50K as the candidate/generated set. FID is computed against statistics estimated from the full ImageNet training set. The neighborhood hyperparameter $k = 1$ and features are extracted using Inception-v3 [40].

	FID↓	**IS**↑	**PR**		**Symmetric**		**Chamfer** *(Ours)*	
Gen. size			Prec.↑	Recall↑	Prec.↑	Recall↑	Fid.↓	Cov.↓
10K	4.13	330.90	0.51	0.52	0.50	0.50	10.59	10.57
30K	1.37	331.75	0.51	0.49	0.25	0.49	10.60	9.90
50K	0.85	331.71	0.51	0.50	0.17	0.50	10.61	9.56

Results Table 3 shows that KNN-ball diversity metrics do not respond correctly to change in sampling-density. Recall is essentially flat, and even decreases from 0.52 at 10K to 0.50 at 50K, although the larger candidate set is a denser sample from the same real distribution. Notably, the same trend holds for Symmetric Recall [17], which was proposed to mitigate the limitations of previous methods. This behavior is consistent with the inflated-ball phenomenon. At 10K candidates, the estimated KNN-ball support already covers a large region, so many reference samples are counted as covered by the binary membership test. Adding more real candidate samples then does not improve the score, even though coverage should improve in the oracle setting. Instead of tracking the sampling density of the candidate set, these metrics primarily reflect membership in an inflated support estimate.

Note on performance of FID: Since FID is the most widely used evaluation metric, it is worthwhile to take a short detour and observe how FID behaves under oracle setting. For FID, each candidate set (10K, 30K, and 50K images) is compared against reference statistics estimated from the full ImageNet training set of 1.2 million images. Even the 10K subset obtains an FID of 4.13, despite

containing less than 1% of the ImageNet training set. Increasing the candidate set size further drives FID down to 1.37 at 30K and 0.85 at 50K. These scores would normally be interpreted as very strong generative performance [36, 45]. This result is significant because it shows that *FID can be satisfied by matching global feature moments over a small finite subset*. A method can therefore achieve a low FID while still having poor distributional coverage, especially if it memorizes, retrieves, or over-represents common modes. In this sense, the experiment demonstrates that low FID is not sufficient evidence that a model has learned the full data distribution.

Additionally, Appendix D reports how the above metrics behave under further distributional changes.

4 Proposed Metric

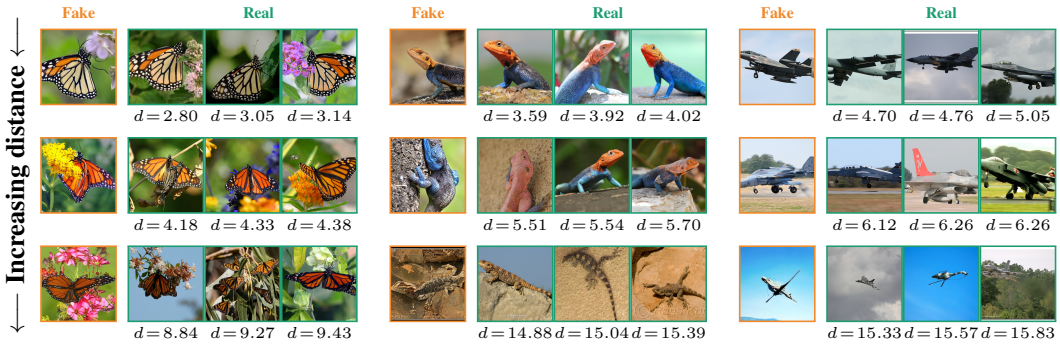


Figure 6: **Cross-set nearest neighbor search distance as a fidelity signal.** Each column group corresponds to a distinct ImageNet class. In each group, the leftmost image is from the generated set, and the three images to its right are its nearest neighbors from the real set. Nearest neighbors are computed in Inception-v3 feature space, and d denotes the distance from each generated image to its nearest real neighbor. Rows are sorted by increasing nearest-neighbor distance. As the distance increases from top to bottom, the generated images become visibly less realistic.

Our aim in this paper is to propose a disentangled metric that reports fidelity and diversity separately, without performing binary membership test on inflated estimated supports. Previous sections suggest that the problem is not nearest-neighbor search itself, but how existing metrics use it. KNN-ball metrics first build an estimated support from *within-set* neighborhoods and then apply a binary membership tests. We instead consider using nearest-neighbor search directly *across* the real and generated sets, giving us a simple distance-based view of both fidelity and diversity.

Figure 6 illustrates the basic intuition for fidelity. For each generated query image, we retrieve its nearest neighbors in the real set. For this visualization, the real set consists of ImageNet-1K images, the generated set consists of samples from IMM-XL (8 steps) [45], and nearest-neighbor distances are computed in the Inception-v3 feature space. If the distance is small, the retrieved real images are similar to the query, implying that the generated sample lies close to the real images. This indicates that these generated samples are realistic (top row in Figure 6). If the distance is large, this implies that the generated sample is far from any real image, suggesting that it is unrealistic (bottom row in Figure 6). Figure 7 illustrates the corresponding intuition for diversity. For each real image, we retrieve its nearest neighbors in the generated set. If the distance is small, the retrieved generated images are similar to the real query, implying that the generator has produced samples near this region of the real data distribution. This indicates that the corresponding real samples are well covered by the model (top row in Figure 7). If the distance is large, then the real query is far from any generated image, suggesting that the generator has failed to cover this region of the real distribution (bottom row in Figure 7).

