

# A Unifying Information-theoretic Perspective on Evaluating Generative Models

Alexis Fox<sup>1</sup>, Samarth Swarup<sup>2</sup>, Abhijin Adiga<sup>2</sup>

<sup>1</sup>Duke University

<sup>2</sup>University of Virginia

alexis.fox@duke.edu, swarup@virginia.edu, abhijin@virginia.edu

## Abstract

Considering the difficulty of interpreting generative model output, there is significant current research focused on determining meaningful evaluation metrics. Several recent approaches utilize “precision” and “recall,” borrowed from the classification domain, to individually quantify the output *fidelity* (realism) and output *diversity* (representation of the real data variation), respectively. With the increase in metric proposals, there is a need for a unifying perspective, allowing for easier comparison and clearer explanation of their benefits and drawbacks. To this end, we unify a class of  $k$ th-nearest neighbors ( $k$ NN)-based metrics under an information-theoretic lens using approaches from  $k$ NN density estimation. Additionally, we propose a tri-dimensional metric composed of Precision Cross-Entropy (PCE), Recall Cross-Entropy (RCE), and Recall Entropy (RE), which separately measure *fidelity* and two distinct aspects of diversity, *inter-* and *intra-class*. Our domain-agnostic metric, derived from the information-theoretic concepts of entropy and cross-entropy, can be dissected for both sample- and mode-level analysis. Our detailed experimental results demonstrate the sensitivity of our metric components to their respective qualities and reveal undesirable behaviors of other metrics.

**Code** — <https://github.com/NSSAC/PrecisionRecallMetric>

**Extended version** — <https://arxiv.org/abs/2412.14340>

## 1 Introduction

**Need for Metrics.** Deep generative models, including Generative Adversarial Networks (GANs) (Goodfellow et al. 2014), Variational Autoencoders (VAEs) (Kingma and Welling 2022), and diffusion models (Sohl-Dickstein et al. 2015), are achieving unprecedented realism in their outputs (Ravuri et al. 2023). Their success highlights an urgent need for robust methods to evaluate their output quality. Unlike discriminative models, which are evaluated against a clear ground truth in the form of labeled test sets, generative models produce a spectrum of plausible outputs based on the learned data distribution. Hence, evaluating generative models is a nontrivial task.

**Traditional 1D Metrics.** Often, an important goal of generative models is to produce outputs that human observers

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

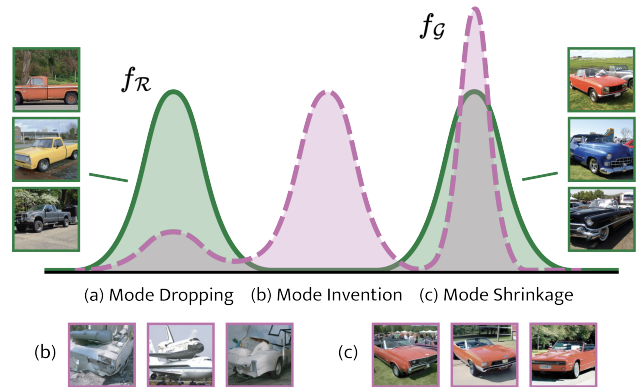


Figure 1: We visualize the various failure modes (a)-(c) for a model distribution, where the real (green) distribution is composed of two modes: sport cars and pickup trucks. (a) represents a lack of generated pickup trucks, (b) and its corresponding examples show unrealistic points, and (c) depicts points clustered at the mode average.

find realistic (Stein et al. 2024). While human assessment can be quantified, manual human evaluation is expensive, time-consuming, and makes continuous assessments impractical. Automatic evaluation metrics such as Inception Score ( $IS$ ) (Salimans et al. 2016) and Fréchet Inception Distance ( $FID$ ) (Heusel et al. 2017) address these inefficiencies by computing the distributional differences of generated and real content.

**Motivation for 2D Metrics.** However, as one-dimensional scores,  $IS$  and  $FID$  cannot distinguish various shortcomings, e.g. when outputs are generated in low-probability regions of the real distribution (low fidelity) or insufficiently represent high-probability regions (low diversity) (Sajjadi et al. 2018). A recent set of two-dimensional metrics, which can be broadly categorized as versions of “precision” and “recall,” has emerged to differentiate the several types of failure modes. Sajjadi et al. (2018) introduce the first of a series of precision and recall metrics based on decomposing the real and generated distributions into common and unique components. Subsequent works propose alternatives to precision and recall, with most utilizing  $k$ -Nearest Neighbor ( $k$ NN) methods (Kynkäänniemi et al. 2019; Naeem et al. 2020; Alaa

et al. 2022; Cheema and Urner 2023; Park and Kim 2023).

**The Gap.** Despite the proliferation of evaluation methods, there remains a need for clearer theoretical comparisons and validation of these metrics. We address this gap by providing probabilistic derivations with information-theoretic interpretations for a subset of metrics, focusing on Precision Recall Cover (Cheema and Urner 2023) and Density & Coverage (Naeem et al. 2020).

We also introduce a simple new information-theoretic metric that, unlike existing metrics, simultaneously distinguishes between the three failure modes visualized in Figure 1. Mode invention (1b) indicates a loss of precision (*fidelity*), occurring when the model generates implausible outputs that do not align with the real distribution. Mode dropping (1a) represents one type of recall (*diversity*) loss, where the model lacks sufficient coverage of certain real regions. Another recall issue, mode shrinkage (1c), happens when a model repeatedly produces highly similar points—lacking *intra-class diversity*—that represent a mode “average.” We summarize our three main contributions below.

**Unifying Existing Metrics.** In Section 4, we provide information-theoretic derivations for Precision Recall Cover and Density & Coverage. By extending lemmas from Noshad et al. (2017), we derive a bias term for the expected value of these metrics, allowing for easier interpretation and comparison and motivating our metric.

**Novel Information-Theoretic Metric.** We propose a new metric for generative models, composed of Precision Cross-Entropy (*PCE*), Recall Cross-Entropy (*RCE*), and Recall Entropy (*RE*), based on the estimator from Leonenko, Pronzato, and Savani (2008). Unlike previous work, we separate the two types of diversity loss. Section 5 details the definition of these scores as well as how they discern the various failure modes mentioned. We also describe how our population-level metric can be dissected for both mode- and sample-level analysis, useful for diagnostic purposes.

