Unveiling Differences in Generative Models: A Scalable Differential Clustering Approach

Jingwei Zhang Mohammad Jalali Cheuk Ting Li Farzan Farnia
The Chinese University of Hong Kong
{jwzhang22, mjalali24, farnia}@cse.cuhk.edu.hk, ctli@ie.cuhk.edu.hk

Abstract

A fine-grained comparison of generative models requires the identification of sample types generated differently by each of the involved models. While quantitative scores have been proposed in the literature to rank different generative models, score-based evaluation and ranking do not reveal the nuanced differences between the generative models in producing different sample types. In this work, we propose solving a differential clustering *problem to detect sample types generated differently by two generative models. To solve the differential clustering problem, we develop a spectral method called* Fourier-based Identification of Novel Clusters (FINC) *to identify modes produced by a generative model with a higher frequency in comparison to a reference distribution. FINC provides a scalable algorithm based on random Fourier features to estimate the eigenspace of kernel covariance matrices of two generative models and utilize the principal eigendirections to detect the sample types present more dominantly in each model. We demonstrate the application of the FINC method to large-scale computer vision datasets and generative modeling frameworks. Our numerical results suggest the scalability of the developed Fourier-based method in highlighting the sample types produced with different frequencies by generative models. The project code is available at* https://github.com/buyeah1109/FINC

1. Introduction

Deep generative models trained by variational autoencoders [29], generative adversarial networks [14], and denoising diffusion models [20] have led to outstanding results on various computer vision, audio, and video datasets. To compare different generative models and rank their performance, several quantitative scores have been proposed in the computer vision community. Metrics such as Inception Score (IS) [51] and Fréchet Inception Distance (FID) [19] have been widely used in the community for comparing the performance of modern generative model frameworks.

While the scores in the literature provide insightful evaluations of generative models, such score-based assessments do not provide a nuanced comparison between the models, which may lead to inconsistent rankings compared to human evaluations. The recent study [55] demonstrated the imperfections of quantitative evaluation scores, showing that their rankings may not align with human-based assessments. These findings highlight the need for more detailed comparisons of generative models' produced data and the reference datasets. Specifically, a comprehensive comparison that identifies the sample types expressed differently by the models could be useful in several applications. Such a comparison can reveal suboptimally generated sample types, which could be used to further improve the performance of generative models.

To have a fine-grained comparison between two generative models, we propose solving a *differential clustering* problem. This approach seeks to identify the novel sample clusters of a generative model that are generated more frequently compared to a reference distribution. By addressing the differential clustering task, we can identify the differently expressed sample types between the two models and perform an interpretable comparison revealing the types of images captured with a higher frequency by each generative model.

To solve the differential clustering problem, one can utilize the scores proposed in the literature to quantify the uncommonness of generated data. As one such novelty score, [15] proposes the sample-based *Rarity score* measuring the distance of a generated sample to the manifold of a reference distribution. In another work, [25] formulates *Feature Likelihood Divergence (FLD)* as a sample-based novelty score based on the sample's likelihood according to the reference distribution. Note that the application of Rarity and FLD scores for differential clustering requires a post-clustering of the identified novel samples. Also, [61] suggests the spectral *Kernel-based Entropic Novelty (KEN)* framework to detect novel sample clusters by the eigendecomposition of a kernel similarity matrix.

The discussed novelty score-based methods can potentially find the sample types differently captured by two generative models. However, the application of these methods to large-scale datasets containing many sample types, e.g. ImageNet [12], will be computationally challenging. The discussed Rarity and FLD-based methods require a post-clustering of the identified novel samples which will be expensive under a relatively large number of clusters. Furthermore, the KEN method involves the eigendecomposition of a $2n \times 2n$ kernel matrix for a generated data size n. However, on the ImageNet dataset with hundreds of image categories, one should perform the analysis on a large sample size n under which performing the eigendecomposition is computationally costly.

In this work, we develop a scalable algorithm, called *Fourier-based Identification of Novel Clusters (FINC)*, to efficiently address the differential clustering problem and find novel sample types of a test generative model compared to another reference model. In the development of FINC, we follow the random Fourier feature framework proposed in [45] and perform the spectral analysis for finding the novel sample types on the covariance matrix characterized by the random Fourier features.

According to the proposed FINC method, we generate a limited number r of random Fourier features to approximate the Gaussian kernel similarity score. We then run a fully stochastic algorithm to estimate the $r \times r$ covariance matrices of the test and reference generative models in the space of selected Fourier features. We prove that applying the spectral decomposition to the resulting $r \times r$ covariance matrix difference will yield the novel clusters of the test model. As a result, the complexity of the algorithm will primarily depend on the Fourier feature size r that can be selected independently of the sample size n.

On the theory side, we prove an approximation guarantee that a number $r = \mathcal{O}\big(\frac{\log n}{\epsilon^4}\big)$ of random features will be sufficient to accurately approximate the principal eigenvalues of the covariance matrices within an ϵ-bounded approximation error, given n samples generated by the generative models. The main challenge in analyzing the kernel method in [61] is the non-Hermitian structure of their kernel matrix, which makes standard eigenvalue perturbation bounds in the literature inapplicable to analyze the perturbations to the eigenspace of the kernel matrix. To address this, we provide a novel eigenvalue perturbation analysis tailored to this non-symmetric matrix, guaranteeing that the entry-level approximations provided by random Fourier features (RFF) yield accurate proxies for the eigenspace. Our analysis ensures the sub-logarithmic growth of the required Fourier feature size r in the dataset size n, showing the scalability of FINC for identifying differently generated modes.

We numerically evaluate the performance of our proposed FINC method on large-scale image datasets and generative models. In our numerical analysis, we attempt to identify the novel modes between state-of-the-art generative models to characterize the relative strengths of each generative modeling scheme. Also, we apply FINC to find the sub-optimally captured modes of generative models with respect to their target image datasets. The identification of novel and more-frequently generated sample clusters could be utilized to detect generated sample types with higher memorization scores and also data types generated with higher alignment scores in text-guided generative models. In the following, we summarize our work's main contributions:

- Proposing a comparison approach between generative models using differential clustering,
- Developing the FINC method to detect differently-expressed sample types by generative models,
- Providing approximation guarantees on the required Fourier feature size for the FINC method,
- Presenting the numerical results of applying FINC to compare image-based generative models.

2. Related Work

Differential Clustering and Random Fourier Features. Several related works [10, 42, 45, 57] explore the application of random Fourier features [45] to clustering and unsupervised learning tasks. Specifically, [10] uses random Fourier features to accelerate kernel clustering by projecting data points into a low-dimension space where the clustering is performed. [57] demonstrates that a Fourier feature-based mapping enhances a neural net's capability to learn high-frequency functions in low-dimensional data domains. However, these references do not study the differential clustering task in our work. Also, [17] studies the RFF framework with non-symmetric kernel functions, targeting a general non-symmetric kernel that does not offer our logarithmically growing bound in Theorem 1. Also, we note that our target differential clustering task is different from the "differential clustering" task for the single-cell RNA sequencing in the computational biology literature [2, 35, 41]. Specifically, the differential clustering proposed in [2] aims at discovering cell types in two different biological conditions using single-cell RNA sequencing data and describing how the transcriptome profile of these cell types alters as the conditions are changing. On the other hand, our differential clustering task focuses on identifying sample types differently generated by probability models.

Evaluation of generative models. The assessment of deep generative models has been the subject of a large body of related works as surveyed in [5]. The proposed metrics can be categorized into the following groups: 1) distance measures between the generative model and data distribution, including FID [19] and KID [3], 2) scores assessing the quality and diversity of generative models including Inception score [51], precision/recall [30, 50], GAN-train/GAN-

test [53], density/coverage [37], Vendi [13], RKE [24], FKEA-Vendi [39], distributed KID-avg [59], and online multi-armed bandit evaluation schemes [22, 23, 47], 3) measures for the train-to-test generalization evaluation of GANs [1, 25, 36]. Specifically, [13, 24, 39] utilize the eigenspectrum of a kernel matrix to quantify diversity. Unlike our work, these works do not aim for a comparison of generative models in generating different sample types. As for the generalizability evaluation, the mentioned references compare a likelihood [1, 25] or distance [36] measure of the generated data between the training and test datasets. The distribution-based novel mode identification in our work is more general than the train-to-test generalizability, since the reference distribution in our analysis can be different from the generative model's training set.

Novelty assessment of generative models. The recent works [15, 25, 61] study the evaluation of novelty of samples and modes produced by generative models. [15] empirically demonstrates that rare samples are distantly located from the reference data manifold, and proposes the rarity score as the distance to the nearest neighbor for quantifying the uncommonness of a sample. [25] measures the mismatch between the likelihoods evaluated for generated samples with respect to the training set and a reference dataset. We remark that [15, 25] offer *sample-based* novelty scores which needs to be combined with a post-clustering step to address our target differential clustering task. However, the post-clustering step would be computationally costly under a significant number of clusters. Also, [61] proposes a spectral method for measuring the entropy of the novel modes of a generative model with respect to a reference model, which reduces to the eigendecomposition of a kernel similarity matrix. Consequently, [61]'s algorithm leads to an $\mathcal{O}(n^3)$ computational complexity given n generated samples, hindering the method's application to large-scale datasets. In our work, we aim to leverage random Fourier features to reduce the spectral approach's computational complexity and to provide a solution to the differential clustering task.

3. Preliminaries

3.1. Sample-based Evaluation of Generative Models

Given a generative model G generating samples from P_G, the goal of a quantitative evaluation is to use a group of n generated data $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim P_G$ to compute a score quantifying a desired property of the generated data such as visual quality, diversity, generalizability, and novelty. Note that in such a sample-based evaluation, we do not have access to the cumulative distribution function (CDF) of P_G and only observe a group of n samples generated by the model. As discussed in Section 2, multiple scores have been proposed in the literature to assess different properties of generative models. We note that many novelty scores require selecting a reference dataset as a baseline to evaluate the novelty of the test distribution. Similarly, in our work, we select a test-reference distribution pair to assess how the test distribution differs from the reference distribution.

3.2. Kernel Function and Kernel Covariance Matrix

Consider a kernel function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ assigning the similarity score $k(\mathbf{x}, \mathbf{x}')$ to data vectors $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$. The kernel function is assumed to satisfy the positive-semidefinite (PSD) property requiring that for every integer $n \in \mathbb{N}$ and vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$, the $n \times n$ kernel matrix $K = \big[k(\mathbf{x}_i, \mathbf{x}_j)\big]_{n\times n}$ is a PSD matrix. Throughout this paper, we commonly use the Gaussian kernel $k_{\text{Gaussian}(\sigma^2)}$ which for a bandwidth parameter σ^2 is defined as:

$$k_{\text{Gaussian}(\sigma^2)}\big(\mathbf{x}, \mathbf{x}'\big) := \exp\Big(\frac{-\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\sigma^2}\Big) \tag{1}$$

For every kernel function k, there exists a feature map $\phi : \mathbb{R}^d \to \mathbb{R}^s$ such that $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^\top \phi(\mathbf{x}')$ is the inner product of ϕ-based transformations of inputs $\mathbf{x}$ and $\mathbf{x}'$. Given the feature map ϕ corresponding to kernel k, we define the empirical kernel covariance matrix $\widehat{C}_X \in \mathbb{R}^{s\times s}$:

$$\widehat{C}_X := \frac{1}{n}\sum_{i=1}^{n} \phi\big(\mathbf{x}_i\big)\phi\big(\mathbf{x}_i\big)^\top \tag{2}$$

It can be seen that the kernel covariance matrix $\widehat{C}_X$ shares the same eigenvalues with the normalized kernel matrix $\frac{1}{n}K = \frac{1}{n}\big[k(\mathbf{x}_i, \mathbf{x}_j)\big]_{n\times n}$. Assuming a normalized kernel satisfying $k(\mathbf{x}, \mathbf{x}) = 1$ for every $\mathbf{x} \in \mathbb{R}^d$, a condition that holds for the Gaussian kernel in (1), the eigenvalues of $\widehat{C}_X$ will satisfy the probability axioms and result in a probability model. The resulting probability model, as theoretically shown in [61], will estimate the frequencies of the major modes of a mixture distribution.

3.3. Shift-Invariant Kernels and Random Fourier Features

A kernel function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is called *shift-invariant* if there exists a function $\kappa : \mathbb{R}^d \to \mathbb{R}$ such that for every $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$ we have $k\big(\mathbf{x}, \mathbf{x}'\big) = \kappa\big(\mathbf{x} - \mathbf{x}'\big)$. In other words, the output of a shift-invariant kernel $k(\mathbf{x}, \mathbf{x}')$ only depends on the difference $\mathbf{x} - \mathbf{x}'$. For example, the Gaussian kernel in (1) represents a shift-invariant kernel. Bochner's theorem [4] proves that a function $\kappa : \mathbb{R}^d \to \mathbb{R}$ results in a shift-invariant normalized kernel $k(\mathbf{x}, \mathbf{x}') = \kappa(\mathbf{x} - \mathbf{x}')$ if and only if the Fourier transform of κ is a probability density function (PDF) taking non-negative values and integrating to 1. Recall that the Fourier transform of $\kappa : \mathbb{R}^d \to \mathbb{R}$, which we denote by $\widehat{\kappa} : \mathbb{R}^d \to \mathbb{R}$, is defined as

$$\widehat{\kappa}(\boldsymbol{\omega}) := \frac{1}{(2\pi)^d}\int \kappa(\mathbf{x}) \exp\big(-i\boldsymbol{\omega}^\top \mathbf{x}\big)\mathrm{d}\mathbf{x}$$

For example, in the case of Gaussian kernel (1) with bandwidth parameter σ^2, the Fourier transform will be

the PDF of Gaussian distribution $\mathcal{N}(\mathbf{0}, \frac{1}{\sigma^2} I_d)$ with zero mean vector and isotropic covaraince matrix $\frac{1}{\sigma^2} I_d$. Based on Bochner's theorem, [45, 46, 56] suggest using r random Fourier features $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_r \sim \widehat{\kappa}$ drawn independently from PDF $\widehat{\kappa}$, and approximating the shift-invariant kernel function $\kappa(\mathbf{x} - \mathbf{x}')$ as follows where $\phi_r(\mathbf{x}) = \frac{1}{\sqrt{r}}\big[\cos(\boldsymbol{\omega}_1^\top \mathbf{x}), \sin(\boldsymbol{\omega}_1^\top \mathbf{x}), \ldots, \cos(\boldsymbol{\omega}_r^\top \mathbf{x}), \sin(\boldsymbol{\omega}_r^\top \mathbf{x})\big]$ denotes the normalized feature map characterized by the Fourier features: $k\big(\mathbf{x}, \mathbf{x}'\big) \approx \phi_r(\mathbf{x})^\top \phi_r(\mathbf{x}')$

4. A Scalable Algorithm for the Identification of Novel Sample Clusters

To provide an interpretable comparison between the test generative model P_G and the reference distribution P_{ref}, we aim to find the sample clusters generated significantly more frequently by the test model P_G compared to the reference model P_{ref}. We suppose we have access to n independent samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from P_G as well as m independent samples $\mathbf{y}_1, \ldots, \mathbf{y}_m$ from P_{ref}. To address the comparison task, we propose solving a *differential clustering* problem to find clusters of test samples that have a significantly higher likelihood according to P_G than according to P_{ref}.