We now turn this nearest-neighbor intuition into a statistical measure. Let P and Q be arbitrary distributions, and let $P_n = \{x_1, \ldots, x_n\}$ and $Q_m = \{y_1, \ldots, y_m\}$ be finite samples from them. For a query point $x \sim P$, we measure its discrepancy from Q by averaging the distances to its K nearest neighbors in Q_m. The total discrepancy from P to Q is then obtained by averaging this quantity over

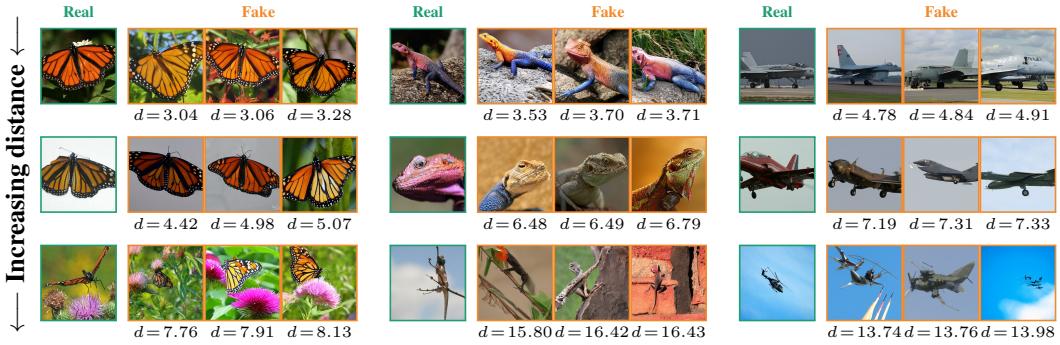


Figure 7: **Cross-set nearest neighbor search distance as a diversity signal.** Each column group corresponds to a distinct ImageNet class. In each group, the leftmost image is from the real set, and the three images to its right are its nearest neighbors from the generated set. Nearest neighbors are computed in Inception-v3 feature space, and d denotes the distance from each real image to its nearest generated neighbor. Rows are sorted by increasing nearest-neighbor distance. As the distance increases from top to bottom, real images become less well represented by the generated set, indicating poorer coverage of the data.

points drawn from P:

$$d(P_n, Q_m) = \mathop{\mathbb{E}}_{x \sim P_n} \left[\frac{1}{K} \sum_{j=1}^{K} \left\| f(x) - f(y_j^*) \right\|_2 \right], \tag{1}$$

where $y_1^*, \ldots, y_K^*$ are the K nearest neighbors of x in Q_m, ordered by distance in feature space.

This construction is a direct generalization of the one-sided Chamfer distance [1, 2, 4, 10], a standard measure for comparing point clouds in 3D computer vision. When $K = 1$ and P_n and Q_m are finite 3D point sets, the symmetrized version recovers the classical Chamfer distance:

$$d_{\text{Sym-Chamfer}}(P_n, Q_m) = \frac{1}{|P_n|} \sum_{x \in P_n} \min_{y \in Q_m} \|x - y\|_2 + \frac{1}{|Q_m|} \sum_{y \in Q_m} \min_{x \in P_n} \|y - x\|_2.$$

For this reason, we refer to our metric as Chamfer: it preserves the nearest-neighbor interpretation of Chamfer distance while extending it to K-nearest-neighbor distances in high-dimensional feature spaces. For generative model evaluation, we take P_n to be samples from the generated distribution P and Q_m to be samples from the real distribution Q. In this direction, $d(P_n, Q_m)$ measures how close generated samples are to real samples, so we call it **Chamfer-Fidelity**. Reversing the direction measures how well the generated distribution covers the real distribution, so $d(Q_m, P_n)$ is **Chamfer-Coverage**. Both metrics are distances, so lower values indicate better performance.

Before moving to experiments, we summarize how our proposed metric avoids the issues with the popular metrics discussed above. Unlike distributional metrics, it reports fidelity and diversity as two separate metrics. Unlike KNN-ball metrics, it does not estimate a support or rely on binary membership tests. As a result, off-manifold samples are penalized by their actual distance to the real set rather than being accepted simply because they fall inside an inflated support estimate. For example, our metrics are non-zero for the two pentagons in Figure 3a, correctly indicating that the two distributions are not identical, whereas other metrics in Table 1 report perfect scores. Additionally, because the final score is an average over per-sample distances, the metric is naturally diagnostic: each term corresponds to an individual generated or real sample, allowing practitioners to identify low-fidelity generations and under-covered regions of the data distribution.

5 Experiments

Sensitivity to Image Distortions We return to the image distortion experiment presented in Section 3.1: despite visible degradation in the distorted images, KNN-ball based metrics remained stagnant. As shown in Table 2, our proposed metrics Chamfer-Fidelity and Chamfer-Coverage increase monotonically with noise level, correctly reflecting the degradation in image quality.

Table 4: **Sensitivity to number of neighbors k:** We evaluate 10K samples from iMeanFlow [11] (NFE=1) against 10K ImageNet reference images using Inception features.

k	Prec.	Recall	Density	Cov.	Chamfer-Fid.	Chamfer-Cov.
1	0.52	0.44	1.06	0.50	10.50	10.65
3	0.78	0.70	1.08	0.88	11.04	11.17
5	0.87	0.80	1.07	0.97	11.47	11.56
10	0.95	0.91	1.07	1.00	12.32	12.37
20	0.98	0.95	1.08	1.00	13.37	13.44
50	0.99	0.97	1.09	1.00	14.53	14.63
100	0.99	0.98	1.10	1.00	15.12	15.24

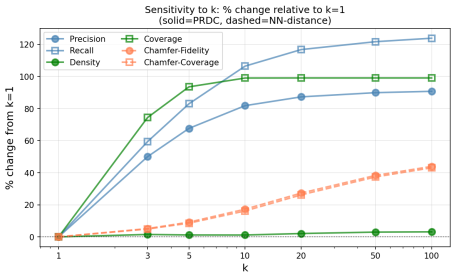


Figure 8: **Sensitivity to number of neighbors.** While commonly used ball-based evaluation metrics exhibit high sensitivity to k, our metrics show gradual change.

Table 5: **Wall-clock computation time.** We report post feature-extraction runtime (in seconds) for FID, Precision/Recall, and Chamfer distance across increasing reference and candidate set sizes.