**Experimental Results.** We list desiderata for generative model metrics in Section 5.1. Through our experimental results in Section 6, we show how our metrics align with these ideals, as well as how other existing metrics are less effective with respect to one or more of these. Specifically, we analyze the sensitivity of the measures to the different failure modes mentioned above, as well as examine their correlation with human scoring. As with most prior work, we focus on image evaluation, but our methods can be easily generalized to other data modalities, as seen in the extended version’s Appendix.

## 2 Background and Related Work

Several metrics have been proposed to quantify the performance of generative models. Below we summarize relevant metrics and their respective benefits and shortcomings. We start with common background information.

**Background.** Calculating distances in raw input spaces, such as image pixels, is often impractical due to the high dimensionality and sensitivity to noise. To address this, many metrics — including ours — embed data into a feature space that provides a more semantic representation, where Euclidean distances correlate better with human-perceived similarity. While expensive, human perception can be used as a

“ground truth” for other metrics for realism. One benchmark, Human eYe Perceptual Evaluation (*HYPE*), measures human error rate when identifying fake and real images, serving as a proxy for realism. (Zhou et al. 2019).

**Classic Metrics.** Inception Score and log-likelihood have historically been popular metrics but are criticized for poor correlation with human judgment. (Theis, Oord, and Bethge 2015; Kolchinski et al. 2019). Fréchet Inception Distance (*FID*), which we later compare experimentally, is another widely used metric for evaluating generated images. *FID* evaluates images by comparing the Fréchet distance of generated and real images in an Inception-based feature space, modeled as multivariate Gaussians (Heusel et al. 2017). We use the generalized abbreviation *FD* to refer to the Fréchet distance metric hereafter, as justified in Section 6. *FD* is consistent with human judgment (Stein et al. 2024); however, it assumes Gaussian-distributed features, which may not always hold (Borji 2022). Additionally, as a one-dimensional metric, *FD* cannot distinguish failure modes, which is a motivation for the following multi-dimensional metrics.

**Early Notion of Precision and Recall.** Sajjadi et al. (2018) introduce the novel Precision-Recall for Distributions (*PRD*) metric, which compares embedded samples from the real and generated distribution, using clustering assignments to assess whether the differences in their common support should be attributed to precision (accuracy) or recall (diversity) loss. Recognizing the limitations of the *PRD* metric, which relies on relative densities and yields a continuum of values along a PR-curve, Kynkäänniemi et al. (2019) propose an Improved version of Precision and Recall (*IP & IR*) leveraging *k*NN distances of samples embedded in a feature space. *IP* measures precision by assessing if each generated image is within the real image manifold, estimated with the union of *k*NN balls centered on each image. Symmetrically, *IR* evaluates if each real image is within the generated image manifold.

**Recent Extensions.** Several subsequent works can be seen as extensions of the approaches by Sajjadi et al. (2018) and Kynkäänniemi et al. (2019). While superficially different, most of these build upon common underlying computational ideas. We focus in particular on the work of Cheema and Urner (2023), who propose Precision Coverage and Recall Coverage (*PC & RC*, collectively defined as *PRC*), which differs from previous works by parameterizing what is considered “sufficient” coverage, and on the work of Naeem et al. (2020), who observed that *IP & IR* are sensitive to outliers and fail to identify identical distributions. They developed the more robust Density and Coverage measures (*D & C*). We focus on these two because they are either recent or have particular limitations (on which, more in Section 4). However, other recent approaches can be seen in a similar information-theoretic light (Alaa et al. 2022; Djolonga et al. 2020; Liu et al. 2021), which we comment upon briefly in the discussion. Outside the precision and recall framework, the Rarity Score (Han et al. 2022) quantifies the uniqueness of individual generated samples, which, by comparing histograms, can be implicitly seen as an intra-class diversity measure.

We note that no metric discussed above simultaneously distinguishes between inter- and intra-class diversity, which is a strength of our metric detailed in Section 5.

### 3 Preliminaries

Given a (multivariate) random variable  $\mathcal{X}$  with distribution  $f_{\mathcal{X}}$ , let  $X = \{X_1, \dots, X_{N_X}\}$  be a set of  $N_X$  i.i.d. samples of  $\mathcal{X}$ , with each  $X_i \in \mathbb{R}^d$ . In this work, we will frequently compare two distributions, letting  $f_{\mathcal{X}}$  and  $f_{\mathcal{Y}}$  represent two general distributions, while  $f_{\mathcal{R}}$  and  $f_{\mathcal{G}}$  specifically represent the real and generated distributions, respectively. Given sets of samples  $X$  and  $Y$  of  $\mathcal{X}$  and  $\mathcal{Y}$ ,  $Z$  is the combined set of points from  $X$  and  $Y$ , denoted by  $Z = X \cup Y$ .

#### 3.1 Information Theoretic Concepts

For completeness, we present some foundational concepts from information theory, which we will later integrate into the analysis of generative metrics. **Shannon entropy** of a random variable  $\mathcal{X}$  is defined as  $H(\mathcal{X}) = \mathbb{E}_{x \sim \mathcal{X}} \log \frac{1}{f_{\mathcal{X}}(x)}$  while **cross-entropy** of  $\mathcal{X}$  relative to  $\mathcal{Y}$  is  $CE(\mathcal{X}, \mathcal{Y}) = \mathbb{E}_{x \sim \mathcal{X}} \log \frac{1}{f_{\mathcal{Y}}(x)}$ . **KL-divergence** of  $\mathcal{X}$  from  $\mathcal{Y}$  is defined as  $D_{KL}(\mathcal{X} \parallel \mathcal{Y}) = \mathbb{E}_{x \sim \mathcal{X}} \log \frac{f_{\mathcal{X}}(x)}{f_{\mathcal{Y}}(x)}$ .

**Rényi divergence**, parameterized by order  $\alpha$  ( $\alpha \neq 1$  and  $\alpha > 0$ ), is a generalization of KL-divergence that similarly quantifies the informational difference between two distributions (Rényi 1961). The formal definition is as follows:

$$D_{\alpha}(\mathcal{X} \parallel \mathcal{Y}) = \frac{1}{\alpha - 1} \log \int f_{\mathcal{X}}(x)^{\alpha} f_{\mathcal{Y}}(x)^{1-\alpha} dx \quad (1)$$

$$= \frac{1}{\alpha - 1} \log J_{\alpha}(f_{\mathcal{X}}, f_{\mathcal{Y}}), \quad (2)$$

where  $J_{\alpha}(\mathcal{X}, \mathcal{Y}) = \mathbb{E}_{x \sim \mathcal{Y}} \left[ \left( \frac{f_{\mathcal{X}}(x)}{f_{\mathcal{Y}}(x)} \right)^{\alpha} \right]$  (Noshad et al. 2017). As  $\alpha \rightarrow 1$ , Rényi divergence approaches the KL-divergence. We note that  $\alpha = 1$  is defined for  $J_{\alpha}(\mathcal{X}, \mathcal{Y})$ .