To address the differential clustering task, we follow the spectral approach proposed in [61], where for a parameter $\rho \geq 1$ we define the *ρ-conditional kernel covariance matrix* $\Lambda_{\mathbf{X}|\rho\mathbf{Y}}$ as the difference of kernel covariance matrices $\widehat{C}_{\mathbf{X}}$, $\widehat{C}_{\mathbf{Y}}$ of test and reference data, respectively, as in (2):

$$\Lambda_{\mathbf{X}|\rho\mathbf{Y}} := \widehat{C}_{\mathbf{X}} - \rho\, \widehat{C}_{\mathbf{Y}} \tag{3}$$

As proven in [61], under multi-modal distributions P_G and P_{ref} with well-separable modes, the eigenvectors corresponding to the positive eigenvalues of the above matrix will reveal the modes of $\mathbf{X}$ which has a frequency ρ-times greater than the mode's frequency in $\mathbf{Y}$. We remark that the constant ρ can be interpreted as the novelty threshold, which the novelty evaluation requires the detected modes of P_G to possess compared to P_{ref}. Therefore, our goal is to find the eigenvectors corresponding to the maximum eigenvalues of the conditional covariance matrix.

The reference [61] applies the kernel trick to solve the problem which leads to the eigendecomposition of an $(m+n) \times (m+n)$ kernel similarity matrix which shares the same eigenvalues with $\Lambda_{\mathbf{X}|\rho\mathbf{Y}}$. However, the resulting eigendecompositon task will be computationally challenging for significantly large m and n sample sizes. Furthermore, while the eigenvectors of [61]'s kernel matrix can cluster the test and reference samples, they do not offer a clustering rule applicable to fresh samples from P_G which would be desired if the learner wants to find the cluster of newly-generated samples from P_G.

In this work, we propose to leverage the framework of random Fourier features (RFF) [45] to find a scalable solution to the eigendecomposition of $\Lambda_{\mathbf{X}|\rho\mathbf{Y}}$. Unlike [61], we do not utilize the kernel trick and aim to directly apply the spectral decomposition to $\Lambda_{\mathbf{X}|\rho\mathbf{Y}}$. However, under the assumed Gaussian kernel similarity measure, the dimensions of the target ρ-conditional kernel covariance matrix is not finite. To resolve this challenge, we approximate the eigenspace of the original $\Lambda_{\mathbf{X}|\rho\mathbf{Y}}$ with that of the covariance matrix of the random Fourier features.

Algorithm 1 Fourier-based Identification of Novel Clusters (FINC)

Input: Sample sets $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and $\{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$, Gaussian kernel bandwidth σ^2, parameter ρ, size r
Draw r Gaussian random vectors $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_r \sim \mathcal{N}\big(\mathbf{0}, \frac{1}{\sigma^2} I_{d\times d}\big)$
Create the map $\widetilde{\phi}_r(\mathbf{x}) = \frac{1}{\sqrt{r}}\big[\cos(\boldsymbol{\omega}_1^\top \mathbf{x}), \sin(\boldsymbol{\omega}_1^\top \mathbf{x}), \ldots, \cos(\boldsymbol{\omega}_r^\top \mathbf{x}), \sin(\boldsymbol{\omega}_r^\top \mathbf{x})\big]$
Initialize $\widetilde{C}_X, \widetilde{C}_Y = \mathbf{0}_{r\times r}$
for $i \in \big\{1, \ldots, \max\{n, m\}\big\}$ **do**
  Update $\widetilde{C}_X \leftarrow \widetilde{C}_X + \frac{1}{n}\widetilde{\phi}_r(\mathbf{x}_i)\widetilde{\phi}_r(\mathbf{x}_i)^\top$
  Update $\widetilde{C}_Y \leftarrow \widetilde{C}_Y + \frac{1}{m}\widetilde{\phi}_r(\mathbf{y}_i)\widetilde{\phi}_r(\mathbf{y}_i)^\top$
end for
Compute $\widetilde{\Lambda}_{\mathbf{X}|\mathbf{Y}} = \widetilde{C}_X - \rho\widetilde{C}_Y$
Apply spectral eigendecomposition to obtain $\widetilde{\Lambda}_{\mathbf{X}|\mathbf{Y}} = U^\top \mathrm{diag}(\boldsymbol{\lambda}) U$
Find the positive eigenvalues $\lambda_1, \ldots, \lambda_t > 0$
Output: Eigenspace $\lambda_1, \ldots, \lambda_t$, $\mathbf{u}_1, \ldots, \mathbf{u}_t$, similarity score $s_{\mathbf{u}}(\mathbf{x}) = \sum_{i=1}^{r} u_{2i-1}\cos(\boldsymbol{\omega}_i^\top \mathbf{x}) + u_{2i}\sin(\boldsymbol{\omega}_i^\top \mathbf{x})$ for every sample $\mathbf{x}$ and mode represented by $\mathbf{u}$.

Specifically, assuming a Gaussian kernel $k_{\text{Gaussian}(\sigma^2)}$ with bandwidth σ^2, we sample r independent vectors $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_r \sim \mathcal{N}(\mathbf{0}, \frac{1}{\sigma^2} I_d)$ from a zero-mean Gaussian distribution with covariance matrix $\frac{1}{\sigma^2} I_d$ where I_d denotes the d-dimensional identity matrix. Given the randomly-sampled frequency parameters, we consider the following feature map $\widetilde{\phi}_r : \mathbb{R}^d \to \mathbb{R}^{2r}$, where $\widetilde{\phi}_r(\mathbf{x})$ is defined as

$$\frac{1}{\sqrt{r}}\Big[\cos\big(\boldsymbol{\omega}_1^\top \mathbf{x}\big), \sin\big(\boldsymbol{\omega}_1^\top \mathbf{x}\big), \ldots, \cos\big(\boldsymbol{\omega}_r^\top \mathbf{x}\big), \sin\big(\boldsymbol{\omega}_r^\top \mathbf{x}\big)\Big]. \tag{4}$$

Then, we compute the positive eigenvalues and corresponding eigenvectors of the following RFF-based ρ-conditional kernel covariance matrix $\widetilde{\Lambda}_{\mathbf{X}|\rho\mathbf{Y}} \in \mathbb{R}^{2r\times 2r}$:

$$\widetilde{\Lambda}_{\mathbf{X}|\rho\mathbf{Y}} := \widetilde{C}_{\mathbf{X}} - \rho\widetilde{C}_{\mathbf{Y}}, \tag{5}$$

$$\text{where } \widetilde{C}_{\mathbf{X}} := \frac{1}{n}\sum_{i=1}^{n} \widetilde{\phi}_r(\mathbf{x}_i)\widetilde{\phi}_r(\mathbf{x}_i)^\top,$$

$$\widetilde{C}_{\mathbf{Y}} := \frac{1}{m}\sum_{j=1}^{m} \widetilde{\phi}_r(\mathbf{y}_j)\widetilde{\phi}_r(\mathbf{y}_j)^\top.$$

Therefore, instead of the original $\Lambda_{\mathbf{X}|\rho\mathbf{Y}}$ with an infinite dimension, we only need to run the spectral eigendecomposition algorithm on the $2r \times 2r$ matrix $\widetilde{\Lambda}_{\mathbf{X}|\rho\mathbf{Y}}$. Furthermore, the computation of both $\widetilde{C}_{\mathbf{X}}$ and $\widetilde{C}_{\mathbf{Y}}$ can be performed using a stochastic algorithm for averaging the $2r \times 2r$ sample-based matrix $\widetilde{\phi}_r(\mathbf{x}_i)\widetilde{\phi}_r(\mathbf{x}_i)^\top$'s and $\widetilde{\phi}_r(\mathbf{y}_j)\widetilde{\phi}_r(\mathbf{y}_j)^\top$'s.

Algorithm 1, which we call the *Fourier-based Identification of Novel Clusters (FINC)*, contains the steps explained above to reach a scalable algorithm for approximating the top more-frequently generated modes of $\mathbf{X} \sim P_G$ with respect to $\mathbf{Y} \sim P_{\text{ref}}$. The following theorem proves that by choosing $r = \mathcal{O}(\frac{\log(n+m)}{\epsilon^4})$, the approximation error of the Fourier-based method is bounded by ϵ.

Theorem 1. *Suppose $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_r \sim \widehat{\kappa}$ are drawn i.i.d. according to the Fourier transform of normalized shift-invariant kernel κ. Define $r' = \min\{2r, m+n\}$ and let $\widetilde{\lambda}_1 \geq \cdots \geq \widetilde{\lambda}_{r'}$ be the sorted top-r' eigenvalues of $\widetilde{\Lambda}_{\mathbf{X}|\rho\mathbf{Y}}$ with corresponding eigenvectors $\widetilde{\mathbf{v}}_1, \ldots, \widetilde{\mathbf{v}}_{r'}$. Similarly, define $\lambda_1 \geq \cdots \geq \lambda_{r'}$ as the sorted top-r' eigenvalues of $\Lambda_{\mathbf{X}|\rho\mathbf{Y}}$. Then, for every $\delta > 0$, with probability at least $1 - \delta$ we will have the following approximation guarantee with $\epsilon = \big(2(1+\rho)\big)^{3/2} \sqrt[4]{\frac{\log((m+n)/\delta)}{r}}$:*

$$\sum_{i=1}^{r'} \big(\widetilde{\lambda}_i - \lambda_i\big)^2 \leq \epsilon^2,$$

$$\max_{1 \leq i \leq r'} \Big\| \Lambda_{\mathbf{X}|\rho\mathbf{Y}} \mathbf{v}_i - \lambda_i \mathbf{v}_i \Big\|_2 \leq 2(1+\rho)\epsilon,$$

where for every i we define $\mathbf{v}_i = \sum_{t=1}^{n} \big(\widetilde{\mathbf{v}}_i^\top \widetilde{\phi}_r(\mathbf{x}_i)\big)\phi(\mathbf{x}_i) + \sum_{s=1}^{m} \big(\widetilde{\mathbf{v}}_i^\top \widetilde{\phi}_r(\mathbf{y}_s)\big)\phi(\mathbf{y}_s)$ as the proxy eigenvector corresponding to the given $\widetilde{\mathbf{v}}_i$.

Theorem 1 proves that the eigenvalues and eigenvectors of the Fourier-based approximation $\widetilde{\Lambda}_{\mathbf{X}|\rho\mathbf{Y}}$ will be ϵ-close to those of the original conditional kernel covariance matrix $\Lambda_{\mathbf{X}|\rho\mathbf{Y}}$, conditioned that we have $r = \mathcal{O}(\frac{\log(m+n)}{\epsilon^4})$ number of random Fourier features. Therefore, for a fixed ϵ, the required number of Fourier features grows only logarithmically with the test and reference sample size $m+n$. This guarantee shows that the complexity of FINC is significantly lower than the kernel method in [61] applying eigendecomposition to an $(m+n) \times (m+n)$ kernel matrix.

5. Numerical Results

5.1. Experiment Setting

Datasets. We performed the numerical experiments on the following standard image datasets: 1) ImageNet-1K [12] containing 1.4 million images with 1000 labels. The ImageNet-dogs subset used in our experiments includes 20k dog images from 120 different breeds, 2) CelebA [34] containing 200k human face images from celebrities, 3) FFHQ [27] including 70k human-face images, 4) AFHQ [11] including 15k dogs, cats, and wildlife animal-face images.

Feature extraction: Following the standard in the literature of generative model evaluation, we utilized the DINOv2 model [55] to perform feature extraction.

Hyper-parameters selection and sample size: We chose the kernel bandwidth σ^2 by searching for the smallest σ value resulting in a variance bounded by 0.01. Also, we conducted the experiments using at least $m, n = 50$k sample sizes for every generative model. For Fourier feature size r, we found by grid search that $r = 2000$ fulfills the goal of approximation error estimate below 1%.

5.2. Identification of More-Frequent and Novel Modes via FINC

Sanity check on FINC-detected novel modes. We empirically tested the validity of FINC's detected novel modes in controlled settings. We performed an experiment on a colored MNIST dataset. To do this, we divided the MNIST training set into two halves, one half for the reference distribution and the other for the test distribution. As a result, every grayscale *sample* in the test distribution is not included in the reference dataset. Then, we colorized half of the samples in the test sample set, by assigning an exclusive color to every digit. Consequently, colored samples form the ground-truth novel sample clusters of the test distribution compared to the grayscale reference sample set. We tested whether the FINC detected top-10 novel modes match the known ground-truth novelty.

As shown in Figure 1, FINC could detect the ten types of colored digits as its top-10 novel modes. We also repeated experiments on the colorized Fashion-MNIST dataset, and observed similar results which are put in the Appendix B. We note that we used the original pixel embedding in these experiments. In addition, we investigated the image dataset pair, AFHQ and ImageNet-dogs, using the DINOv2 embedding, and as shown in Figure 2, observed FINC-detected novel modes are consistent with our knowledge of the ground-truth novel modes between the datasets.