Ref. Size N	Cand. Size N	FID	Precision/Recall	Chamfer
10K	10K	8.36	1.71	**1.00**
	50K	9.67	6.84	**1.15**
100K	100K	25.95	52.12	**4.20**

Sensitivity to Sampling Density Chamfer distance correctly responds to sampling density. Under the oracle setting of Table 3, where the generated set consists of real ImageNet images at three sizes (10K, 30K, 50K), Chamfer-Coverage decreases monotonically from 10.57 to 9.56, reflecting that a denser generated set places its samples closer to the reference on average.

Sensitivity to Number of Neighbors KNN-ball metrics are highly sensitive to the neighborhood parameter k, since increasing k directly enlarges the estimated support. We quantify this dependence by evaluating 10K iMeanFlow [11] samples (NFE$=1$) against 10K ImageNet references using Inception features, while sweeping $k \in \{1, 3, 5, 10, 20, 50, 100\}$. As shown in Table 4 and Figure 8, changing k from $1 \to 3$ raises Precision from 0.52 to 0.78, Recall from 0.43 to 0.69, and Coverage from 0.50 to 0.87. As k grows, these binary membership scores approach saturation; Coverage reaches 1.0 by $k = 20$, after which it no longer distinguishes larger neighborhoods. In practice for ball-based metrics, k is a *critical hyperparameter* that must be tuned for each evaluation setting. In contrast, our metrics change smoothly across the sweep, indicating lower sensitivity to k.

Computational Efficiency We benchmark post feature-extraction runtime for FID, Precision/Recall, and Chamfer distance in Table 5 using a single NVIDIA L40S. All metrics are computed from the same 2048-dimensional Inception-v3 features, so the reported times exclude feature extraction. Chamfer distance is considerably faster than FID and Precision/Recall, allowing it to scale to much larger sets of reference and generated images. This efficiency arises from the fact that Chamfer distance relies on nearest-neighbor lookups, which can be efficiently implemented using GPU-accelerated libraries like FAISS [9] (see Appendix J for complexity analysis).

5.1 Benchmarking Modern Generative Models

We now turn to evaluating trained generative models. Table 6 reports evaluation results across popular models using 100K generated samples against 100K reference images from ImageNet. Evaluation protocols are provided in Appendix I. We observe that our proposed metric agrees with many accepted notions in the literature. For example, GANs which are prone to suffer from mode collapse, have poor Chamfer-Coverage. For IMM, both Chamfer-Fidelity and Chamfer-Coverage decrease monotonically as NFE increases, consistent with the expectation that additional sampling steps refine generated samples. More informative are the cases where existing metrics saturate or disagree, and Chamfer resolves what they leave ambiguous:

- BigGAN obtains a better FID than IMM (1 step), yet IMM has a much better Chamfer-Coverage (34.75 against 40.09). This agrees with the samples in Figures 2 and 9, where IMM is visibly more diverse. This reveals that FID does not reflect a difference in diversity.
- For IMM, Recall peaks at 2 steps and stagnates thereafter, suggesting no further diversity gains. Chamfer-Coverage, however, continues to improve monotonically with NFE, showing that additional sampling steps improve the diversity.
- Recall ranks IMM (2 steps) above VAR, but both FD-based metrics and Chamfer-Coverage agree that VAR is the better model, which is not surprising as VAR has much higher number of parameters.
- $\text{FD}_{\text{DINOv2}}$ ranks IMM (2 steps) above RAR-XL (230.07 against 263.36), even though RAR-XL has a far better FID (1.19 against 3.67). Chamfer ranks RAR-XL ahead on both fidelity and coverage.
- StyleGAN-XL and DiT-XL/2 have similar FIDs (2.08 and 1.92), yet DiT-XL/2 is better on *both* Chamfer metrics.
- LlamaGen-L and RAR-XL have similar $\text{FD}_{\text{DINOv2}}$, yet each is better on a different axis: LlamaGen-L has better Chamfer-Fidelity and RAR-XL has better Chamfer-Coverage. A single symmetric score like $\text{FD}_{\text{DINOv2}}$ cannot show the axis along which the models are weak.
- Precision rates StyleGAN-XL lowest (0.29), although StyleGAN-XL beats BigGAN on FID and $\text{FD}_{\text{DINOv2}}$. It rates TiTok-L-32 third (0.49), above Priming-L (0.47), which has much better FID and $\text{FD}_{\text{DINOv2}}$. Chamfer-Fidelity agrees with FID and $\text{FD}_{\text{DINOv2}}$ in both cases.

Table 6: Generative Model Evaluation on ImageNet 256 (100K samples, $k = 3$). Values are mean $\pm$ std over 3 runs; each run draws a new set of 100K generated samples with a different random seed. A std shown as 0.00 is below 0.005.