#### 3.2 $k$ NN-based Estimators

In practical applications, the full distribution of  $\mathcal{X}$  or  $\mathcal{Y}$  is often not known, or the integrals required to compute measures such as entropy or divergence are intractable. As a result, direct computation of these measures is not feasible, necessitating the estimation of these values from available data. Multiple methods (Noshad et al. 2017; Póczos and Schneider 2011; Leonenko, Pronzato, and Savani 2008) have been proposed that utilize the  $k$ th-nearest neighbors of sample points to estimate probability density ratios. Below we introduce notation and definitions for such estimators.

**$k$ NN Density Estimation.** Let  $B_{k,X}(X_i)$  be a hypersphere in  $\mathbb{R}^d$  centered at the point  $X_i$ , with its radius determined by the distance to the  $k$ th-nearest neighbor in the set  $X$ . Both the radius and number of points within a  $k$ NN ball can be used for density estimation. We define  $\#(Y, B_{k,X}(X_i))$  as the number of points from set  $Y$  found within the ball  $B_{k,X}(X_i)$ . To accommodate the distance-based nature of other estimators, we define  $D_{k,Y}(X_i)$  as the distance from  $X_i$  to its  $k$ th-nearest neighbor in the set  $Y$ .

**Entropy and Cross-Entropy Estimator.** Leonenko, Pronzato, and Savani (2008) propose direct, nonparametric estimators for cross-entropy and entropy. These estimators leverage  $k$ NN distances as a proxy for probability density.

The estimators for entropy,  $\check{H}_k(X)$ , and cross-entropy,  $\check{CE}_k(X, Y)$ , are defined below:

$$\check{H}_k(X) = \frac{1}{N_X} \sum_{i=1}^{N_X} \log \left( (N_X - 1) e^{-\Psi(k)} \bar{c}(D_{k,X}(X_i))^d \right) \quad (3)$$

$$\check{CE}_k(X, Y) = \frac{1}{N_X} \sum_{i=1}^{N_X} \log \left( N_Y e^{-\Psi(k)} \bar{c}(D_{k,Y}(X_i))^d \right). \quad (4)$$

Here,  $\Psi(z) = \Gamma'(z)/\Gamma(z)$  represents the digamma function and  $\bar{c}$  is the volume of a unit ball in  $\mathbb{R}^d$ .

**Divergence Estimator.** Noshad et al. (2017) define an estimator for Rényi divergence utilizing nearest neighbor ratios. Their method calculates the ratio of the number of points from  $X$  to that of  $Y$  found within the hypersphere,  $B_{k,Z}(Y_i)$ , to estimate  $\frac{f_{\mathcal{X}}(x)}{f_{\mathcal{Y}}(x)}$ . Noshad et al. (2017) derive the following estimators for  $J_{\alpha}(\mathcal{X}, \mathcal{Y})$  and  $D_{\alpha}(\mathcal{X} \parallel \mathcal{Y})$ , with  $\eta = N_Y/N_X$ :

$$\hat{J}_{\alpha}(X, Y) = \frac{\eta^{\alpha}}{N_Y} \sum_{i=1}^{N_Y} \left( \frac{\#(X, B_{k,Z}(Y_i))}{\#(Y, B_{k,Z}(Y_i)) + 1} \right)^{\alpha} \quad (5)$$

$$\tilde{D}_{\alpha}(X, Y) = \frac{1}{(\alpha - 1)} \log \hat{J}_{\alpha}(X, Y). \quad (6)$$

## 4 Unifying Information Theoretic Perspective

Below we examine the formulae for two metrics—Precision Recall Cover and Density & Coverage—in order to express them within a unified information theoretic perspective. We then use our results to motivate our metric.

### 4.1 Precision Recall Cover

Cheema and Uerner (2023) propose the Precision Recall Cover (PRC) metric, consisting of Precision Coverage (PC) and Recall Coverage (RC). For their population-level analogue, they define two parameters,  $\alpha$  and  $\beta$ , which allow for flexibility in determining what is considered “sufficiently” covered for both generated and real regions. Cheema and Uerner (2023) additionally provide a  $k$ NN-based empirical definition, which we describe below.

**Empirical  $(k, k')$ -PRC Definition.** For two positive integers  $k = Ck'$  where  $C$  is a positive integer as well, let  $\alpha = k'/N_G$  and  $\beta = k/N_R$ . The  $(k, k')$ -Precision Coverage of sample set  $R$  with respect to  $G$  is defined by constructing  $k$ - and  $k'$ -nearest neighbor balls over the sample sets:

$$PC_{k,k'}(R, G) = \frac{1}{N_G} \sum_{i=1}^{N_G} \mathbf{1}[\#(R, B_{k,G}(G_i)) \geq k'], \quad (7)$$

where  $\mathbf{1}[\cdot]$  is the indicator variable of the specified event. This method calculates the proportion of  $k$ NN balls in  $G$  that contain at least  $k'$  points from  $R$ , normalized by the total points in  $G$ .  $(k, k')$ -Recall Coverage is defined similarly:

$$RC_{k,k'}(R, G) = \frac{1}{N_R} \sum_{i=1}^{N_R} \mathbf{1}[\#(G, B_{k,R}(R_i)) \geq k']. \quad (8)$$

**Connecting PRC.** To set up its definition to be compatible with our perspective, we rearrange the expression of  $PC$ . Results for  $RC$  generalize symmetrically. By definition of  $\#(\cdot)$ , the quantity  $\#(G, B_{k,G}(G_i))$  refers to the number of points from  $G$  within the  $k$ -nearest neighborhood of  $G_i$ , which is exactly  $k$ . The term in Equation 7 can be expressed as:

$$\#(R, B_{k,G}(G_i)) \geq k' \Rightarrow \frac{\#(R, B_{k,G}(G_i))}{\#(G, B_{k,G}(G_i))} \geq \frac{1}{C}. \quad (9)$$