FINC-detected novel modes in generative models. Next, we applied FINC to detect novel modes between various generative models trained on ImageNet-1K and FFHQ. We tested generative models following standard frameworks, including GAN-based BigGAN [6], GigaGAN [26], StyleGAN-XL [52], InsGen [60], diffusion-based LDM [48], DiT-XL [40], and VAE-based RQ-Transformer [32], VDVAE [9]. For generative models trained on the ImageNet training set, we first demonstrate their overrepresented modes in Figure 3. Then, we illustrate novel modes between generative models. Figure 4 displays FINC-detected more frequently generated modes of LDM compared to the reference generative models with threshold $\rho = 1$. We observe the higher-frequency modes of LDM

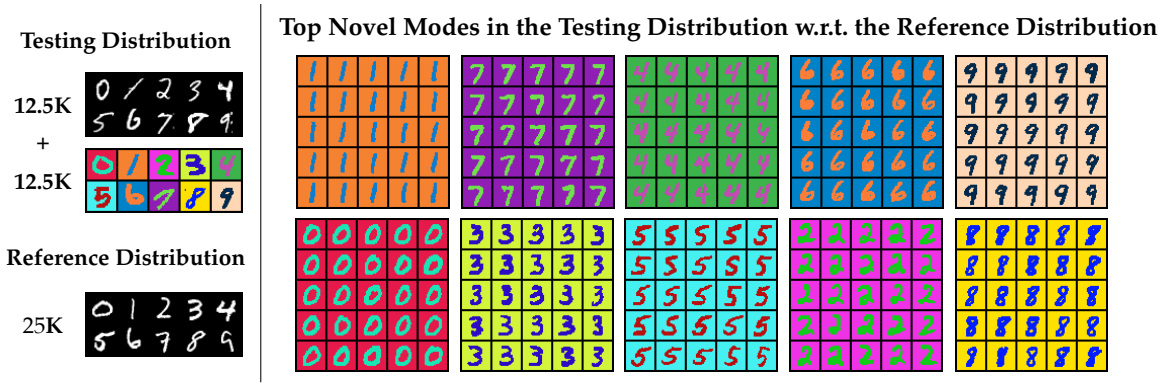

Figure 1. FINC-identified top 10 novel modes between differently-colored MNIST datasets.

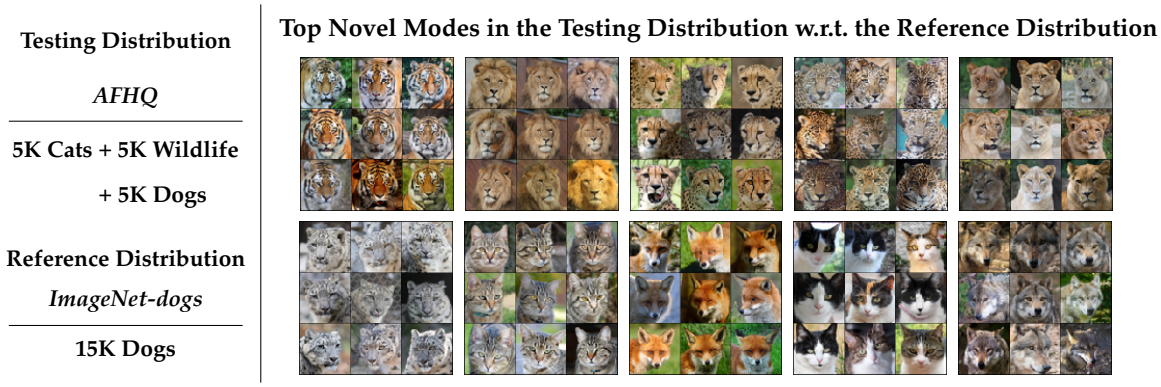

Figure 2. FINC-identified top novel modes between AFHQ & ImageNet-dogs, DINOv2 embedding is used.

is moderately consistent to its overrepresented modes (e.g. mode "koala"). Meanwhile, we can reverse the role of test and reference distribution to detect underrepresented and less frequently generated modes of generative models. Due to the page limit, we defer these results as well as results similar to Figure 3 and 4 with the FFHQ dataset, video datasets, and other generative models to the Appendix B.

FINC mode-based similarity score. Also, we used the eigenfunctions corresponding to the computed eigenvectors to assign a mode-based similarity score to images. As Theorem 1 implies, the mode-scoring function for the eigenvector $\widetilde{\mathbf{v}}$ is $s_{\widetilde{\mathbf{v}}}(\mathbf{x}) = \sum_{i=1}^{r} \tilde{v}_{2i-1} \cos(\boldsymbol{\omega}_i^\top \mathbf{x}) + \tilde{v}_{2i} \sin(\boldsymbol{\omega}_i^\top \mathbf{x})$. This likelihood quantification can apply to both train data and fresh unseen data not contained in the sample sets for running FINC. As a result, we may retrieve the samples with largest scores assigned by top eigenvectors (ranked by associated eiganvalues) to visualize top novel modes. According to our empirical study, the FINC-based assigned scores indicate a semantically meaningful ranking of the images' likelihood of belonging to the displayed modes. In the Appendix B, we illustrate the evaluated similarity scores under two pairs of datasets, where we measured and ranked a group of unseen data according to the FINC-detected modes' scoring function.

Scalability of FINC in the Identification of Novel Modes. Scalability is a key property of a differential clustering algorithm in application to a data distribution with a significant number of modes, such as ImageNet. To evaluate the scalability of the proposed FINC method compared to the baselines, we recorded the time (in seconds) in Table 1 of running the algorithm on the ImageNet-1K dataset. In addition to the computational efficiency, we also qualitatively evaluate the novel modes detected by baselines and the FINC. For baselines, we downloaded the implementation of the Rarity score [15], FLD score [25], and KEN method [61], and measured their time for detecting the novel samples followed by a spectral-clustering step performed by the standard TorchCluster package.

Quantitative: Table 1 shows the time taken by FINC with $r = 2000$ and the baselines on different sample sizes (we suppose $n = m$), suggesting an almost linearly growing time with sample size in the case of FINC. In contrast, the baselines' spent time grew at a superlinear rate, and for sample sizes larger than 10k and 50k all the baselines ran into a memory overflow on the GPU and CPU processors.

Qualitative: We utilize the largest applicable sample size in our GPU processors for baselines. According to our empirical experiments, FINC seems to better detect novel modes compared to baselines in the ImageNet experiments. We defer the detailed results to the Appendix B.

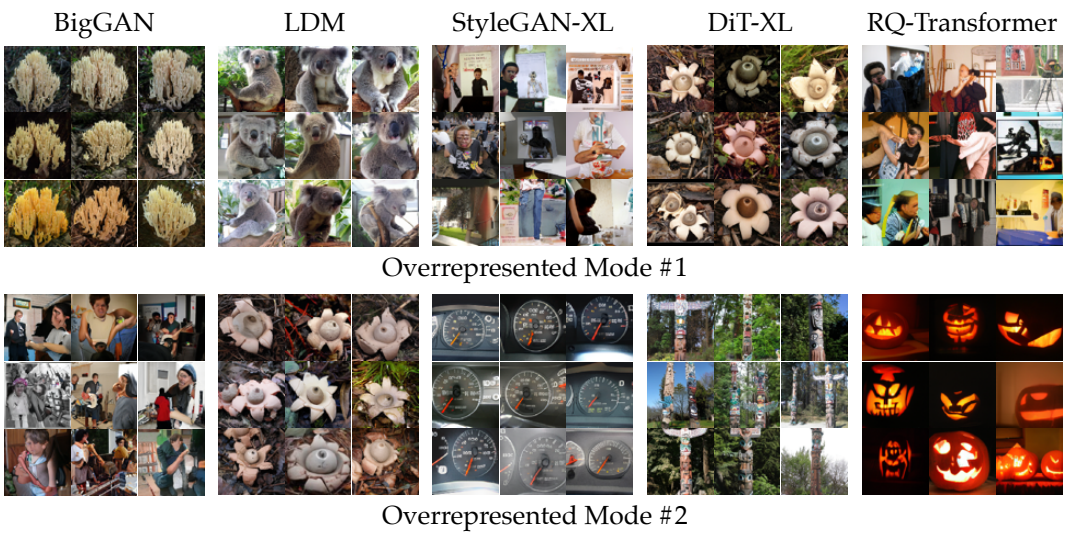

Figure 3. The FINC-identified top two overrepresented modes expressed by test generative models with a higher frequency than in the reference ImageNet dataset. DINOv2 embedding is used.

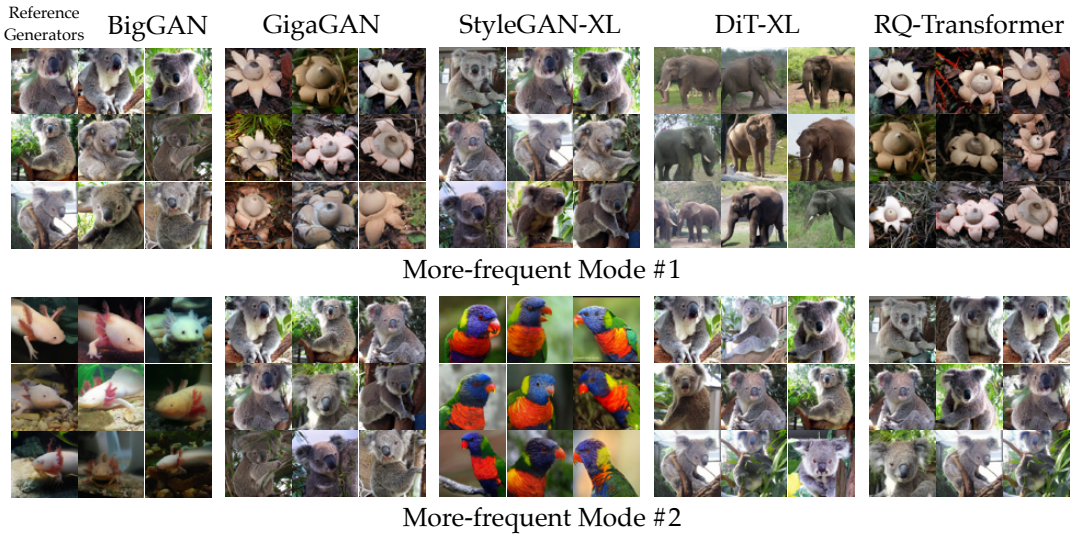

Figure 4. Representative samples from FINC-detected top-2 modes with the maximum frequency gap between LDM (higher frequency) and reference generative models. DINOv2 embedding is used.

5.3. Detection of High Memorization-Score Modes

In the literature, memorization scores have been proposed to detect memorizing training samples by generative models. Although sample-level memorization metrics have been used to quantify and rank the similarity of every generated sample to training data, these scores do not directly reveal which sample types could be memorized more frequently by a generative model. To address this task, we apply FINC to conduct differential clustering on the groups of samples with low and high memorization FLD scores.

To do this, we evaluated and ranked the sample-level memorization score for all the generated samples using the FLD memorization metric with respect to the ImageNet training samples. Subsequently, we separated the two groups f samples with the top 20% and bottom 20% of FLD memorization scores. We performed differential clustering via FINC between the two groups, and used the eigenvalues/eigenvectors computed by Algorithm 1 to quantify and rank mode-level memorization scores. Figure 5 displays the top two detected sample types that occur more frequently in

Table 1. Time complexity (in seconds) of solving the differential clustering problem on the ImageNet dataset using the proposed FINC method vs. the baselines. The dash - means failure due to memory overflow, and $> 24h$ implies not complete in 24 hours. For baseline Rarity and FLD, post-clustering method spectral clustering is adopted. For KEN score, both eigen and Cholesky decomposition are used.

Processor	Method	Sample Size							
		1k	2.5k	5k	10k	25k	50k	100k	250k
CPU	Rarity [15]	1.4	9.1	37.6	153.5	1030	4616	-	-
	FLD [25]	2.7	14.3	59.4	224.5	1408	6024	-	-
	KEN [61]	3.5	26.3	122.8	1130	14354	>24h	-	-
	KEN (Cholesky) [61]	2.6	16.1	61.7	266.0	2045	10934	-	-
	FINC ($r = 2000$)	13.1	31.6	62.8	125.5	311.2	617.6	1248	3090
GPU	Rarity [15]	0.1	4.3	10.7	27.6	-	-	-	-
	FLD [25]	0.2	3.7	10.8	25.1	-	-	-	-
	KEN [61]	1.1	11.6	52.1	455.6	-	-	-	-
	KEN (Cholesky) [61]	0.1	0.7	3.2	16.7	-	-	-	-
	FINC ($r = 2000$)	0.5	1.0	1.6	3.1	7.6	15.0	29.8	74.3

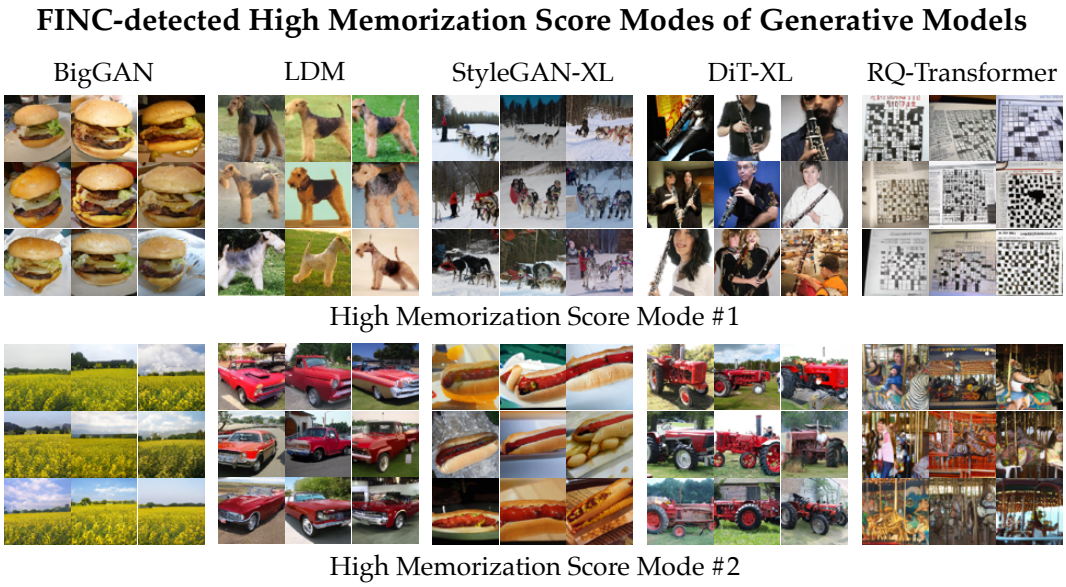

Figure 5. The FINC-identified top-2 generated sample clusters with higher FLD memorization scores w.r.t. the ImageNet training dataset. The top 9 images with the maximum alignment with the detected mode's eigenvectors are shown.

the high-memorization samples of the generative models.

5.4. Detection of Sample Clusters with Higher and Lower CLIPScores

The CLIPScore [18] is a widely-used sample-based metric to measure the alignment of the generated image by a text-to-image model. To identify the sample types with the highest and lowest CLIPScores, we ran a differential clustering experiment between the CLIPScore-based top 20% and the bottom 20% of samples generated by three text-to-image models, GigaGAN, SD-XL, and Kandinsky in response to text prompts from MS-COCO dataset. We show the top-3 and bottom-3 FINC-identified modes with the maximum and minimum CLIPScores in Appendix B.