			DINOv2-Giant (ViT-g/14)					Inception-v3
Model	**Params (M)**	**NFE ↓**	**$\text{FD}_{\text{DINOv2}}$ ↓**	**Precision ↑**	**Recall ↑**	**Chamfer-Fid. ↓**	**Chamfer-Cov. ↓**	**FID ↓**
BigGAN [5]	112	1	$743.29_{\pm 0.85}$	$0.47_{\pm 0.00}$	$0.14_{\pm 0.01}$	$44.59_{\pm 0.05}$	$40.09_{\pm 0.06}$	$6.67_{\pm 0.05}$
StyleGAN-XL [35]	166	1	$317.99_{\pm 0.81}$	$0.29_{\pm 0.00}$	$0.40_{\pm 0.00}$	$42.32_{\pm 0.01}$	$35.92_{\pm 0.02}$	$2.08_{\pm 0.03}$
DiT-XL/2 (cfg = 1.5) [29]	675	250	$108.55_{\pm 0.86}$	$0.56_{\pm 0.00}$	$0.58_{\pm 0.00}$	$32.72_{\pm 0.01}$	$31.18_{\pm 0.01}$	$1.92_{\pm 0.01}$
SiT-XL/2 (cfg = 1.5) [23]	675	250	$151.16_{\pm 0.72}$	$0.48_{\pm 0.00}$	$0.59_{\pm 0.00}$	$35.39_{\pm 0.07}$	$31.71_{\pm 0.01}$	$1.74_{\pm 0.01}$
IMM (1 step, cfg = 1.5) [45]	676	1	$384.37_{\pm 1.77}$	$0.32_{\pm 0.00}$	$0.62_{\pm 0.00}$	$44.29_{\pm 0.04}$	$34.75_{\pm 0.03}$	$8.11_{\pm 0.08}$
IMM (2 steps, cfg = 1.5)	676	2	$230.07_{\pm 0.61}$	$0.38_{\pm 0.00}$	$0.68_{\pm 0.00}$	$40.11_{\pm 0.03}$	$33.02_{\pm 0.02}$	$3.67_{\pm 0.05}$
IMM (4 steps, cfg = 1.5)	676	4	$163.24_{\pm 0.86}$	$0.43_{\pm 0.00}$	$0.68_{\pm 0.00}$	$37.51_{\pm 0.06}$	$32.25_{\pm 0.03}$	$2.19_{\pm 0.01}$
IMM (8 steps, cfg = 1.5)	676	8	$128.80_{\pm 0.31}$	$0.47_{\pm 0.00}$	$0.67_{\pm 0.00}$	$35.76_{\pm 0.01}$	$31.81_{\pm 0.03}$	$1.69_{\pm 0.01}$
iMeanFlow [11]	610	1	$138.65_{\pm 0.41}$	$0.48_{\pm 0.00}$	$0.57_{\pm 0.00}$	$34.65_{\pm 0.05}$	$31.51_{\pm 0.02}$	$1.48_{\pm 0.01}$
Drifting-L [7]	463	1	$213.14_{\pm 1.14}$	$0.39_{\pm 0.00}$	$0.56_{\pm 0.00}$	$38.40_{\pm 0.06}$	$32.89_{\pm 0.01}$	$1.24_{\pm 0.01}$
Priming-L [27]	405	1	$88.12_{\pm 0.18}$	$0.47_{\pm 0.00}$	$0.66_{\pm 0.00}$	$34.32_{\pm 0.02}$	$31.34_{\pm 0.00}$	$0.93_{\pm 0.01}$
VAR [41]	2000	10	$104.07_{\pm 0.20}$	$0.56_{\pm 0.00}$	$0.63_{\pm 0.00}$	$32.99_{\pm 0.03}$	$30.70_{\pm 0.00}$	$1.63_{\pm 0.01}$
LlamaGen-L [39]	343	256	$269.79_{\pm 0.63}$	$0.46_{\pm 0.00}$	$0.42_{\pm 0.00}$	$36.04_{\pm 0.06}$	$33.40_{\pm 0.01}$	$4.13_{\pm 0.01}$
Open-MAGVIT2-XL [21]	1500	256	$145.70_{\pm 0.25}$	$0.48_{\pm 0.00}$	$0.58_{\pm 0.00}$	$34.95_{\pm 0.01}$	$31.79_{\pm 0.01}$	$2.09_{\pm 0.01}$
RAR-XL [43]	955	256	$263.36_{\pm 0.11}$	$0.44_{\pm 0.00}$	$0.49_{\pm 0.00}$	$37.61_{\pm 0.00}$	$32.76_{\pm 0.01}$	$1.19_{\pm 0.00}$
RandAR-XL [26]	775	88	$243.24_{\pm 0.59}$	$0.48_{\pm 0.00}$	$0.49_{\pm 0.00}$	$35.74_{\pm 0.02}$	$32.51_{\pm 0.01}$	$1.90_{\pm 0.01}$
TiTok-L-32 [44]	177	8	$377.30_{\pm 0.67}$	$0.49_{\pm 0.00}$	$0.42_{\pm 0.00}$	$41.65_{\pm 0.02}$	$33.39_{\pm 0.03}$	$2.50_{\pm 0.03}$

6 Discussions and Conclusion

Future Work This work studies existing generative evaluation metrics from an empirical perspective and proposes Chamfer distance as an alternative that avoids their limitations. We note that Chamfer distance is encoder-agnostic and can be paired with feature spaces that better capture human perception, aesthetic quality, semantic alignment, or domain-specific similarity. Studying these encoder choices is an important direction left for future work.

Conclusion Widely used evaluation metrics often conflate fidelity and diversity into a single symmetric score. Existing metrics that report the two separately rely on inflated support estimates, which makes them unreliable. We propose Chamfer distance, a disentangled metric based on cross-set nearest-neighbor, that provides interpretable signals for fidelity and coverage. Across our analyses, Chamfer distance is less sensitive to the neighborhood parameter, responds more meaningfully to image corruption and sampling density, and remains computationally efficient.

Acknowledgements This research was enabled in part by NSERC, the Canada Foundation for Innovation, the Canada CIFAR AI Chairs program, the BC DRI Group and the Digital Research Alliance of Canada.

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds, 2018. URL https://arxiv.org/abs/1707.02392.

[2] H. G. Barrow, J. M. Tenenbaum, R. C. Bolles, and H. C. Wolf. Parametric correspondence and chamfer matching: two new techniques for image matching. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'77, page 659–663, San Francisco, CA, USA, 1977. Morgan Kaufmann Publishers Inc.

[3] Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans, 2021. URL https://arxiv.org/abs/1801.01401.

[4] Gunilla Borgefors. Hierarchical chamfer matching: A parametric edge matching algorithm. *IEEE Trans. Pattern Anal. Mach. Intell.*, 10(6):849–865, 1988. doi: 10.1109/34.9107. URL https://doi.org/10.1109/34.9107.

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1xsqj09Fm.

[6] Min Jin Chong and David Forsyth. Effectively unbiased fid and inception score and where to find them, 2020. URL https://arxiv.org/abs/1911.07023.

[7] Mingyang Deng, He Li, Tianhong Li, Yilun Du, and Kaiming He. Generative modeling via drifting. *arXiv preprint arXiv:2602.04770*, 2026.