Therefore, noting that  $\eta = \frac{N_G}{N_R}$ ,

$$PC_{k,k'}(R, G) = \frac{1}{N_G} \sum_{i=1}^{N_G} \mathbf{1} \left[ \frac{\eta \cdot \#(R, B_{k,G}(G_i))}{\#(G, B_{k,G}(G_i))} \geq \frac{\eta}{C} \right] \quad (10)$$

By (5), with  $\alpha = 1$  and  $U = R \cup G$ :

$$\widehat{J}_1(R, G) = \frac{1}{N_G} \sum_{i=1}^{N_G} \frac{\eta \cdot \#(R, B_{k,U}(G_i))}{\#(G, B_{k,U}(G_i)) + 1}. \quad (11)$$

We define an indicator function version of the above and its corresponding empirical version as:

$$\widehat{J}'_1(R, G) = \frac{1}{N_G} \sum_{i=1}^{N_G} \mathbf{1} \left[ \frac{\eta \cdot \#(R, B_{k,U}(G_i))}{\#(G, B_{k,U}(G_i)) + 1} \geq \frac{\eta}{C} \right] \quad (12)$$

$$J'_1(\mathcal{R}, \mathcal{G}) = \mathbb{E}_{x \sim \mathcal{G}} \left[ \mathbf{1} \left[ \frac{f_{\mathcal{R}}(x)}{f_{\mathcal{G}}(x)} \geq \frac{\eta}{C} \right] \right]. \quad (13)$$

Since the indicator function is not Lipschitz continuous, as required by Lemma 3.2 of Noshad et al. (2017), we define a close sigmoid approximation, where  $\sigma(x; \theta) = \frac{1}{1 + e^{-\theta(x - \frac{\mathcal{G}}{2})}}$ , with  $\theta$  controlling the steepness of the sigmoid (higher  $\theta$  approaching closer to the indicator function). As seen above, the empirical definition of  $PC$  (Eqn. 10) and the estimator  $\widehat{J}'_1$  (Eqn. 12) differ in only two aspects:  $PC$  calculates the nearest neighbors within the set  $G$  instead of  $U$ , and  $PC$  has no added 1 in the denominator. In the Appendix, we show how the described differences affect the bias. We also define an information theoretic intuition for  $PC$  through rearrangement of Eqn. 13 and defining  $C_2 = \log(\frac{C}{\eta})$ :

$$J'_1(\mathcal{R}, \mathcal{G}) = \mathbb{E}_{x \sim \mathcal{G}} \left[ \mathbf{1} \left[ \log \left( \frac{f_{\mathcal{G}}(x)}{f_{\mathcal{R}}(x)} \right) \leq C_2 \right] \right]. \quad (14)$$

The interior expression checks if the number of extra bits needed to encode the generated data point using the optimal real encoding scheme instead of generated is no greater than  $C_2$ . By symmetry, we can relate Recall Coverage to  $J'_1(\mathcal{G}, \mathcal{R})$ .

## 4.2 Density & Coverage

Naeem et al. (2020) define their metric, called Density (not to be confused with probability density) & Coverage, with the motivation of building upon the work of Kynkäänniemi et al. (2019), by creating a version of precision & recall more resistant to outliers. Their definitions are described below.

**Density & Coverage Definition.** Density, used to measure the fidelity of generated samples, can be defined as:

$$\text{Density} = \frac{1}{N_G} \sum_{i=1}^{N_R} \frac{\#(G, B_{k,R}(R_i))}{k}. \quad (15)$$

Density counts the number of neighborhood spheres of real samples that contain the generated data point, normalized by  $N_G$ . Their recall counterpart, Coverage, can be defined as:

$$\text{Coverage} = \frac{1}{N_R} \sum_{i=1}^{N_R} \mathbf{1}[\#(G, B_{k,R}(R_i)) \geq 1]. \quad (16)$$

Naeem et al. (2020) base their recall measure on the real sample set, justified by its lower outlier presence compared to generated sets." Coverage averages the number of real  $k$ NN balls that contain one or more generated samples.

**Connecting Density.** We rearrange the expression of Density to define it within our framework. Noting that  $k = \#(R, B_{k,R}(R_i))$ ,

$$\text{Density} = \frac{1}{N_G} \sum_{i=1}^{N_R} \frac{\#(G, B_{k,R}(R_i))}{\#(R, B_{k,R}(R_i))}. \quad (17)$$

We then define  $\widehat{J}_1(G, R)$  and its theoretical value  $J_1(\mathcal{G}, \mathcal{R})$ :

$$\widehat{J}_1(G, R) = \frac{1}{N_G} \sum_{i=1}^{N_R} \frac{\#(G, B_{k,U}(R_i))}{\#(R, B_{k,U}(R_i)) + 1} \quad (18)$$

$$J_1(\mathcal{G}, \mathcal{R}) = \mathbb{E}_{x \sim \mathcal{R}} \left[ \frac{f_{\mathcal{G}}(x)}{f_{\mathcal{R}}(x)} \right] = 1. \quad (19)$$

The difference between Density and  $\widehat{J}_1(G, R)$  is the added 1 in the denominator, a shared difference with that of  $PC$  and  $\widehat{J}'_1(R, G)$ . We highlight the fact that  $J_1(\mathcal{G}, \mathcal{R}) = 1$ , which implies that as  $N_G$  increases, and with  $k$  a growing function of  $N_G$  (Noshad et al. 2017), Density and  $\widehat{J}_1(G, R)$  are both asymptotically 1. Because of differing convergence rates, we note that Density can still be a meaningful measure when comparing output with the same number of drawn samples. However, as our Appendix experiments show, Density is an unreliable measure when working with an increasingly large number of samples.