6. Conclusion

In this work, we developed a scalable algorithm to find the differently expressed modes of two generative models. The proposed Fourier-based approach attempts to solve a differential clustering problem to find the sample clusters generated with significantly different frequencies by the test generative model and a reference distribution. We demonstrated the application of the proposed methodology in finding the suboptimally captured modes in a generative model compared to a reference distribution. An interesting future direction to this work is to extend the comparison framework to more than two generative models, where instead of pairwise comparison, one can simultaneously compare the modes generated by multiple generative models.

Acknowledgments

The work of Farzan Farnia is partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China, Project 14209920, and is partially supported by CUHK Direct Research Grants with CUHK Project No. 4055164 and 4937054. The work of Cheuk Ting Li is partially supported by two grants from the Research Grants Council of the Hong Kong Special Administrative Region, China [Project No.s: CUHK 24205621 (ECS), CUHK 14209823 (GRF)]. The authors also thank the anonymous reviewers for their constructive feedback and useful suggestions.

References

[1] Ahmed Alaa, Boris Van Breugel, Evgeny S Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International Conference on Machine Learning*, pages 290–306. PMLR, 2022. 3

[2] Martin Barron, Siyuan Zhang, and Jun Li. A sparse differential clustering algorithm for tracing cell type changes via single-cell rna-sequencing data. *Nucleic acids research*, 46 (3):e14–e14, 2018. 2

[3] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 2

[4] Salomon Bochner. Lectures on fourier integrals. *Princeton University Press*, 1949. 3

[5] Ali Borji. Pros and cons of gan evaluation measures: New developments. *Computer Vision and Image Understanding*, 215:103329, 2022. 2

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 5

[7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments, 2021. 16

[8] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017. 16

[9] Rewon Child. Very deep vaes generalize autoregressive models and can outperform them on images. *arXiv preprint arXiv:2011.10650*, 2020. 5

[10] Radha Chitta, Rong Jin, and Anil K Jain. Efficient kernel clustering using random fourier features. In *2012 IEEE 12th International Conference on Data Mining*, pages 161–170. IEEE, 2012. 2

[11] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020. 5

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 5

[13] Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022. 3

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1

[15] Jiyeon Han, Hwanil Choi, Yunjey Choi, Junho Kim, Jung-Woo Ha, and Jaesik Choi. Rarity score : A new metric to evaluate the uncommonness of synthesized images. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 3, 6, 8, 16

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 16

[17] Mingzhen He, Fan He, Fanghui Liu, and Xiaolin Huang. Random fourier features for asymmetric kernels. *Machine Learning*, 113(11):8459–8485, 2024. 2

[18] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 8, 16

[19] 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. *Advances in neural information processing systems*, 30, 2017. 1, 2

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[21] Alan J Hoffman and Helmut W Wielandt. The variation of the spectrum of a normal matrix. In *Selected Papers Of Alan J Hoffman: With Commentary*, pages 118–120. World Scientific, 2003. 14

[22] Xiaoyan Hu, Ho-fung Leung, and Farzan Farnia. An online learning approach to prompt-based selection of generative models. *arXiv preprint arXiv:2410.13287*, 2024. 3

[23] Xiaoyan Hu, Ho-fung Leung, and Farzan Farnia. An optimism-based approach to online evaluation of generative models. *arXiv preprint arXiv:2406.07451*, 2024. 3

[24] Mohammad Jalali, Cheuk Ting Li, and Farzan Farnia. An information-theoretic evaluation of generative models in learning multi-modal distributions. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3, 16

[25] Marco Jiralerspong, Joey Bose, Ian Gemp, Chongli Qin, Yoram Bachrach, and Gauthier Gidel. Feature likelihood score: Evaluating the generalization of generative models using samples. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1, 3, 6, 8, 16

[26] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 5, 16

[27] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 5

[28] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017. 16

[29] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1

[30] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[31] Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of imagenet classes in fréchet inception distance. In *The Eleventh International Conference on Learning Representations*, 2023. 15, 16

[32] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11523–11532, 2022. 5

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 16

[34] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 5

[35] Malte D Luecken and Fabian J Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6):e8746, 2019. 2

[36] Casey Meehan, Kamalika Chaudhuri, and Sanjoy Dasgupta. A non-parametric test to detect data-copying in generative models. In *International Conference on Artificial Intelligence and Statistics*, 2020. 3

[37] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, pages 7176–7185. PMLR, 2020. 3

[38] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 16

[39] Azim Ospanov, Jingwei Zhang, Mohammad Jalali, Xuenan Cao, Andrej Bogdanov, and Farzan Farnia. Towards a scalable reference-free evaluation of generative models. *arXiv preprint arXiv:2407.02961*, 2024. 3

[40] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022. 5

[41] Lihong Peng, Xiongfei Tian, Geng Tian, Junlin Xu, Xin Huang, Yanbin Weng, Jialiang Yang, and Liqian Zhou. Single-cell rna-seq clustering: datasets, models, and algorithms. *RNA biology*, 17(6):765–783, 2020. 2

[42] Ninh Pham and Rasmus Pagh. Fast and scalable polynomial kernels via explicit feature maps. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 239–247, 2013. 2

[43] John Phillips. On the uniform continuity of operator functions and generalized powers-stormer inequalities. Technical report, 1987. 14

[44] 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. 16

[45] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007. 2, 4

[46] Ali Rahimi and Benjamin Recht. Uniform approximation of functions with random bases. In *2008 46th annual allerton conference on communication, control, and computing*, pages 555–561. IEEE, 2008. 4

[47] Parham Rezaei, Farzan Farnia, and Cheuk Ting Li. Be more diverse than the most diverse: Online selection of diverse mixtures of generative models. *arXiv preprint arXiv:2412.17622*, 2024. 3

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 5

[49] Masaki Saito, Shunta Saito, Masanori Koyama, and Sosuke Kobayashi. Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal gan. *International Journal of Computer Vision*, 128(10): 2586–2606, 2020. 16

[50] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 2018. 2

[51] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2016. 1, 2

[52] 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. 5

[53] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. How good is my gan? In *Proceedings of the European conference on computer vision (ECCV)*, pages 213–229, 2018. 3

[54] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild, 2012. 16

[55] George Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Leigh Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, J. Eric T. Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models, 2023. 1, 5, 15, 16

[56] Danica J Sutherland and Jeff Schneider. On the error of random fourier features. *arXiv preprint arXiv:1506.02785*, 2015. 4

[57] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020. 2

[58] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges, 2019. 16

[59] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8684–8694, 2020. 3

[60] Ceyuan Yang, Yujun Shen, Yinghao Xu, and Bolei Zhou. Data-efficient instance generation from instance discrimination. *arXiv preprint arXiv:2106.04566*, 2021. 5

[61] Jingwei Zhang, Cheuk Ting Li, and Farzan Farnia. An interpretable evaluation of entropy-based novelty of generative models. *arXiv preprint arXiv:2402.17287*, 2024. 1, 2, 3, 4, 5, 6, 8, 13, 16

A. Proofs

A.1. Proof of Theorem 1

First, note that the Gaussian probability density function (PDF) $\mathcal{N}\big(\mathbf{0}, \sigma^{-2} I_d\big)$ is proportional to the Fourier transform of the Gaussian kernel in (1), because for the corresponding shift-invariant kernel producing function $\kappa_\sigma(\mathbf{x}) = \exp\big(\frac{-\|\mathbf{x}\|_2^2}{2\sigma^2}\big)$ we have the following Fourier transform:

$$\begin{aligned}
\widehat{\kappa_\sigma}(\boldsymbol{\omega}) &= \frac{1}{(2\pi)^d} \int \kappa_\sigma(\mathbf{x}) \exp(-i\boldsymbol{\omega}^\top \mathbf{x}) \mathrm{d}\mathbf{x} \\
&= \frac{1}{(2\pi)^d} \Big(2\pi\sigma^2\Big)^{d/2} \exp\Big(-\frac{\sigma^2 \|\boldsymbol{\omega}\|_2^2}{2}\Big) \\
&= \underbrace{\Big(\frac{1}{2\pi\sigma^{-2}}\Big)^{d/2} \exp\Big(-\frac{\|\boldsymbol{\omega}\|_2^2}{2\sigma^{-2}}\Big)}_{\text{PDF of } \mathcal{N}(\mathbf{0}, \sigma^{-2} I)}.
\end{aligned}$$

On the other hand, according to the synthesis property of Fourier transform we can write

$$\begin{aligned}
k_{\text{Gaussian}(\sigma^2)}\big(\mathbf{x}, \mathbf{x}'\big) &= \kappa_\sigma\big(\mathbf{x} - \mathbf{x}'\big) \\
&\overset{(a)}{=} \int \widehat{\kappa_\sigma}(\boldsymbol{\omega}) \exp\big(i\boldsymbol{\omega}^\top (\mathbf{x} - \mathbf{x}')\big) \mathrm{d}\boldsymbol{\omega} \\
&\overset{(b)}{=} \int \widehat{\kappa_\sigma}(\boldsymbol{\omega}) \cos\big(\boldsymbol{\omega}^\top (\mathbf{x} - \mathbf{x}')\big) \mathrm{d}\boldsymbol{\omega} \\
&\overset{(c)}{=} \mathbb{E}_{\boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0}, \sigma^{-2} I)} \Big[\cos\big(\boldsymbol{\omega}^\top (\mathbf{x} - \mathbf{x}')\big)\Big] \\
&= \mathbb{E}_{\boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0}, \sigma^{-2} I)} \Big[\cos\big(\boldsymbol{\omega}^\top \mathbf{x}\big) \cos\big(\boldsymbol{\omega}^\top \mathbf{x}'\big) + \sin\big(\boldsymbol{\omega}^\top \mathbf{x}\big) \sin\big(\boldsymbol{\omega}^\top \mathbf{x}'\big)\Big]
\end{aligned}$$

In the above, (a) follows from the synthesis equation of Fourier transform. (b) uses the fact that $\widehat{\kappa_\sigma}$ is an even function, leading to a zero imaginary term in the Fourier synthesis equation. (c) use the relationship between $\widehat{\kappa_\sigma}$ and the PDF of $\mathcal{N}(\mathbf{0}, \sigma^{-2} I)$. Therefore, since $\big|\cos\big(\boldsymbol{\omega}^\top \mathbf{y}\big)\big| \leq 1$ holds for every $\boldsymbol{\omega}$ and $\mathbf{y}$, we can apply Hoeffding's inequality to show that for independently sampled $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_r \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \sigma^{-2} I)$ the following probabilistic bound holds:

$$\mathbb{P}\bigg(\Big|\frac{1}{r} \sum_{i=1}^{r} \cos\big(\boldsymbol{\omega}_i^\top (\mathbf{x} - \mathbf{x}')\big) - \mathbb{E}_{\boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0}, \sigma^{-2} I)} \Big[\cos\big(\boldsymbol{\omega}^\top (\mathbf{x} - \mathbf{x}')\big)\Big]\Big| \geq \epsilon\bigg) \leq 2 \exp\Big(-\frac{r\epsilon^2}{2}\Big)$$

However, note that the identity $\cos(a - b) = \cos(a)\cos(b) + \sin(a)\sin(b)$ implies that $\frac{1}{r} \sum_{i=1}^{r} \cos\big(\boldsymbol{\omega}_i^\top (\mathbf{x} - \mathbf{x}')\big) = \widetilde{\phi}_r(\mathbf{x})^\top \widetilde{\phi}_r(\mathbf{x}')$ and so we can rewrite the above bound as

$$\mathbb{P}\bigg(\Big|\widetilde{\phi}_r(\mathbf{x})^\top \widetilde{\phi}_r(\mathbf{x}') - k_{\text{Gaussian}(\sigma^2)}(\mathbf{x}, \mathbf{x}')\Big| \geq \epsilon\bigg) \leq 2 \exp\Big(-\frac{r\epsilon^2}{2}\Big).$$

Note that both $k_{\text{Gaussian}(\sigma^2)}$ and $\widetilde{k}_r(\mathbf{x}, \mathbf{x}') = \widetilde{\phi}_r(\mathbf{x})^\top \widetilde{\phi}_r(\mathbf{x}')$ are normalized kernels, implying that

$$\forall \mathbf{x} \in \mathbb{R}^d : \quad \widetilde{\phi}_r(\mathbf{x})^\top \widetilde{\phi}_r(\mathbf{x}) - k_{\text{Gaussian}(\sigma^2)}(\mathbf{x}, \mathbf{x}) = 0.$$

In addition, the application of the union bound shows that over the test sample set $\mathbf{x}_1, \ldots, \mathbf{x}_n$, we can write

$$\mathbb{P}\bigg(\max_{1 \leq i,j \leq n} \Big(\widetilde{\phi}_r(\mathbf{x}_i)^\top \widetilde{\phi}_r(\mathbf{x}_j) - k_{\text{Gaussian}(\sigma^2)}(\mathbf{x}_i, \mathbf{x}_j)\Big)^2 \geq \epsilon^2\bigg) \leq 2\binom{n}{2} \exp\Big(-\frac{r\epsilon^2}{2}\Big).$$

Similarly, we can show the following concerning the reference sample set $\{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$

$$\mathbb{P}\bigg(\max_{1 \leq i \leq n,\, 1 \leq j \leq m} \Big(\widetilde{\phi}_r(\mathbf{x}_i)^\top \widetilde{\phi}_r(\mathbf{y}_j) - k_{\text{Gaussian}(\sigma^2)}(\mathbf{x}_i, \mathbf{y}_j)\Big)^2 \geq \epsilon^2\bigg) \leq 2mn \exp\Big(-\frac{r\epsilon^2}{2}\Big),$$

$$\mathbb{P}\Big(\max_{1\le i,j\le m}\Big(\widetilde{\phi}_r(\mathbf{y}_i)^\top\widetilde{\phi}_r(\mathbf{y}_j)-k_{\text{Gaussian}(\sigma^2)}(\mathbf{y}_i,\mathbf{y}_j)\Big)^2\ge\epsilon^2\Big)\le 2\binom{m}{2}\exp\Big(-\frac{r\epsilon^2}{2}\Big)$$