[8] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. URL https://arxiv.org/abs/2105.05233.

[9] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library, 2025. URL https://arxiv.org/abs/2401.08281.

[10] Haoqiang Fan, Hao Su, and Leonidas Guibas. A point set generation network for 3d object reconstruction from a single image, 2016. URL https://arxiv.org/abs/1612.00603.

[11] Zhengyang Geng, Yiyang Lu, Zongze Wu, Eli Shechtman, J Zico Kolter, and Kaiming He. Improved mean flows: On the challenges of fastforward generative models. *arXiv preprint arXiv:2512.02012*, 2025.

[12] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL https://arxiv.org/abs/1406.2661.

[13] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

[14] 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. URL https://arxiv.org/abs/1706.08500.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

[16] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation, 2024. URL https://arxiv.org/abs/2401.09603.

[17] Mahyar Khayatkhoei and Wael AbdAlmageed. Emergent asymmetry of precision and recall for measuring fidelity and diversity of generative models in high dimensions, 2023. URL https://arxiv.org/abs/2306.09618.

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. URL https://arxiv.org/abs/1312.6114.

[19] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models, 2019. URL https://arxiv.org/abs/1904.06991.

[20] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

[21] Zhuoyan Luo, Fengyuan Shi, Yixiao Ge, Yujiu Yang, Limin Wang, and Ying Shan. Open-magvit2: An open-source project toward democratizing auto-regressive visual generation, 2025. URL https://arxiv.org/abs/2409.04410.

[22] Lorenzo Luzi, Carlos Ortiz Marrero, Nile Wynar, Richard G. Baraniuk, and Michael J. Henry. Evaluating generative networks using gaussian mixtures of image features, 2022. URL https://arxiv.org/abs/2110.05240.

[23] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pages 23–40. Springer, 2024.

[24] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models, 2020. URL https://arxiv.org/abs/2002.09797.

[25] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

[26] Ziqi Pang, Tianyuan Zhang, Fujun Luan, Yunze Man, Hao Tan, Kai Zhang, William T. Freeman, and Yu-Xiong Wang. Randar: Decoder-only autoregressive visual generation in random orders, 2025. URL https://arxiv.org/abs/2412.01827.

[27] Woobin Park, Jaeyeon Kim, Yilun Du, Sitan Chen, and Bohyung Han. Are we overthinking one-step generation?, 2026. URL https://arxiv.org/abs/2610.20331.

[28] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, 2022.

[29] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[30] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2016. URL https://arxiv.org/abs/1511.06434.

[31] 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, 2021. URL https://arxiv.org/abs/2103.00020.

[32] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows, 2016. URL https://arxiv.org/abs/1505.05770.

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2015. URL https://arxiv.org/abs/1409.0575.

[35] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022.

[36] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets, 2022. URL https://arxiv.org/abs/2202.00273.

[37] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. URL https://arxiv.org/abs/2210.08402.

[38] 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. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[39] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

[40] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions, 2014. URL https://arxiv.org/abs/1409.4842.

[41] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

[42] Jiawei Yang, Zhengyang Geng, Xuan Ju, Yonglong Tian, and Yue Wang. Representation fréchet loss for visual generation, 2026. URL https://arxiv.org/abs/2604.28190.

[43] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Randomized autoregressive visual generation, 2024. URL https://arxiv.org/abs/2411.00776.

[44] Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation, 2024. URL https://arxiv.org/abs/2406.07550.

[45] Linqi Zhou, Stefano Ermon, and Jiaming Song. Inductive moment matching. In *Forty-second International Conference on Machine Learning*, 2025.

A Limitations of FID

Despite its widespread use, FID has several known limitations, including sensitivity to sample size [6], implementation details [28], and the choice of feature encoder [38, 42]. We focus on a more fundamental issue: FID approximates complex feature distributions with a single Gaussian. Given real and generated samples, FID embeds both sets using a feature extractor and approximates each resulting feature distribution by a single Gaussian. The score then compares only the fitted means and covariance matrices. Thus, FID does not compare the full feature distributions; it compares a two-moment summary of them. This creates a fundamental blind spot. Distinct distributions can have identical first and second moments while differing substantially in shape, support, modality, and tail behaviour. This failure is not tied to a particular encoder: any metric (including $\text{FD}_{\text{DINOv2}}$ [38]) that reduces a feature distribution to a single mean and covariance matrix is oblivious to discrepancies beyond second-order structure.

To make this limitation explicit, consider two distributions in $\mathbb{R}^d$: $P = \mathcal{N}(0, I_d)$ and $Q = \text{Unif}([-\sqrt{3}, \sqrt{3}]^d)$. These distributions are clearly different: P has unbounded support and Gaussian tails, while Q is supported on a compact hypercube. However, they have identical first and second moments. Under the Gaussian approximation used by FID, the moment-matched Gaussians for P and Q are identical, so their Fréchet distance under the Gaussian approximation is zero despite $P \neq Q$. This is further supported by the results in Table 1, where the two distinct sets achieve a perfect FD score.

B Additional Samples from BigGAN and IMM

Figure 9 shows random samples from BigGAN and IMM (1 step) for five additional ImageNet classes, in the same layout as Figure 2.