**Intuition.** Here, we describe the intuition behind why the Density metric is asymptotically 1 under appropriate conditions. For clarity, we present definitions of  $J_1(\mathcal{G}, \mathcal{R})$  and its estimator  $\widehat{J}_1(G, R)$  that are simply rearrangements of their formulae in Section 3:

$$J_1(\mathcal{G}, \mathcal{R}) = \int f_{\mathcal{R}}(x) \frac{f_{\mathcal{G}}(x)}{f_{\mathcal{R}}(x)} dx = 1 \quad (20)$$

$$\widehat{J}_1(G, R) = \frac{1}{N_R} \sum_{i=1}^{N_R} \frac{N_R \cdot \#(G, B_{k,U}(R_i))}{N_G \cdot (\#(R, B_{k,U}(R_i)) + 1)}. \quad (21)$$

Monte Carlo integration  $\frac{1}{N_X} \sum_{i=1}^{N_X} f_Y(X_i)$  serves as an unbiased estimator for  $\int f_Y(x) f_X(x) dx$ . Applying this principle,  $\frac{1}{N_R} \sum_{i=1}^{N_R} \frac{f_{\mathcal{G}}(R_i)}{f_{\mathcal{R}}(R_i)}$  is an unbiased estimator of  $J_1(\mathcal{G}, \mathcal{R})$  (Eqn. 20). Dissecting  $\widehat{J}_1(G, R)$ , the average  $\frac{1}{N_R} \sum_{i=1}^{N_R}$  approximates  $f_{\mathcal{R}}(x)$ , while the scaled ratio of  $k$ NN points approximates the probability density ratio  $\frac{f_{\mathcal{G}}(R_i)}{f_{\mathcal{R}}(R_i)}$ . Combining these two ideas,  $\widehat{J}_1(G, R)$  can be constructed to estimate  $J_1(\mathcal{G}, \mathcal{R})$ ,

which is equal to 1. We therefore affirm that  $\widehat{J}_1(G, R)$  and density asymptotically approach 1.

**Connecting Coverage.** We observe that Coverage is a special case of Recall Coverage (see the empirical  $(k, k')$ -*PRC* definition earlier and Eqn. 8), where  $C = k$  and  $k' = 1$ . Hence, leveraging  $k = \#(R, B_{k,R}(R_i))$ , we can define Coverage as:

$$\text{Coverage} = \frac{1}{N_R} \sum_{i=1}^{N_R} \mathbf{1} \left[ \frac{\#(G, B_{k,R}(R_i))}{\#(R, B_{k,R}(R_i))} \geq \frac{1}{k} \right]. \quad (22)$$

Thus, Coverage is also an estimator for  $J'_1(\mathcal{G}, \mathcal{R})$ .

### 4.3 Implication

As demonstrated from our mathematical analysis, these metrics approximate the general form of a statistical divergence  $D(\mathcal{X}||\mathcal{Y}) = \mathbb{E}_{x \sim \mathcal{X}} \left[ g \left( \frac{f_{\mathcal{Y}}(x)}{f_{\mathcal{X}}(x)} \right) \right]$ , where *PC*, *RC*, and *Cov* use  $g(z) = \mathbf{1}[z \geq \frac{1}{C}]$ , *Density* uses  $g(z) = z$ , and *KL-divergence* is  $g(z) = -\log(z)$ . We observe that in the context of precision, the form  $D(\mathcal{G}||\mathcal{R})$  is ineffective due to its dependence on relative densities. For instance,  $J'_1(\mathcal{G}, \mathcal{R})$  (related to *PC*) includes a  $f_{\mathcal{G}}(x)$  term in the denominator, which unfairly punishes for when the generated region is dense — an aspect which should be of neutral relevance. In the context of recall, the form  $D(\mathcal{R}||\mathcal{G})$  is more reliable. Regions with high probability in the real distribution should correspondingly require dense coverage by generated samples; conversely, dense generated regions should not necessitate denser real regions to be considered precise.

We substantiate our observations, noting  $D_{KL}(\mathcal{G}||\mathcal{R}) = CE(\mathcal{G}, \mathcal{R}) - H(\mathcal{G})$ , by comparing the closed-form expression for all three terms when increasing the variance of  $\mathcal{G}$ . We analyze two multivariate Gaussian distributions,  $\mathcal{R} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  and  $\mathcal{G} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$  in  $\mathbb{R}^{10}$ , where  $\mathbf{0}$  represents the zero mean vector and  $\mathbf{I}$  denotes the identity matrix.

$\sigma^2$	$D_{KL}(\mathcal{G}  \mathcal{R})$	$CE(\mathcal{G}, \mathcal{R})$	$H(\mathcal{G})$
0.25	3.18	10.44	7.26
1	0.00	14.19	14.19
2.5	2.92	21.69	18.77

Table 1: KL-divergence  $D_{KL}(\mathcal{G}||\mathcal{R})$ , cross-entropy  $CE(\mathcal{G}, \mathcal{R})$ , and entropy  $H(\mathcal{G})$  for varying values of  $\sigma^2$ , the variance of the generated distribution of  $\mathcal{G}$ .

Precision measures should observe decreased fidelity as  $\mathcal{G}$  increases its variance relative to that of  $\mathcal{R}$ . Notably, the  $D_{KL}$  values at  $\sigma^2 = 0.25$  and 2.5 are approximately symmetric around the  $\sigma^2 = 1$  case, illustrating how both underestimation and overestimation of variance relative to the true model contribute comparably to the divergence. We highlight that the basis of our metric defined in Section 5,  $CE(\mathcal{G}, \mathcal{R})$ , trends in the right direction. These insights motivate the construction of our asymmetric cross-entropy measures.

## 5 Information-theoretic Precision and Recall Metric

Here, we introduce the three components of our comprehensive metric: Precision Cross-Entropy (*PCE*), Recall Cross-Entropy (*RCE*), and Recall Entropy (*RE*).

### 5.1 Desiderata for Generative Model Metrics

To clarify the optimal behavior of generative model metrics, we first present the following desiderata for precision and recall measures.

*Precision* should (P1) reflect high fidelity for generated samples that are likely within the real distribution, (P2) indicate low fidelity for those that are improbable, highlighting cases of mode invention (Borji 2018). For certain domains such as image generation, where realism is effectively judged by human evaluation, precision should (P3) align with human evaluation (Stein et al. 2024).

*Recall* should (R1) reward the case where high-probability regions of the real distribution are sufficiently covered by generated samples, addressing mode dropping, and where (R2) the generated samples capture the full variation within each mode, limiting mode shrinkage.

### 5.2 Theoretical Metric Definitions

We exploit the asymmetric nature of (cross-)entropy, as defined in Section 3.1, to construct an information-theoretic interpretation of precision and recall. Below, we define our metric and justify its use in accordance with the labeled desiderata in Section 5.1.

The three components of our metric are defined below. Note that for all components, the entropy of the real data distribution,  $H(f_{\mathcal{R}})$ , serves as a constant baseline, so that for identical distributions,  $PCE = RCE = RE = 0$ . Because  $H(f_{\mathcal{R}})$  is independent of the generated distribution, the quality of the metric scores does not depend on its value.