Applying the union bound, we can utilize the above inequalities to show

$$\begin{aligned}
&\mathbb{P}\Big(\max_{1\le i,j\le n}\Big(\widetilde{\phi}_r(\mathbf{x}_i)^\top\widetilde{\phi}_r(\mathbf{x}_j)-k_{\text{Gaussian}(\sigma^2)}(\mathbf{x}_i,\mathbf{x}_j)\Big)^2\ge\epsilon^2\\
&\text{OR }\max_{1\le i,j\le m}\Big(\widetilde{\phi}_r(\mathbf{y}_i)^\top\widetilde{\phi}_r(\mathbf{y}_j)-k_{\text{Gaussian}(\sigma^2)}(\mathbf{y}_i,\mathbf{y}_j)\Big)^2\ge\epsilon^2\\
&\text{OR }\max_{1\le i\le n,\,1\le j\le m}\Big(\widetilde{\phi}_r(\mathbf{x}_i)^\top\widetilde{\phi}_r(\mathbf{y}_j)-k_{\text{Gaussian}(\sigma^2)}(\mathbf{x}_i,\mathbf{y}_j)\Big)^2\ge\epsilon^2\Big)\\
&\le\Big(2\binom{m}{2}+2\binom{n}{2}+2mn\Big)\exp\Big(-\frac{r\epsilon^2}{2}\Big)\\
&<(m+n)^2\exp\Big(-\frac{r\epsilon^2}{2}\Big)
\end{aligned}\tag{6}$$

On the other hand, according to Theorem 3 from [61], we know that the non-zero eigenvalues of the ρ-conditional covariance matrix $\Lambda_{\mathbf{X}|\rho\mathbf{Y}}:=\widehat{C}_{\mathbf{X}}-\rho\widehat{C}_{\mathbf{Y}}$ are shared with the ρ-conditional kernel matrix

$$K_{\mathbf{X}|\rho\mathbf{Y}}:=\begin{bmatrix}\frac{1}{n}K_{XX} & \sqrt{\frac{\rho}{mn}}K_{XY}\\ -\sqrt{\frac{\rho}{mn}}K_{XY}^\top & \frac{-\rho}{m}K_{YY}\end{bmatrix}$$

where the kernel function is the Gaussian kernel $k_{\text{Gaussian}(\sigma^2)}$. Then, defining the diagonal matrix $D=\operatorname{diag}\big([\underbrace{1,\ldots,1}_{n\text{ times}},\underbrace{-1,\ldots,-1}_{m\text{ times}}]\big)$, we observe that

$$K_{\mathbf{X}|\rho\mathbf{Y}}\;=\;D\cdot\underbrace{\begin{bmatrix}\frac{1}{n}K_{XX} & \sqrt{\frac{\rho}{mn}}K_{XY}\\ \sqrt{\frac{\rho}{mn}}K_{XY}^\top & \frac{\rho}{m}K_{YY}\end{bmatrix}}_{K_{\mathbf{X},\rho\mathbf{Y}}}$$

where we define $K_{\mathbf{X},\rho\mathbf{Y}}$ as the joint kernel matrix that is guaranteed to be a symmetric PSD matrix. Therefore, there exists a unique symmetric and PSD matrix $\sqrt{K_{\mathbf{X},\rho\mathbf{Y}}}$ which is the square root of $K_{\mathbf{X},\rho\mathbf{Y}}$. Since for every matrices A, B where AB and BA are well-defined, AB and BA share the same non-zero eigenvalues, we conclude that $K_{\mathbf{X}|\rho\mathbf{Y}}=DK_{\mathbf{X},\rho\mathbf{Y}}$ has the same non-zero eigenvalues as $\underline{K}=\sqrt{K_{\mathbf{X},\rho\mathbf{Y}}}\,D\,\sqrt{K_{\mathbf{X},\rho\mathbf{Y}}}$. As a result, $\Lambda_{\mathbf{X}|\rho\mathbf{Y}}$ shares the same non-zero eigenvalues with the defined $\underline{K}$.

We repeat the above definitions (considering the Gaussian kernel $k_{\text{Gaussian}(\sigma^2)}$) for the proxy kernel function $\widetilde{k}_r(\mathbf{x},\mathbf{x}')=\widetilde{\phi}_r(\mathbf{x})^\top\widetilde{\phi}_r(\mathbf{x}')$ to obtain proxy kernel matrices $\widetilde{K}_{\mathbf{X}|\rho\mathbf{Y}}$, $\widetilde{K}_{\mathbf{X},\rho\mathbf{Y}}$, $\underline{\widetilde{K}}=\sqrt{\widetilde{K}_{\mathbf{X},\rho\mathbf{Y}}}\,D\,\sqrt{\widetilde{K}_{\mathbf{X},\rho\mathbf{Y}}}$. We note that Equation (6) implies that

$$\begin{aligned}
&\mathbb{P}\Big(\big\|\widetilde{K}_{\mathbf{X},\rho\mathbf{Y}}-K_{\mathbf{X},\rho\mathbf{Y}}\big\|_F^2\;\ge\;n^2\frac{\epsilon^2}{n^2}+2nm\frac{\rho\epsilon^2}{nm}+m^2\frac{\rho^2\epsilon^2}{m^2}\Big)\\
&\le(m+n)^2\exp\Big(-\frac{r\epsilon^2}{2}\Big).\\
\implies\;&\mathbb{P}\Big(\big\|\widetilde{K}_{\mathbf{X},\rho\mathbf{Y}}-K_{\mathbf{X},\rho\mathbf{Y}}\big\|_F^2\;\ge\;\epsilon^2(1+\rho)^2\Big)\;\le\;(m+n)^2\exp\Big(-\frac{r\epsilon^2}{2}\Big).\\
\implies\;&\mathbb{P}\Big(\big\|\widetilde{K}_{\mathbf{X},\rho\mathbf{Y}}-K_{\mathbf{X},\rho\mathbf{Y}}\big\|_F\;\ge\;\epsilon\Big)\;\le\;(m+n)^2\exp\Big(-\frac{r\epsilon^2}{2(1+\rho)^2}\Big).
\end{aligned}\tag{7}$$

The above inequality will imply the following where $\|\cdot\|_2$ denotes the spectral norm

$$\mathbb{P}\Big(\big\|\widetilde{K}_{\mathbf{X},\rho\mathbf{Y}}-K_{\mathbf{X},\rho\mathbf{Y}}\big\|_2\;\ge\;\epsilon\Big)\;\le\;\mathbb{P}\Big(\big\|\widetilde{K}_{\mathbf{X},\rho\mathbf{Y}}-K_{\mathbf{X},\rho\mathbf{Y}}\big\|_F\;\ge\;\epsilon\Big)$$

$$\leq (m+n)^2 \exp\Big(-\frac{r\epsilon^2}{2(1+\rho)^2}\Big). \tag{8}$$

Since for every PSD matrices $A, B \succeq \mathbf{0}$, $\big\|\sqrt{A}-\sqrt{B}\big\|_2 \leq \sqrt{\|A-B\|_2}$ holds [43], the above inequality also proves that

$$\mathbb{P}\Big(\big\|\sqrt{\widetilde{K}_{\mathbf{X},\rho\mathbf{Y}}}-\sqrt{K_{\mathbf{X},\rho\mathbf{Y}}}\big\|_2 \geq \epsilon\Big) \leq \mathbb{P}\Big(\sqrt{\big\|\widetilde{K}_{\mathbf{X},\rho\mathbf{Y}}-K_{\mathbf{X},\rho\mathbf{Y}}\big\|_2} \geq \epsilon\Big) \tag{9}$$

$$\leq (m+n)^2 \exp\Big(-\frac{r\epsilon^4}{2(1+\rho)^2}\Big). \tag{10}$$

Next, we note the following bound on the Frobenius norm of the difference between $\underline{K}$ and $\widetilde{\underline{K}}$:

$$\begin{aligned}
\Big\|\widetilde{\underline{K}}-\underline{K}\Big\|_F &= \Big\|\sqrt{\widetilde{K}_{\mathbf{X},\rho\mathbf{Y}}}\,D\,\sqrt{\widetilde{K}_{\mathbf{X},\rho\mathbf{Y}}}-\sqrt{K_{\mathbf{X},\rho\mathbf{Y}}}\,D\,\sqrt{K_{\mathbf{X},\rho\mathbf{Y}}}\Big\|_F \\
&\overset{(d)}{\leq} \Big\|\sqrt{\widetilde{K}_{\mathbf{X},\rho\mathbf{Y}}}\,D\,\sqrt{\widetilde{K}_{\mathbf{X},\rho\mathbf{Y}}}-\sqrt{\widetilde{K}_{\mathbf{X},\rho\mathbf{Y}}}\,D\,\sqrt{K_{\mathbf{X},\rho\mathbf{Y}}}\Big\|_F \\
&\quad + \Big\|\sqrt{\widetilde{K}_{\mathbf{X},\rho\mathbf{Y}}}\,D\,\sqrt{K_{\mathbf{X},\rho\mathbf{Y}}}-\sqrt{K_{\mathbf{X},\rho\mathbf{Y}}}\,D\,\sqrt{K_{\mathbf{X},\rho\mathbf{Y}}}\Big\|_F \\
&\overset{(e)}{\leq} \Big\|\sqrt{\widetilde{K}_{\mathbf{X},\rho\mathbf{Y}}}\,D\Big\|_F\Big\|\sqrt{\widetilde{K}_{\mathbf{X},\rho\mathbf{Y}}}-\sqrt{K_{\mathbf{X},\rho\mathbf{Y}}}\Big\|_2 \\
&\quad + \Big\|\sqrt{\widetilde{K}_{\mathbf{X},\rho\mathbf{Y}}}-\sqrt{K_{\mathbf{X},\rho\mathbf{Y}}}\Big\|_2\Big\|D\sqrt{K_{\mathbf{X},\rho\mathbf{Y}}}\Big\|_F \\
&\overset{(f)}{=} \mathrm{Tr}\Big(\widetilde{K}_{\mathbf{X},\rho\mathbf{Y}}\Big)\Big\|\sqrt{\widetilde{K}_{\mathbf{X},\rho\mathbf{Y}}}-\sqrt{K_{\mathbf{X},\rho\mathbf{Y}}}\Big\|_2 \\
&\quad + \Big\|\sqrt{\widetilde{K}_{\mathbf{X},\rho\mathbf{Y}}}-\sqrt{K_{\mathbf{X},\rho\mathbf{Y}}}\Big\|_2\mathrm{Tr}\Big(K_{\mathbf{X},\rho\mathbf{Y}}\Big) \\
&\overset{(g)}{=} 2(1+\rho)\Big\|\sqrt{\widetilde{K}_{\mathbf{X},\rho\mathbf{Y}}}-\sqrt{K_{\mathbf{X},\rho\mathbf{Y}}}\Big\|_2.
\end{aligned}$$

Here, (d) follows from the triangle inequality for the Frobenius norm. (e) follows from the inequality $\|AB\|_F \leq \min\big\{\|A\|_F\|B\|_2, \|A\|_2\|B\|_F\big\}$ for every square matrices $A, B \in \mathbb{R}^{d\times d}$. (f) holds because $\|D\sqrt{K}\|_F = \mathrm{Tr}(\sqrt{K}D^2\sqrt{K}) = \mathrm{Tr}(K)$, and (g) comes from the fact that $\mathrm{Tr}(K_{\mathbf{X},\rho\mathbf{Y}}) = 1+\rho$. Combining the above results shows the following probabilistic bound:

$$\mathbb{P}\Big(\big\|\widetilde{\underline{K}}-\underline{K}\big\|_F \geq \epsilon\Big) \leq \mathbb{P}\Big(\big\|\sqrt{\widetilde{K}_{\mathbf{X},\rho\mathbf{Y}}}-\sqrt{K_{\mathbf{X},\rho\mathbf{Y}}}\big\|_2 \geq \frac{\epsilon}{2(1+\rho)}\Big)$$

$$\leq (m+n)^2 \exp\Big(-\frac{r\epsilon^4}{32(1+\rho)^6}\Big). \tag{11}$$

According to the eigenvalue-perturbation bound in [21], the following holds

$$\sqrt{\sum_{i=1}^{r'}(\widetilde{\lambda}_i-\lambda_i)^2} \leq \|\widetilde{\underline{K}}-\underline{K}\|_F$$

which proves that

$$\mathbb{P}\Big(\sqrt{\sum_{i=1}^{r'}(\widetilde{\lambda}_i-\lambda_i)^2} \geq \epsilon\Big) \leq (m+n)^2 \exp\Big(-\frac{r\epsilon^4}{4(1+\rho)^3}\Big) \tag{12}$$

Defining $\delta = (m+n)^2\exp\Big(-\frac{r\epsilon^4}{32(1+\rho)^6}\Big)$ implying that $\epsilon = \sqrt{8(1+\rho)^3}\sqrt[4]{\frac{\log\big((m+n)/\delta\big)}{r}}$ will then result in the following which completes the proof of the first part of the theorem:

$$\mathbb{P}\Big(\sqrt{\sum_{i=1}^{r'}(\widetilde{\lambda}_i-\lambda_i)^2} \geq \epsilon\Big) \leq \delta. \tag{13}$$

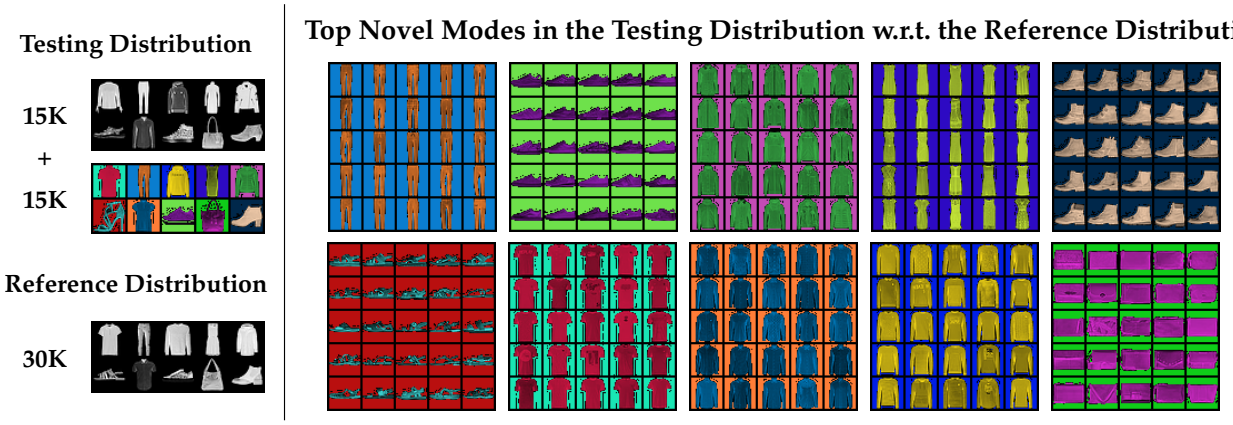

Figure 6. FINC-identified top 10 novel modes between Fashion-MNIST datasets.