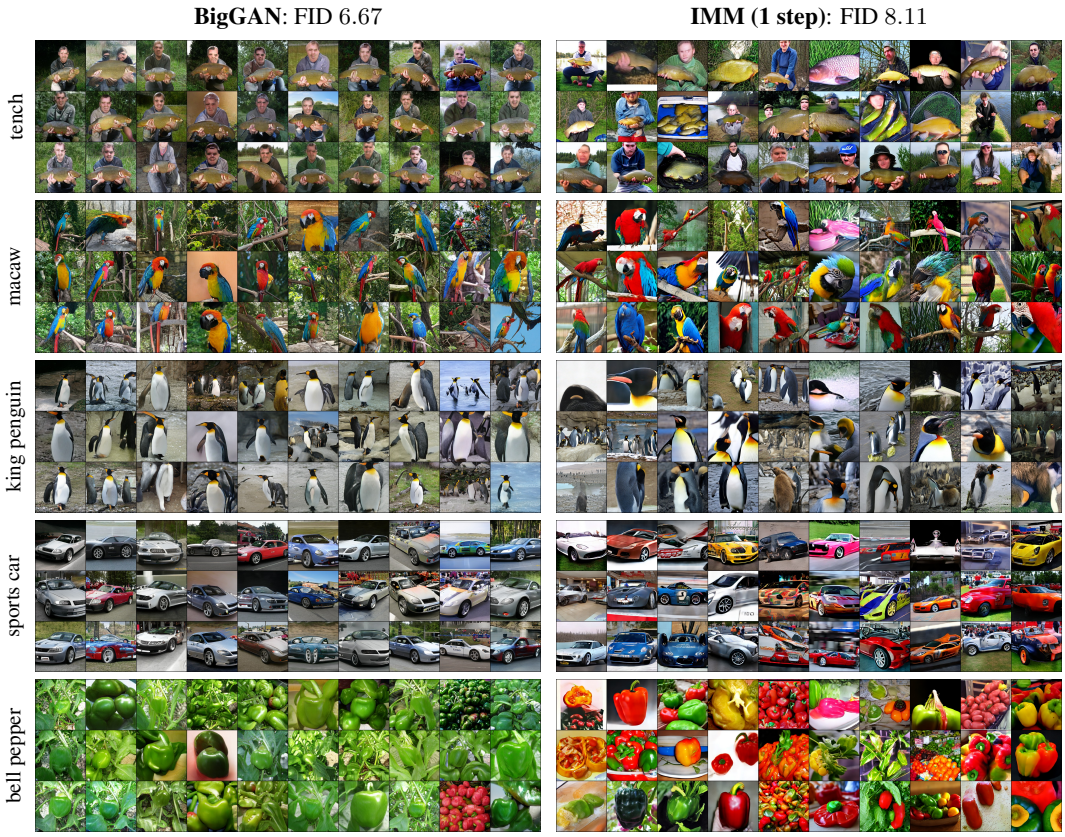


Figure 9: **Random samples for five additional ImageNet classes.** Same layout as Figure 2.

C Construction of the Pentagon Example

The reference set X (■) in Figure 3a consists of the five vertices of a regular pentagon inscribed in the unit circle:

$$X = \{(\cos\theta_i,\ \sin\theta_i)\}_{i=1}^{5}, \qquad \theta_i = \frac{\pi}{2} + \frac{2\pi i}{5}.$$

The candidate set Y (★) consists of the origin and the five vertices of a second regular pentagon, rotated by $\frac{\pi}{5}$ and scaled by $\rho = \frac{\sqrt{5}}{2}$:

$$Y = \{(0,0)\} \cup \left\{\rho\left(\cos\left(\theta_i + \frac{\pi}{5}\right),\ \sin\left(\theta_i + \frac{\pi}{5}\right)\right)\right\}_{i=1}^{5}.$$

Both sets have zero mean, and the same sample covariance, $\frac{5}{8}I$.

D Evaluation on Distribution Shifts

The main paper studies a controlled shift, *sensitivity to sampling density*, where the candidate set is progressively enlarged while the reference set is held fixed and the behaviour of each metric is measured (Section 3). In this section, we evaluate different metrics on other controlled distribution shifts. All experiments follow the oracle setting from Table 3: the reference set is $10K$ images sampled uniformly from ImageNet, and the candidate set is $50K$. For all the experiments below, we use $k = 3$ and Inception-v3 as the feature encoder. This setup isolates the distributional shift, since both sets are drawn from real images and only the sampling of the candidate set is varied.

D.1 Density Reweighting

We partition the 1000 classes into a group of 100 and a group of 900, keeping both present in the sample set so the support is unchanged. We progressively increase the mass of the 100-class group while correspondingly decreasing the mass of the 900-class group. This shifts the density while holding support fixed. Below, "Head %" is the fraction of the sample-set mass assigned to the 100-class group and "Factor" its upweighting relative to uniform (10% / $1\times$ is uniform).

Table 7: **Density reweighting.** The 1000 classes are split into a head group of 100 and a tail group of 900. Both remain present in the sample set, so the support is unchanged and only the probability mass is shifted.

			PR		**Chamfer** *(Ours)*	
Head %	**Factor**	**FID** ↓	Prec. ↑	Recall ↑	Fid. ↓	Cov. ↓
10%	1× (uniform)	0.76	0.757	0.711	11.21	10.09
20%	2×	1.32	0.760	0.714	11.20	10.10
30%	3×	2.73	0.761	0.711	11.21	10.14
40%	4×	4.89	0.762	0.712	11.20	10.20
50%	5×	7.72	0.763	0.712	11.21	10.26
60%	6×	11.28	0.764	0.713	11.21	10.35
70%	7×	15.55	0.766	0.717	11.21	10.49
80%	8×	20.62	0.766	0.723	11.22	10.69
90%	9×	26.64	0.768	0.740	11.23	11.14

As the head group is upweighted, Recall fails to register the shift: it in fact moves in the wrong direction, increasing slightly ($0.711 \to 0.740$) even as the distribution becomes more concentrated. Chamfer-Coverage, in contrast, responds monotonically ($10.09 \to 11.14$), correctly reflecting that the downweighted 900 classes become progressively less well covered. Chamfer-Fidelity remains flat, as expected, since the reference set is unchanged.

D.2 Sample Duplication

We hold the sample set at 50K while reducing the number of unique images ($50K \to 10K$), filling the shortfall with exact copies. Replication is applied within each class, so every class contributes uniformly.

Table 8: **Sample duplication.** The sample set is held at 50K while the number of unique images is reduced, with the shortfall filled by exact copies.