#### Precision Cross-Entropy (*PCE*):

$$\begin{aligned} PCE &= CE(\mathcal{G}, \mathcal{R}) - H(\mathcal{R}) \\ &= \mathbb{E}_{x \sim \mathcal{G}} \log \frac{1}{f_{\mathcal{R}}(x)} - \mathbb{E}_{x \sim \mathcal{R}} \log \frac{1}{f_{\mathcal{R}}(x)}. \end{aligned} \quad (23)$$

*PCE* is minimized when generated points are in high-probability regions of the real manifold (P1). Lower values of *PCE* correspond to higher levels of fidelity and less mode invention (P2).

#### Recall Cross-Entropy (*RCE*):

$$RCE = CE(\mathcal{R}, \mathcal{G}) - H(\mathcal{R}) = \mathbb{E}_{x \sim \mathcal{R}} \log \frac{f_{\mathcal{R}}(x)}{f_{\mathcal{G}}(x)}. \quad (24)$$

Minimizing *RCE* requires that the regions where the real probability is large also exhibit large values of the generated probability (R1). Low values of *RCE* indicates high inter-class diversity and therefore limited mode dropping. Note that *RCE* is equivalent to  $D_{KL}(\mathcal{R}||\mathcal{G})$ .

#### Recall Entropy (*RE*):

$$\begin{aligned} RE &= H(\mathcal{G}) - H(\mathcal{R}) \\ &= \mathbb{E}_{x \sim \mathcal{G}} \log \frac{1}{f_{\mathcal{G}}(x)} - \mathbb{E}_{x \sim \mathcal{R}} \log \frac{1}{f_{\mathcal{R}}(x)}. \end{aligned} \quad (25)$$

The more similar generated points are to each other (i.e. *lower uniqueness*), the *lower RE* is. Low *RE* coupled with low *PCE* indicates mode shrinkage, as points clustered at the average of a mode are highly precise, but not unique (R2). *RE* can be seen as the intra-class diversity complement to *RCE*. *RE* and *PCE* are normalized versions of the two components of  $D_{KL}(\mathcal{G}||\mathcal{R}) = CE(\mathcal{G}, \mathcal{R}) - H(\mathcal{G})$ .

### 5.3 Empirical Definitions

For calculations on sample sets, we refer back to the *k*NN-based estimators defined in Section 3.2 (see Eqn. 3 and 4) and define the empirical version of our metric as:

$$PCE_k = \check{C}\check{E}_k(G, R) - \check{H}_k(R) \quad (26)$$

$$RCE_k = \check{C}\check{E}_k(R, G) - \check{H}_k(R) \quad (27)$$

$$RE_k = \check{H}_k(G) - \check{H}_k(R). \quad (28)$$

From a high-level perspective,  $PCE_k$  calculates the average log-normalized distance to the *k*th-nearest neighbor in the real set for all generated samples. Similarly,  $RCE_k$  is minimized when the average normalized distance of real points to their generated neighbors is small. Conversely,  $RE_k$  is sensitive to when generated points are clustered together. In this empirical context,  $\check{H}_k(R)$  can also be seen as a normalization factor dependent on the dimension and choice of *k* (Park and Kim 2023). These scores leverage the same principle as other *k*NN metrics and *FD*; distances in a feature space convey meaningful information, with shorter distances signifying greater similarity between points.

Unlike non-*k*NN-based metrics, such as *FD*, our scores can be dissected for mode- and sample-level analysis, as the estimators are additive in the contributions of different data points. Hence, we can see which samples contribute the most to the overall score and what modes the generative model is struggling to represent. We note that in cases where  $PCE_k$  is significantly negative for a sample, memorization—when the model generates points (near-)identical to a training sample—may be present. Examples are provided in the Appendix.

Because of its favorable sample-level nature, our method can be easily related to other works. For instance, generated samples can be sorted by contribution to  $PCE_k$ , and the highest scoring can be discarded; this is a form of model-auditing (Alaa et al. 2022), where samples are judged for quality and the most unrealistic samples are rejected. We can also extend our method to Precision-Recall curves (Djolonga et al. 2020), where we can visualize a trade-off between  $PCE_k$  and  $RE_k$  when auditing levels are increased. When the less realistic samples, which likely reside in the fringes of modes, are removed, intra-class diversity decreases while fidelity increases.

## 6 Experiments

Here we analyze the empirical behavior of Density & Coverage (*D* & *C*), Precision Recall Cover ( $PC_{k,k'}$  &  $RC_{k,k'}$ ), Fréchet distance (*FD*), and our metric ( $PCE_k$ ,  $RCE_k$ , &  $RE_k$ ), specifically focusing on how well they align with the desiderata listed in Section 5.1. We set *k* to 5, *k'* = 15, and omit the subscripts from the metric abbreviations for

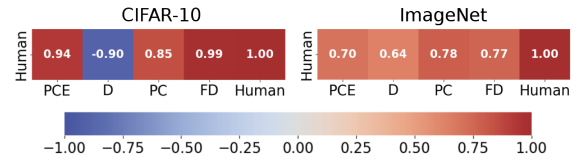


Figure 2: Pearson correlation heatmap between the precision measures and human error rate, displayed separately for the CIFAR-10 and ImageNet datasets. The direction of the scores where higher values indicate poorer precision was flipped so that all metric scores rank in the same direction.

brevity hereafter. We follow the recommendation of Stein et al. (2024) to embed the images with the DINOv2-ViT-L/14 encoder (Oquab et al. 2024), which they claim provides a richer representation space than the commonly used Inception network, which may unfairly punish diffusion models. This motivates our generalized abbreviation *FD*.

**Dataset Descriptions.** We use both ImageNet (Deng et al. 2009) and CIFAR-10 (Krizhevsky, Hinton et al. 2009) image datasets for our analysis. The sampled training set for ImageNet contains 1000 classes with 100 images each, while CIFAR-10 has 10 classes with 4500 images each — more information about dataset usage is available in the Appendix.