Regarding the theorem's approximation guarantee for the eigenvectors, we note that for each eigenvectors $\widetilde{\mathbf{v}}_i$ of the proxy conditional kernel covariance matrix $\widetilde{K}_{\mathbf{X}|\rho\mathbf{Y}}$, we can write the following considering the conditional kernel covariance matrix $K_{\mathbf{X}|\rho\mathbf{Y}}$ and the corresponding eigenvalue λ_i:

$$\begin{aligned}
\left\|K_{\mathbf{X}|\rho\mathbf{Y}}\widetilde{\mathbf{v}}_i - \lambda_i\widetilde{\mathbf{v}}_i\right\|_2 &\leq \left\|K_{\mathbf{X}|\rho\mathbf{Y}}\widetilde{\mathbf{v}}_i - \tilde{\lambda}_i\widetilde{\mathbf{v}}_i\right\|_2 + \left\|\tilde{\lambda}_i\widetilde{\mathbf{v}}_i - \lambda_i\widetilde{\mathbf{v}}_i\right\|_2 \\
&= \left\|K_{\mathbf{X}|\rho\mathbf{Y}}\widetilde{\mathbf{v}}_i - \widetilde{K}_{\mathbf{X}|\rho\mathbf{Y}}\widetilde{\mathbf{v}}_i\right\|_2 + \left|\tilde{\lambda}_i - \lambda_i\right| \\
&\leq \left\|K_{\mathbf{X}|\rho\mathbf{Y}} - \widetilde{K}_{\mathbf{X}|\rho\mathbf{Y}}\right\|_2\left\|\widetilde{\mathbf{v}}_i\right\|_2 + \left|\tilde{\lambda}_i - \lambda_i\right| \\
&= \left\|K_{\mathbf{X}|\rho\mathbf{Y}} - \widetilde{K}_{\mathbf{X}|\rho\mathbf{Y}}\right\|_2 + \left|\tilde{\lambda}_i - \lambda_i\right| \\
&\leq \left\|K_{\mathbf{X}|\rho\mathbf{Y}} - \widetilde{K}_{\mathbf{X}|\rho\mathbf{Y}}\right\|_F + \left|\tilde{\lambda}_i - \lambda_i\right|
\end{aligned}$$

Therefore, given our approximation guarantee on the eigenvalues and that $\sqrt{\sum_{i=1}^{r'}(\widetilde{\lambda}_i - \lambda_i)^2} \leq \epsilon$ implies $\max_{1\leq i\leq r'}\left|\widetilde{\lambda}_i - \lambda_i\right| \leq \epsilon$, we can write the following

$$\mathbb{P}\Big(\max_{1\leq i\leq r'}\left\|K_{\mathbf{X}|\rho\mathbf{Y}}\widetilde{\mathbf{v}}_i - \lambda_i\widetilde{\mathbf{v}}_i\right\|_2 \geq 2\epsilon\Big) \leq \delta. \tag{14}$$

Note that the event under which we proved the closeness of the eigenvalues also requires that the Frobenius norm of $\|\widetilde{K}_{\mathbf{X}|\rho\mathbf{Y}} - K_{\mathbf{X}|\rho\mathbf{Y}}\|_F = \|\widetilde{K}_{\mathbf{X},\rho\mathbf{Y}} - K_{\mathbf{X},\rho\mathbf{Y}}\|_F \leq \epsilon$ to hold. On the other hand, we note that $K_{\mathbf{X}|\rho\mathbf{Y}} = DK_{\mathbf{X},\rho\mathbf{Y}} = D\Phi^\top\Phi$ which leads to the conditional kernel covariance matrix $\Lambda_{\mathbf{X}|\rho\mathbf{Y}} = \Phi D\Phi^\top$. Therefore, if we define $\widetilde{\mathbf{u}}_i = \Phi\widetilde{\mathbf{v}}_i$ we will have

$$\begin{aligned}
\left\|\Lambda_{\mathbf{X}|\rho\mathbf{Y}}\widetilde{\mathbf{u}}_i - \lambda_i\widetilde{\mathbf{u}}_i\right\| &= \left\|\Phi K_{\mathbf{X}|\rho\mathbf{Y}}\widetilde{\mathbf{v}}_i - \lambda_i\Phi\widetilde{\mathbf{v}}_i\right\|_2 \\
&\leq \left\|\Phi\right\|_2\left\|K_{\mathbf{X}|\rho\mathbf{Y}}\widetilde{\mathbf{v}}_i - \lambda_i\widetilde{\mathbf{v}}_i\right\|_2 \\
&\leq (1+\rho)\left\|K_{\mathbf{X}|\rho\mathbf{Y}}\widetilde{\mathbf{v}}_i - \lambda_i\widetilde{\mathbf{v}}_i\right\|_2
\end{aligned}$$

We can combine the above results to show the following and complete the proof:

$$\mathbb{P}\Big(\max_{1\leq i\leq r'}\left\|\Lambda_{\mathbf{X}|\rho\mathbf{Y}}\widetilde{\mathbf{u}}_i - \lambda_i\widetilde{\mathbf{u}}_i\right\|_2 \geq 2(1+\rho)\epsilon\Big) \leq \delta. \tag{15}$$

B. Additional Numerical Results

B.1. Detailed Experiment Setting

Feature extraction: Recent studies [31, 55] show current score-based evaluation may be biased in some applications, and the Inception-V3 embedding, which is retrieved from an ImageNet-pretrained network, may lead to unfair comparisons as

well. To this end, these works propose replacing Inception-V3 with CLIP [44] and DINOv2 [38] embeddings. Following [55], we adopt DINOv2 embedding in our image-based differential clustering experiments. Also, we downloaded the pre-trained generative models from DINOv2 [55] repository. In our experiments, we present results both from Inception-V3 and DINOv2 embedding.

Hyper-parameters selection and sample size: Similar to the kernel-based evaluation in [24], we chose the kernel bandwidth σ^2 by searching for the smallest σ value resulting in a variance bounded by 0.01. Also, we conducted the experiments using at least $m, n = 50$k sample sizes for the test and reference generative models, with the exception of dataset AFHQ and ImageNet-dogs, where the dataset's size was smaller than 50k. For these datasets, we used the entire dataset in the differential clustering. For Fourier feature size r, we tested r values in $\{1000, 2000, 4000\}$ to find the size obtaining an approximation error estimate below 1%, and $r = 2000$ consistently fulfilled the goal.

B.2. Qualitative Comparison between baselines and FINC on ImageNet and AFHQ Dataset

To compare the differential clustering results between the novelty score-based baseline methods and our proposed FINC method, we considered the baseline methods following the FLD score [25], Rarity score [15], and KEN score [61]. In the experiments concerning the FLD and Rarity score-based baselines, we used the methods to detect novel samples, and subsequently we performed the spectral clustering from the TorchCluster package. For the baseline methods, we utilized the largest sample size 10k which did not lead to the memory overflow issue in our GPU processors.

Figure 7 presents our numerical results on the ImageNet dataset, suggesting that the novel sample clusters in the Rarity score and FLD baselines may be semantically dissimilar in some cases. On the other hand, we observed that the novel samples within the clusters found by the KEN-based baseline method are more semantically similar. However, a few clusters found by the KEN method may contain more than one mode. However, the proposed FINC method seems to perform a semantically meaningful clustering of the modes in the large-scale ImageNet setting.

We also performed experiments on the AFHQ dataset containing only 15k images that is fewer than 1.4 million images in ImageNet-1K. The baseline methods FLD and KEN methods seemed to perform with higher accuracy on the AFHQ dataset in comparison to the ImageNet-1K dataset. As displayed in Figure 8, FLD seems capable of detecting cat-like modes. The proposed FINC and the baseline KEN method led to nearly identical found modes.

B.3. Experimental Results on Video Datasets

In this section, we compare Kinetics-400 [28] with UCF101 [54] dataset. Kinetics-400 includes 400 human action categories, each with a minimum of 400 video clips depicting the action, while UCF101 consists of 101 actions. Similar to video evaluation metrics [49, 58] we used the I3D pre-trained model [8] which maps each video to a 1024-vector feature. In this experiment, we considered Kinetics-400 as the test and UCF101 as our reference dataset. The top 9 novel modes are shown in Figure 12 and none of these categories are included in UCF101.

B.4. Experimental Results on Text-to-Image Models

B.4.1. Evaluation of text-to-image Models

We assessed GigaGAN [26], a state-of-the-art text-to-image generative model on the MS-COCO dataset [33]. We conducted a zero-shot evaluation on the 30K images from the MS-COCO validation set. As shown in Figure 14, the model showed some novelty in generating pictures of a bathroom and sea with a bird or human on it. However, our FINC-based evaluation suggests that the model may be less capable in producing images of zebras, giraffes, or trains.

B.4.2. Effect of different feature extractors

As shown in the literature [31], Inception-V3 may lead to a biased evaluation metric. In this experiment, instead of using pre-trained Inception-V3 to extract features, we used pre-trained models SwAV [7] and ResNet50 [16] as feature extractors and attempted to investigate the effect of different pre-trained models on our proposed method. Figure 18 shows top-3 novel models of GigaGAN [26] using SwAV and ResNet50 models. The top two novel modes were the same among the different feature extractors indicating that our method is not biased on the pre-trained networks. In the third mode, SwAV and ResNet50 results were different showing that each of these models captured different novel modes based on the features they extracted from the data.

B.5. Detection of Sample Clusters with Higher and Lower CLIPScores

The CLIPScore [18] is a popular sample-based metric to measure the alignment of the generated image by a text-to-image model. To identify the sample types with the highest and lowest CLIPScores, we ran a differential clustering experiment

between the CLIPScore-based top 20% and the bottom 20% of samples generated by three text-to-image models, GigaGAN, SD-XL, and Kandinsky in response to text prompts from MS-COCO dataset. In Figure 15, we show the top-3 and bottom-3 FINC-identified modes with the maximum and minimum CLIPScores for each text-to-image model. We observe similar clusters in the top modes, indicating that CLIPScore performs better in clusters related to zebras or cats/dogs, however, it does not perform well in bottom clusters such as buses or electronic devices.

B.6. Detection of High Memorization-Score Modes

In the literature, memorization scores have been proposed to detect memorizing training samples by generative models. Although sample-level memorization metrics have been used to quantify and rank the similarity of every generated sample to training data, these scores do not directly reveal which sample types could be memorized more frequently by a generative model. To address this task, we apply FINC to conduct differential clustering on the groups of samples with low and high memorization FLD scores.

To do this, we evaluated and ranked the sample-level memorization score for all the generated samples using the FLD memorization metric with respect to the ImageNet training samples. Subsequently, we separated the two groups f samples with the top 20% and bottom 20% of FLD memorization scores. We performed differential clustering via FINC between the two groups, and used the eigenvalues/eigenvectors computed by Algorithm 1 to quantify and rank mode-level memorization scores. Figure 11 displays the top two detected sample types that occur more frequently in the high-memorization samples of the generative models, and the most similar ImageNet training samples to the generated modes.

B.7. Comparative Mode Frequency Analysis between Generative Models

B.7.1. Evaluation of StyleGAN-XL on ImageNet

Figures 19 and 20 illustrate the more frequently and less frequently generated modes of StyleGAN-XL in comparison to the reference models. These figures repeat the experiments presented in Figures 4 and 17 on StyleGAN-XL.

B.7.2. Novel Modes of DiT-XL on ImageNet

Figure 21 illustrates the novel modes of DiT-XL with novelty threshold $\rho = 5$ in comparison to the reference models. For this experiment, we use a sample size of 100k in each distribution.

B.8. Analyzing the Effect of σ Hyperparameter

We examined the effect of σ hyper-parameter of the Gaussian kernel in (1). We used the trace of the covariance matrix to measure the heterogeneity of samples within each mode. Based on our numerical evaluations shown in Figure 10, we observed that smaller values of σ could capture more specific modes containing samples with a seemingly lower variety. On the other hand, a greater σ value seemed to better capture more general modes, displaying diverse samples and a higher variance statistic.

B.9. Computational Software and Hardware

We leverage the PyTorch framework for doing the experiments. All experiments are conducted in a cluster configured with dual 24-core CPU processors with 500 GB memory and 8 RTX 3090 GPU processors with 24 GB memory.