			PR		Chamfer *(Ours)*	
Unique	**50K / Unique**	**FID** ↓	Prec. ↑	Recall ↑	Fid. ↓	Cov. ↓
50K	×1.00	0.76	0.759	0.713	11.21	10.09
40K	×1.25	1.02	0.760	0.674	11.20	10.13
30K	×1.67	1.29	0.758	0.594	11.21	10.17
20K	×2.50	1.82	0.757	0.492	11.22	10.23
10K	×5.00	3.47	0.756	0.000	11.26	10.53

Chamfer-Coverage degrades smoothly and monotonically as uniqueness falls ($10.09 \rightarrow 10.53$), tracking the loss of distinct samples throughout the sweep. Recall also degrades, but fails discontinuously: at 10K unique every sample carries five copies, giving it a zero-radius neighborhood, so Recall reads exactly 0.

D.3 Dropping subsets of classes

We progressively restrict the sample set to fewer classes ($1000 \rightarrow 500 \rightarrow 250 \rightarrow 100$), while the reference set remains the full uniform 10K. As classes are removed from the sample set, those regions of the real distribution become uncovered.

Table 9: **Dropping subsets of classes.** The sample set is progressively restricted to fewer ImageNet classes while the reference set remains the full uniform 10K subset.

			PR		Chamfer *(Ours)*	
Classes	**Imgs/class**	**FID** ↓	Prec. ↑	Recall ↑	Fid. ↓	Cov. ↓
1000	50	0.76	0.757	0.711	11.21	10.09
500	100	7.80	0.754	0.558	11.19	11.23
250	200	17.66	0.748	0.454	11.20	12.11
100	500	35.46	0.752	0.362	11.15	12.99

For this experiment, all metrics respond correctly, including the ball-based ones. The result confirms our decomposition behaves as intended: Chamfer-Coverage worsens monotonically ($10.09 \rightarrow 12.99$) as classes are removed, while Chamfer-Fidelity stays flat ($11.21 \rightarrow 11.15$), since the remaining samples are still real and close to the reference.

E Ablation over Distance Function

Equation 1 is stated with the ℓ_2 norm, following the convention of the classical Chamfer distance. The construction is not restricted to this choice and any distance function on the feature space yields a valid one-sided measure. In this section we ablate over p-norms (ℓ_1, ℓ_2, ℓ_3, ℓ_4) and cosine distance. We repeat the oracle setting of Table 3: both the sample set and the reference set are drawn from real training images, and we vary the sample-set size (10K / 30K / 50K) to observe the effect of sampling density. Because the absolute scale of the metric differs by orders of magnitude across these choices, we also report ΔCov., the relative decrease in Chamfer-Coverage from 10K to 50K. This gives a scale-free summary of how strongly each variant responds to sampling density.

The qualitative behaviour is invariant across all choices, while only the absolute scale of the metric varies. In every case Chamfer-Coverage decreases monotonically with sampling density, by a consistent 9–10% from 10K to 50K, and Chamfer-Fidelity remains stable because the reference set is fixed. We therefore recommend ℓ_2 as the default.

Table 10: **Ablation over the distance function used in Chamfer distance.** Oracle setting on ImageNet-1K with Inception-v3 features and $k = 1$.

	10K		30K		50K		
Distance	Fid. ↓	Cov. ↓	Fid. ↓	Cov. ↓	Fid. ↓	Cov. ↓	**ΔCov.** ↑
ℓ_1	315.10	312.02	314.38	293.09	314.22	284.05	8.97%
ℓ_2	10.65	10.54	10.62	9.88	10.62	9.57	9.18%
ℓ_3	4.01	3.96	4.00	3.71	4.00	3.59	9.44%
ℓ_4	2.64	2.61	2.64	2.44	2.64	2.36	9.74%
Cosine	0.62	0.62	0.62	0.58	0.62	0.56	9.34%

F Ablation over Encoder

Equation 1 requires only a feature map f, and places no constraint on how that map is obtained. We report results using DINOv2-Giant (ViT-g/14) [25] in the main paper, following common practice, but the construction is deliberately encoder-agnostic. It is therefore reasonable to ask whether the behaviour we report is a property of the metric or of a particular feature space.

In this section we ablate across encoders spanning feature dimensions from 384 to 2048: four DINOv2 [25] vision transformers and Inception-v3 [40]. We repeat the *oracle setting* of Table 3. Because the absolute scale of the metric differs across feature spaces, we report ΔCov., the relative decrease in Chamfer-Coverage from 10K to 50K, as a scale-free summary of how strongly each variant responds.

Table 11: **Ablation over the encoder used in Chamfer distance.** Oracle setting on ImageNet-1K with $k = 1$ and the ℓ_2 norm. The reference set is fixed while the sample set grows from 10K to 50K.

		10K		30K		50K		
Backbone	**Dim.**	Fid. ↓	Cov. ↓	Fid. ↓	Cov. ↓	Fid. ↓	Cov. ↓	**ΔCov.** ↑
DINOv2 ViT-S/14	384	40.50	40.45	40.50	37.51	40.48	36.03	10.93%
DINOv2 ViT-B/14	768	37.82	37.61	37.79	34.58	37.76	33.12	11.95%
DINOv2 ViT-L/14	1024	33.97	33.59	33.92	30.60	33.89	29.21	13.03%
DINOv2 ViT-g/14	1536	31.83	31.29	31.78	28.17	31.74	26.80	14.36%
Inception-v3	2048	10.65	10.54	10.62	9.88	10.62	9.57	9.18%

The qualitative behaviour is invariant across all encoders. In every case Chamfer-Coverage decreases monotonically as the sample set grows denser, correctly reflecting improved coverage, while Chamfer-Fidelity remains essentially constant because the reference set is fixed. The conclusions drawn from Table 3 are valid, independent of the choice of feature extractor. Within the DINOv2 family, ΔCov. increases with model capacity.

G Note on Earth Mover's Distance

Chamfer distance is widely used in 3D vision, where it measures the discrepancy between two point clouds. Earth Mover's Distance, also called the Wasserstein distance, is the other standard measure from that literature, and the two are commonly reported together. Since we use one-sided Chamfer distance to evaluate generative models, it is natural to ask how well this second classical point-cloud measure would serve the same purpose.