**Correlation with Human Rankings.** Although there is no clear “ground truth” for overall grading of generative output, human assessment can be seen as the gold standard for the realism component of metrics. Stein et al. (2024) recently evaluated human perception for a multitude of models, using a similar method to the HYPE metric mentioned in Section 2, creating a meaningful baseline for realism scores. Using their recommended DINO-v2 space, they find that only *FD* correlates well with human assessment, with their experiments including density and improved precision (Kynkäänniemi et al. 2019). We extend their experiments on the CIFAR10 and the more complex ImageNet datasets, adding *PC* and our precision measure *PCE* to the analysis. As seen in Figure 2, we find that *PCE*, *PC*, and *FD* all have a reasonably high and similar ( $\geq 0.7$ , all within a range of 0.08 units) positive correlation with human error rate for both datasets. Although Density (*D*) shows a positive correlation on the ImageNet dataset, it shows a heavy negative score on CIFAR-10. This inconsistent relationship with human error rate highlights a significant weakness of the Density measure when measuring realism on practical datasets.

**Mode Shrinkage Experiment.** The classifier-free guidance (CFG) parameter (Ho and Salimans 2022) allows control over trading off intra-class diversity for fidelity. Increasing CFG leads to datasets with worse mode shrinkage but higher precision, similar in effect to Figure 1(c). Adapting the diversity experiment from Stein et al. (2024), we calculate the metrics for image sets generated at five levels of the CFG parameter, steadily decreasing the intra-class diversity of the generated output, as shown in Figure 3. The DiT-XL-2 model, trained on ImageNet, was used for generation. With increasing CFG, precision metrics should show increased fidelity, while recall—specifically *RE* for our metric—should show decreased diversity. However, we find that *PC* decreases for

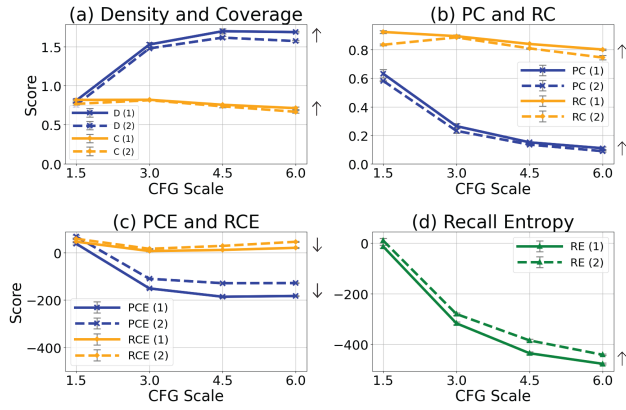


Figure 3: Behavior of metrics at different CFG values for the generation of samples using the DiT-XL-2 model trained on ImageNet. Two sets of 15 random classes were used (as marked by (1) and (2)), with six runs each performed and averaged. Error bars ( $\pm 1$  SD) are plotted but are visibly negligible. Arrows are put aside each measure’s lines - upright arrows mean that higher values indicate better fidelity or diversity (opposite for the downward arrows). See sample images in the Appendix and Figure 1(c).

higher CFG values, while  $D$  and  $PCE$  perform as expected. The recall metrics ( $C$ ,  $RC$ ,  $RCE$ ) see only a slight change - only our intra-class diversity metric ( $RE$ ) shows significant sensitivity to mode shrinkage.

**Mode Dropping Experiment.** To complement our intra-class diversity test, we conduct an inter-class diversity experiment. We start with the original 1000 classes from ImageNet, with 100 images generated per class by the ADMG-ADMU (Dhariwal and Nichol 2021) model. As seen in Figure 4, 100 classes were dropped at a time until 100 classes remained, for a total of 10 intervals. Ideally, mode dropping metrics ( $C$ ,  $RC$ ,  $RCE$ ) should show significant change, while precision and mode shrinkage measures ( $D$ ,  $PC$ ,  $PCE$ ,  $RE$ ) should remain relatively constant. We find that to be the case. Figure 3(d) shows that the one-dimensional  $FD$  cannot distinguish diversity loss from precision, highlighting the benefit of multi-dimensional metrics.

**Summary.** We find through our experiments that:

- Unlike metrics under the precision and recall framework, Fréchet distance ( $FD$ ) cannot differentiate between failure modes, as per its one-dimensional nature.
- Our inter-class diversity measure, Recall Cross-Entropy ( $RCE$ ), Coverage ( $C$ ) and Recall Coverage ( $RC$ ), are all highly sensitive to mode dropping, while their precision counterparts remain relatively constant.
- Out of the recall measures, only our intra-class diversity measure, Recall Entropy ( $RE$ ), is significantly sensitive to mode shrinkage.
- Only our Precision Cross-Entropy ( $PCE$ ) both correctly correlates with human perception and increases with mode shrinkage. Density ( $D$ ) fails to correlate in the correct direction for CIFAR-10, while Precision Coverage ( $PC$ ) decreases as generated samples become more clustered.

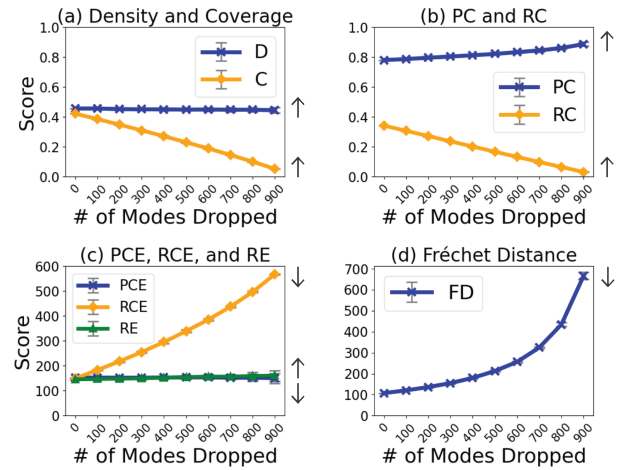


Figure 4: 100 random classes were dropped at a time from the ImageNet ADMG-ADMU generated dataset. The corresponding metric scores are plotted. The scores of 10 runs were averaged and ( $\pm 1$  SD) error bars are plotted, though visually minimal. Similar to Figure 3, arrows are beside each measure’s lines, with the orientation of the arrow corresponding to the direction in which better fidelity/diversity is defined.

## 7 Concluding Remarks

**Limitations.** Quality semantic embedding is needed for our metric to produce meaningful results. However, we emphasize that significant efforts across data modalities have been directed towards improving encoder capabilities and developing automatic, embedding-based evaluation metrics (Zhang et al. 2021; Huang et al. 2020). Specifically, the wide-spread application of Fréchet distance, which requires an embedded space, to text, videos etc. (Xiang et al. 2021; Unterthiner et al. 2019), supports our metric’s application across data types. Another limitation and future research direction is expanding to data with more complex structures, such as graphs and trees, or to specialized generative tasks like surrogates of scientific simulations, where defining an appropriate probability distribution and a meaningful distance measure is nontrivial.