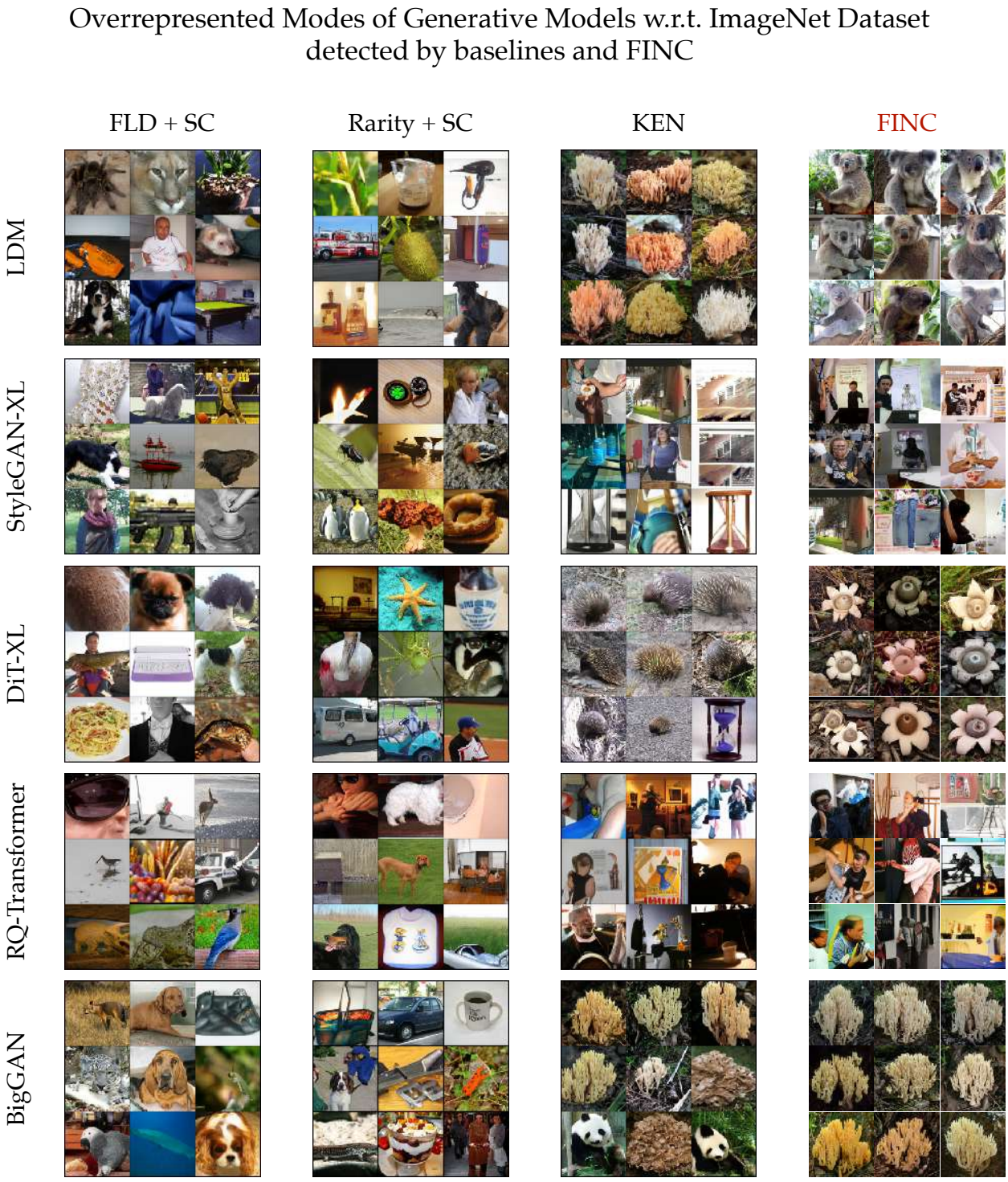

Figure 7. Baseline-identified and FINC-identified top overrepresented modes expressed by test generative models with a considerably higher frequency than in the reference ImageNet dataset. DINOv2 embedding is used. For baseline methods, we use the applicable largest sample size 10k in the GPU computation. For the FINC method, we use 50k samples instead.

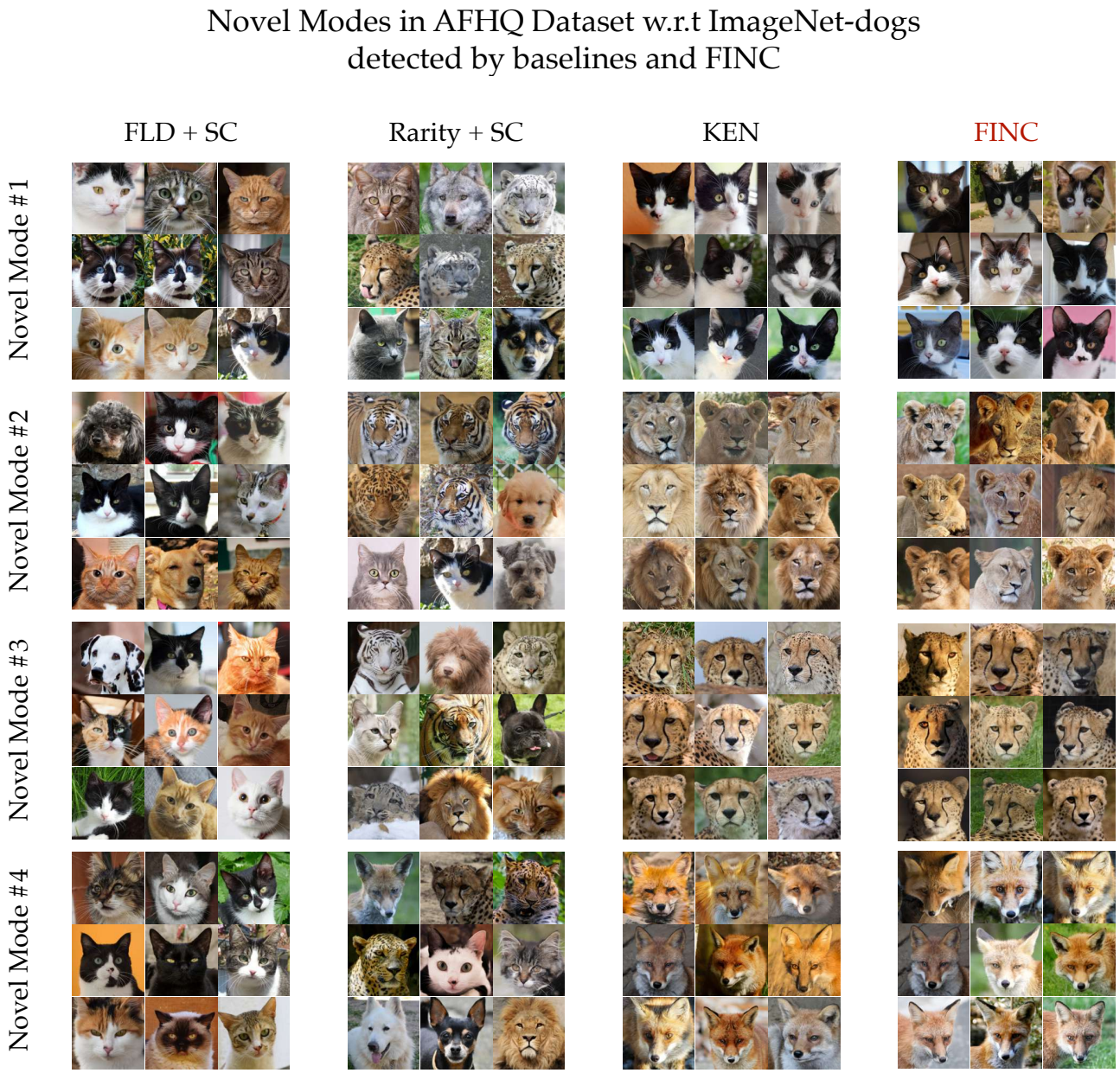

Figure 8. Baseline-identified and FINC-identified novel modes of AFHQ with respect to ImageNet-dog dataset. Inception-V3 embedding is used. For baseline methods, we use the applicable largest sample size 10k in the GPU computation. For the FINC method, we use 50k samples instead.

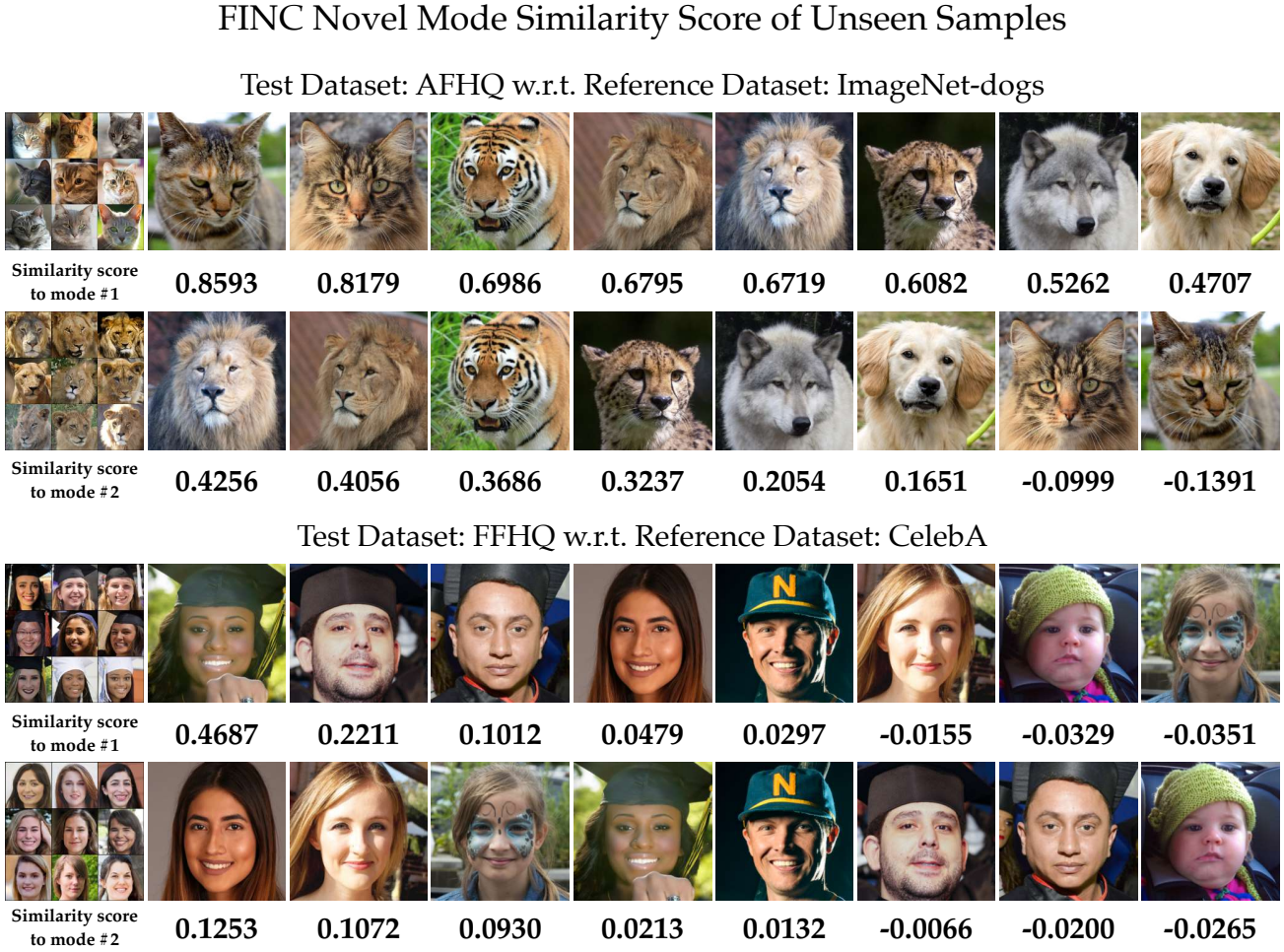

Figure 9. Computing FINC-assigned mode score for 8 unseen samples and 2 detected modes. Images looking more similar to the modes are supposed to have higher scores. (Top-half) AFHQ w.r.t. ImageNet-dogs, (Bottom-half) FFHQ w.r.t. CelebA.

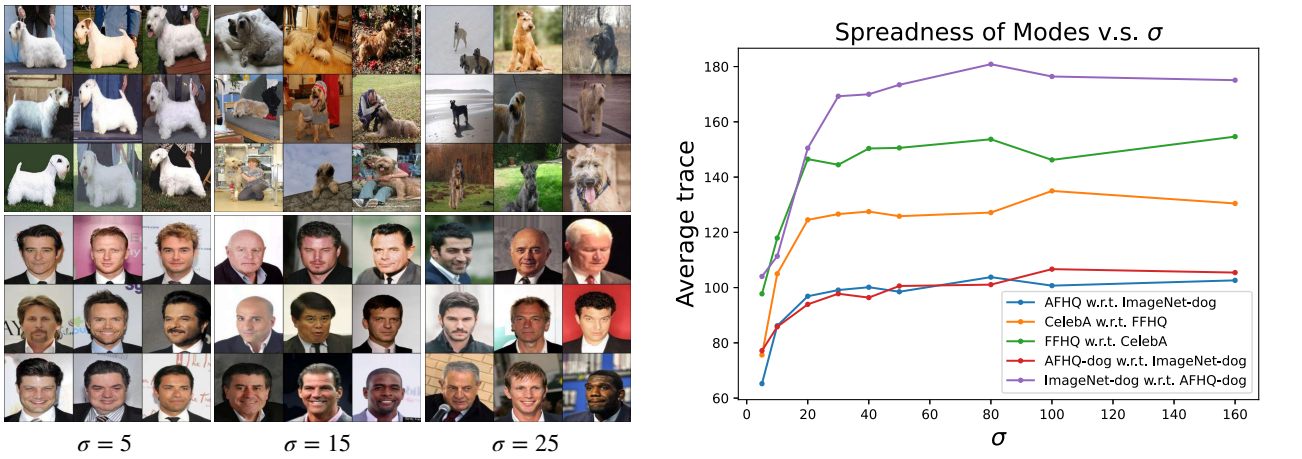

Figure 10. Effect analysis of hyperparameter σ. Modes captured by smaller σ are more specific with similar samples and lower spreadness statistic, while modes captured by larger σ are more general with diverse samples and higher spreadness statistic.

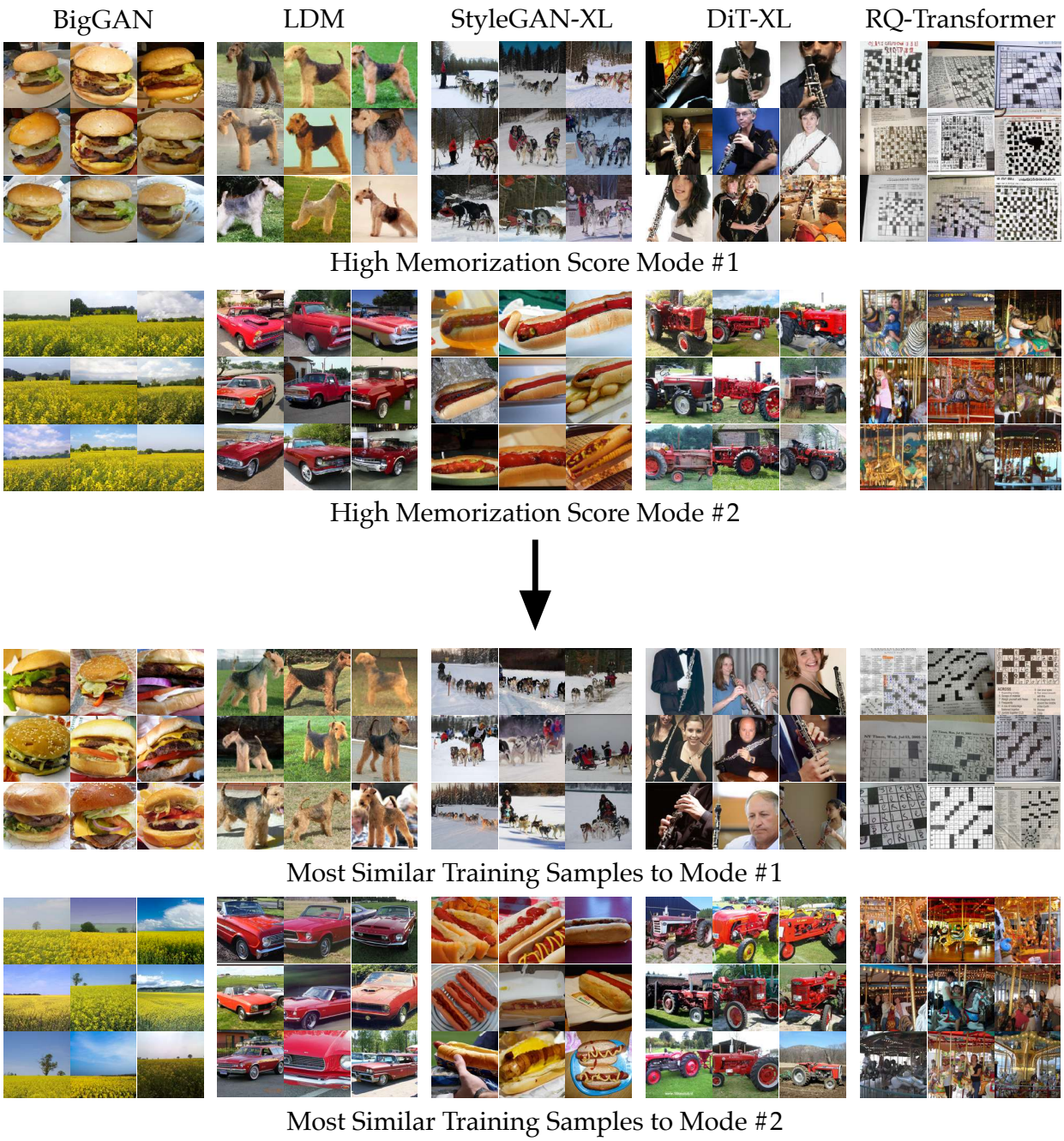

Figure 11. **Top:** The FINC-identified top-2 generated sample clusters with higher FLD memorization scores w.r.t. the ImageNet training dataset. The top 9 images with the maximum alignment with the detected mode's eigenvectors are shown. **Bottom:** The most similar training samples retrieved from ImageNet training set.