Earth Mover's Distance is symmetric, so $\mathrm{EMD}(P_n, Q_m) = \mathrm{EMD}(Q_m, P_n)$. It therefore returns a single scalar and cannot provide disentangled diagnostics. If the Earth Mover's Distance between the real and generated sets is high, we cannot tell whether this is caused by poor fidelity (meaning that samples look unrealistic) or by poor coverage (meaning that modes of the real distribution are missed). The measure we propose is asymmetric, since $d_{\text{Chamfer}}(P_n, Q_m) \neq d_{\text{Chamfer}}(Q_m, P_n)$ in general. This is precisely what allows us to report fidelity and coverage as two separate axes.

Earth Mover's Distance is also expensive to compute, requiring $O(N^3)$ time and $O(N^2)$ space. This makes it infeasible at the scale of our benchmark, where Table 6 uses 100K samples. By contrast,

Table 12: **Sensitivity to Image Distortions (ImageNet).** Metrics are computed using Inception-v3 [40] features for 5K ImageNet images and their distorted counterparts across increasing latent noise levels. Rows correspond to the standard deviation σ associated with the Gaussian noise added.

	PR		**DC**		**Symmetric**		**Chamfer** *(Ours)*	
σ	Prec.↑	Recall↑	Density↑	Cov.↑	Prec.↑	Recall↑	Fid.↓	Cov.↓
0.10	1.00	1.00	1.13	1.00	1.00	1.00	8.79	8.74
0.125	1.00	1.00	1.11	1.00	1.00	1.00	9.05	8.98
0.15	1.00	1.00	1.08	1.00	1.00	1.00	9.37	9.26
0.175	0.99	1.00	1.04	1.00	0.99	1.00	9.74	9.59
0.20	0.98	1.00	0.99	0.99	0.98	0.99	10.14	9.93
0.225	0.96	1.00	0.93	0.98	0.96	0.98	10.57	10.30

Chamfer distance requires only two cross-set nearest-neighbor lookups, which is much cheaper computationally, as discussed in Appendix J.

H Image Distortions on ImageNet

Sensitivity to Image Distortions We repeat the image distortion experiment of Section 3 on 5K randomly sampled ImageNet images, using Inception-v3 [40] features with $k = 3$. Figure 10 shows representative original-distorted pairs, and Table 12 reports the metrics. KNN-ball based metrics again remain nearly unchanged as the noise level increases, while Chamfer-Fidelity and Chamfer-Coverage increase monotonically with the degradation.

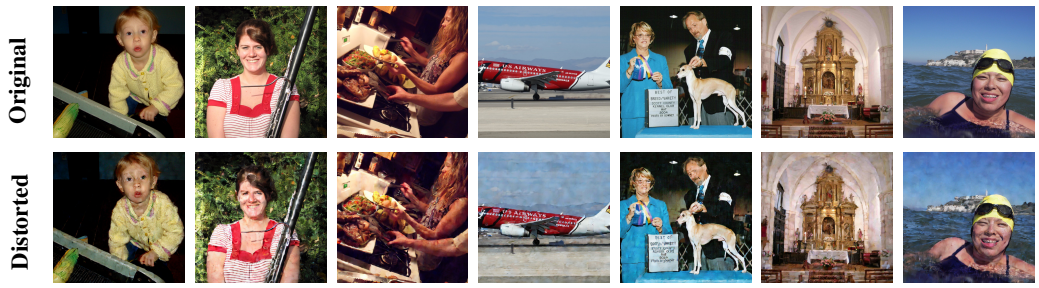


Figure 10: **Original and distorted image pairs.** Each column shows an original ImageNet image (top) and its distorted counterpart obtained by adding element-wise Gaussian noise with mean 0 and standard deviation 0.225 to the VAE latent representation before decoding.

I Benchmarking Modern Generative Models

For Table 6, in order to compute Fréchet metrics (FID and $\text{FD}_{\text{DINOv2}}$), we generate 100K samples from the models and evaluate them against statistics computed from the full ImageNet training set. For the other metrics, namely Precision, Recall, Chamfer-Fidelity and Chamfer-Coverage, we evaluate on 100K generated samples against 100K reference samples, which are randomly drawn from the ImageNet training set. We repeat each evaluation with 3 random seeds, each drawing a new set of 100K generated samples, and report the mean and standard deviation

J Discussion of Computational Complexity

For reference and generated sets of sizes N and M with d-dimensional features, Chamfer distance reduces to two nearest-neighbor searches between the sets, one in each direction. We run these as exact searches with FAISS [9] on the GPU. An exact search takes $O(NMd)$ time, but FAISS computes it as dense matrix multiplications, which GPUs execute at high throughput. FID estimates a $d \times d$ covariance matrix for each set, which takes $O((N + M)d^2)$ time, and then computes a matrix square root, which takes $O(d^3)$ time. The square root does not depend on the number of samples,

which is consistent with FID's runtime in Table 5 changing little between 10K and 50K generated samples (8.36 s and 9.67 s). In every setting of Table 5, Chamfer distance is the fastest of the three metrics. At 100K samples per set, it takes 4.20 s, compared with 25.95 s for FID.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction are supported by the contributions and scope of the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limitations of the work performed by the authors. We mention that we leave human perception, aesthetic quality, semantic alignment evaluation to future work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper fully discloses all the information needed to reproduce the main experimental results of the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide sufficient instructions in the supplemental material to reproduce the main experimental results, and we will release code and data upon publication.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies all the training and test details necessary to understand the results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We do not report error bars for our experiments because it would be too computationally expensive to run multiple trials of the large-scale generative model evaluations we perform.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have provided sufficient information on the computer resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: Not applicable. Since, the paper is focused on evaluation metrics, we do not foresee any direct societal impacts of the work performed.

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Not applicable

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have cited the original papers and datasets used in our paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Not applicable

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Not applicable

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Not applicable

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.