**Conclusion.** In this work, we show that some of the prior approaches to precision and recall metrics for generative models can be reduced to nonparametric measures of information-theoretic divergence. This unifying perspective also allows us to identify a significant gap in existing recall metrics: their limited sensitivity to intra-class diversity loss (mode shrinkage). We then introduce a novel three-dimensional metric for evaluating generative models, comprised of Precision Cross-Entropy, Recall Cross-Entropy, and Recall Entropy. These components address three critical failure modes of generative models: mode invention, mode dropping, and the often-overlooked mode shrinkage. Both empirical evidence, such as correlation with human judgment, and theoretical insights are used to demonstrate the advantages of our proposed metrics over other approaches.

## Acknowledgements

This work was supported by the USDA-NIFA/NSF AI Research Institutes program, under award No. 2021-67021-35344 and by NSF Expeditions in Computing CCF-1918656.

## References

- Alaa, A.; Van Breugel, B.; Saveliev, E. S.; and van der Schaar, M. 2022. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International Conference on Machine Learning*, 290–306. PMLR.
- Borji, A. 2018. Pros and Cons of GAN Evaluation Measures. arXiv:1802.03446.
- Borji, A. 2022. Pros and cons of GAN evaluation measures: New developments. *Computer Vision and Image Understanding*, 215: 103329.
- Cheema, F.; and Uner, R. 2023. Precision Recall Cover: A Method For Assessing Generative Models. In *International Conference on Artificial Intelligence and Statistics*, 6571–6594. PMLR.
- Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.
- Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat GANs on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.
- Djulonga, J.; Lucic, M.; Cuturi, M.; Bachem, O.; Bousquet, O.; and Gelly, S. 2020. Precision-recall curves using information divergence frontiers. In *International Conference on Artificial Intelligence and Statistics*, 2550–2559. PMLR.
- Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Networks. arXiv:1406.2661.
- Han, J.; Choi, H.; Choi, Y.; Kim, J.; Ha, J.-W.; and Choi, J. 2022. Rarity Score : A New Metric to Evaluate the Uncommonness of Synthesized Images. arXiv:2206.08549.
- Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.
- Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598.
- Huang, X.; Khetan, A.; Cvitkovic, M.; and Karnin, Z. 2020. TabTransformer: Tabular Data Modeling Using Contextual Embeddings. arXiv:2012.06678.
- Kingma, D. P.; and Welling, M. 2022. Auto-Encoding Variational Bayes. arXiv:1312.6114.
- Kolchinski, Y. A.; Zhou, S.; Zhao, S.; Gordon, M.; and Ermon, S. 2019. Approximating human judgment of generated image quality. arXiv preprint arXiv:1912.12121.
- Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.
- Kynkäänniemi, T.; Karras, T.; Laine, S.; Lehtinen, J.; and Aila, T. 2019. Improved Precision and Recall Metric for Assessing Generative Models. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Leonenko, N.; Pronzato, L.; and Savani, V. 2008. A class of Rényi information estimators for multidimensional densities.
- Liu, L.; Pillutla, K.; Welleck, S.; Oh, S.; Choi, Y.; and Harchaoui, Z. 2021. Divergence Frontiers for Generative Models: Sample Complexity, Quantization Effects, and Frontier Integrals. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 12930–12942. Curran Associates, Inc.
- Naeem, M. F.; Oh, S. J.; Uh, Y.; Choi, Y.; and Yoo, J. 2020. Reliable Fidelity and Diversity Metrics for Generative Models. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 7176–7185. PMLR.
- Noshad, M.; Moon, K. R.; Sekeh, S. Y.; and Hero, A. O. 2017. Direct estimation of information divergence using nearest neighbor ratios. In *2017 IEEE International Symposium on Information Theory (ISIT)*, 903–907. IEEE.
- Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Assran, M.; Ballas, N.; Galuba, W.; Howes, R.; Huang, P.-Y.; Li, S.-W.; Misra, I.; Rabbat, M.; Sharma, V.; Synnaeve, G.; Xu, H.; Jegou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2024. DINOv2: Learning Robust Visual Features without Supervision. arXiv:2304.07193.
- Park, D.; and Kim, S. 2023. Probabilistic Precision and Recall Towards Reliable Evaluation of Generative Models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE.
- Póczos, B.; and Schneider, J. 2011. On the Estimation of  $\alpha$ -Divergences. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 609–617. JMLR Workshop and Conference Proceedings.
- Ravuri, S.; Rey, M.; Mohamed, S.; and Deisenroth, M. 2023. Understanding Deep Generative Models with Generalized Empirical Likelihoods. arXiv:2306.09780.
- Rényi, A. 1961. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, 547–562. University of California Press.
- Sajjadi, M. S. M.; Bachem, O.; Lucic, M.; Bousquet, O.; and Gelly, S. 2018. Assessing Generative Models via Precision and Recall. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.
- Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training

gans. *Advances in neural information processing systems*, 29.

Sohl-Dickstein, J.; Weiss, E. A.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. *arXiv:1503.03585*.

Stein, G.; Cresswell, J.; Hosseinzadeh, R.; Sui, Y.; Ross, B.; Vilecroze, V.; Liu, Z.; Caterini, A. L.; Taylor, E.; and Loaiza-Ganem, G. 2024. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in Neural Information Processing Systems*, 36.

Theis, L.; Oord, A. v. d.; and Bethge, M. 2015. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*.

Unterthiner, T.; van Steenkiste, S.; Kurach, K.; Marinier, R.; Michalski, M.; and Gelly, S. 2019. Towards Accurate Generative Models of Video: A New Metric & Challenges. *arXiv:1812.01717*.

Xiang, J.; Liu, Y.; Cai, D.; Li, H.; Lian, D.; and Liu, L. 2021. Assessing Dialogue Systems with Distribution Distances. *arXiv:2105.02573*.

Zhang, C.; Sedoc, J.; D'Haro, L. F.; Banchs, R.; and Rudnicky, A. 2021. Automatic Evaluation and Moderation of Open-domain Dialogue Systems. *arXiv:2111.02110*.

Zhou, S.; Gordon, M.; Krishna, R.; Narcomey, A.; Fei-Fei, L. F.; and Bernstein, M. 2019. HYPE: A benchmark for human eye perceptual evaluation of generative models. *Advances in neural information processing systems*, 32.