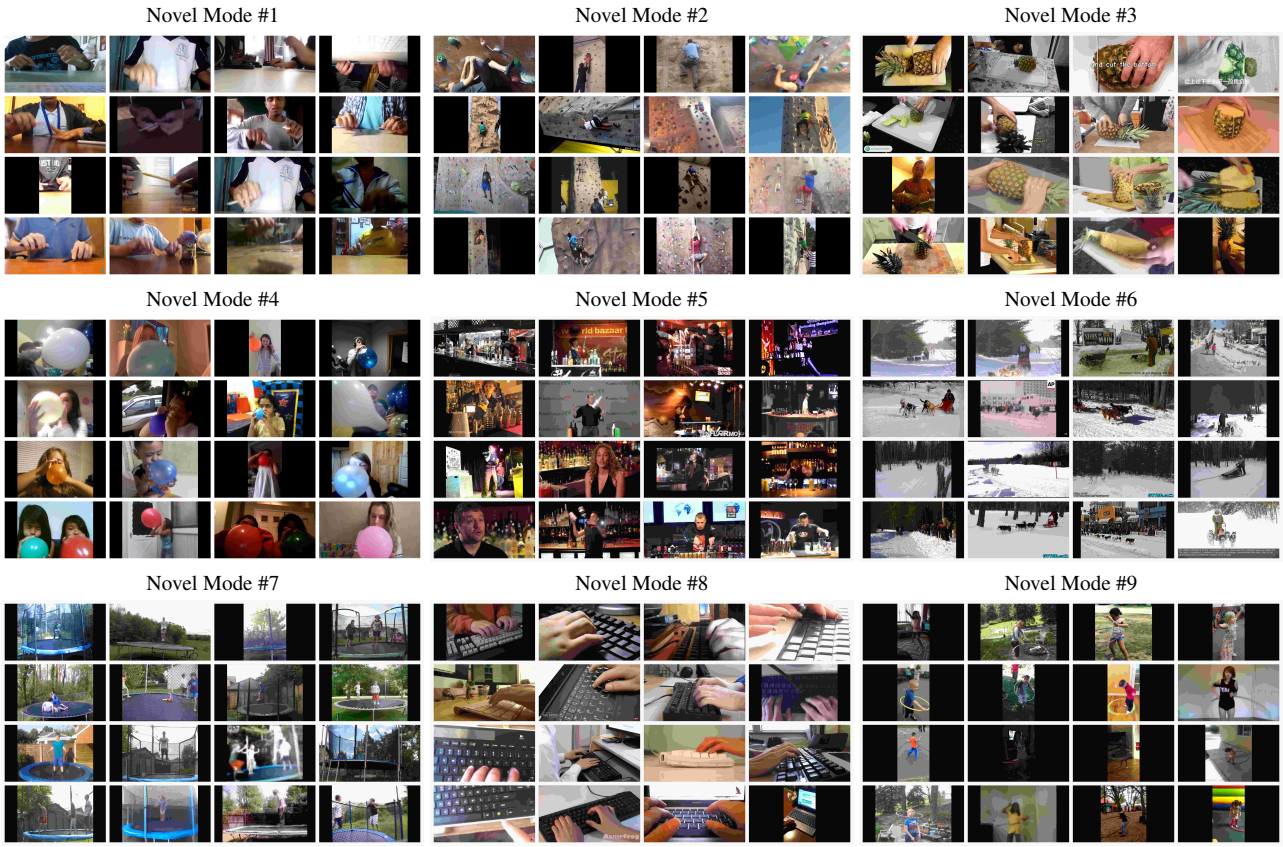

Figure 12. Top 9 novel modes of Kinetics-400 dataset based on FINC evaluation with UCF101 as the reference dataset.

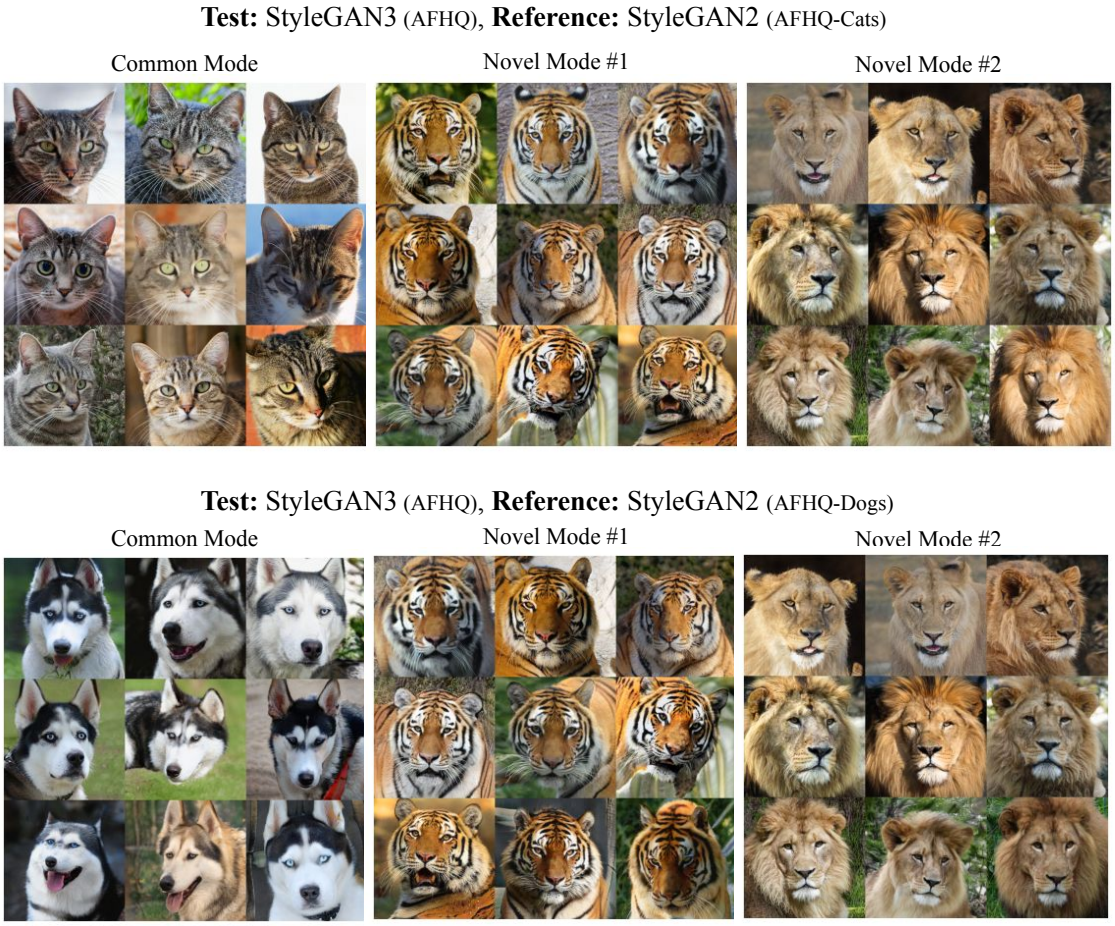

Figure 13. Identified top-2 novel modes and the most common mode between the StyleGAN3 and StyleGAN2 trained on AFHQ

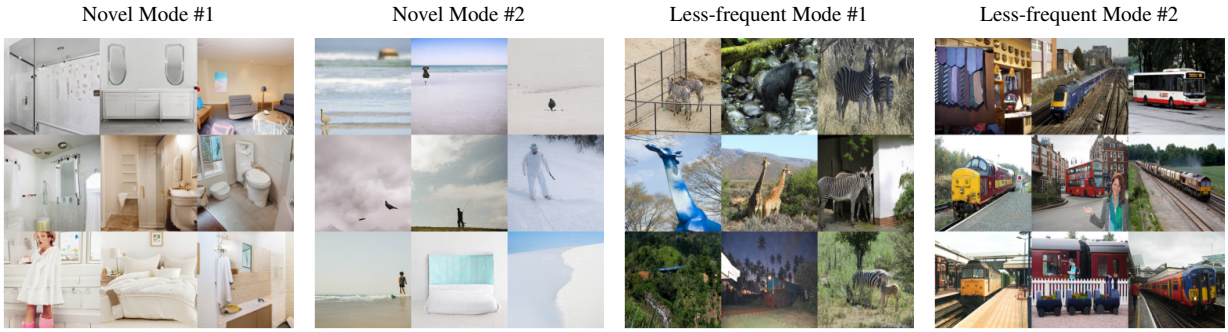

Figure 14. Identified top-2 novel modes and top-2 less-frequent modes on the GigaGAN text-to-image model generated on the MS-COCO dataset.

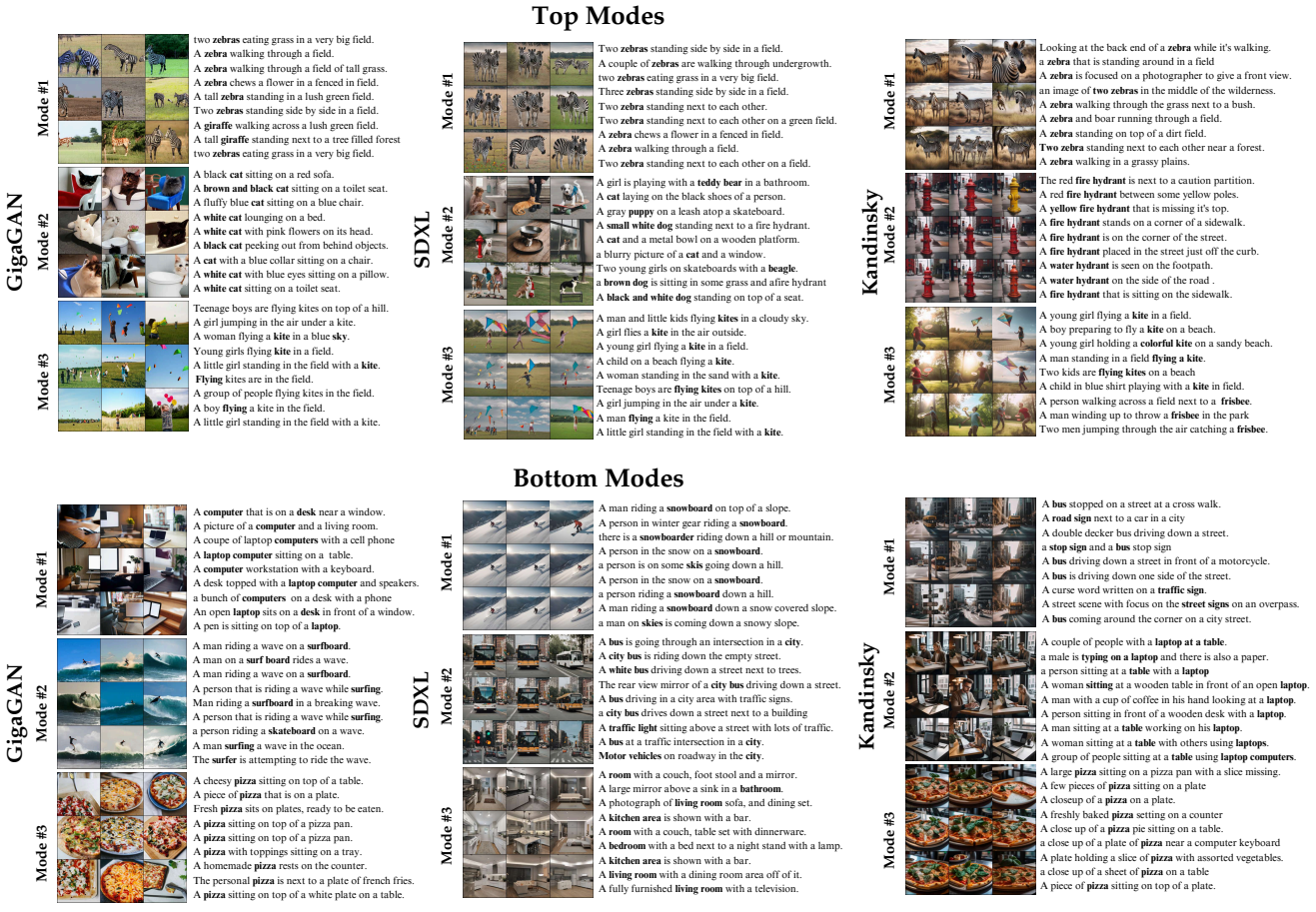

Figure 15. The FINC-identified top-3 and bottom-3 sample clusters according to the CLIPScore of text-to-image generative models's data.

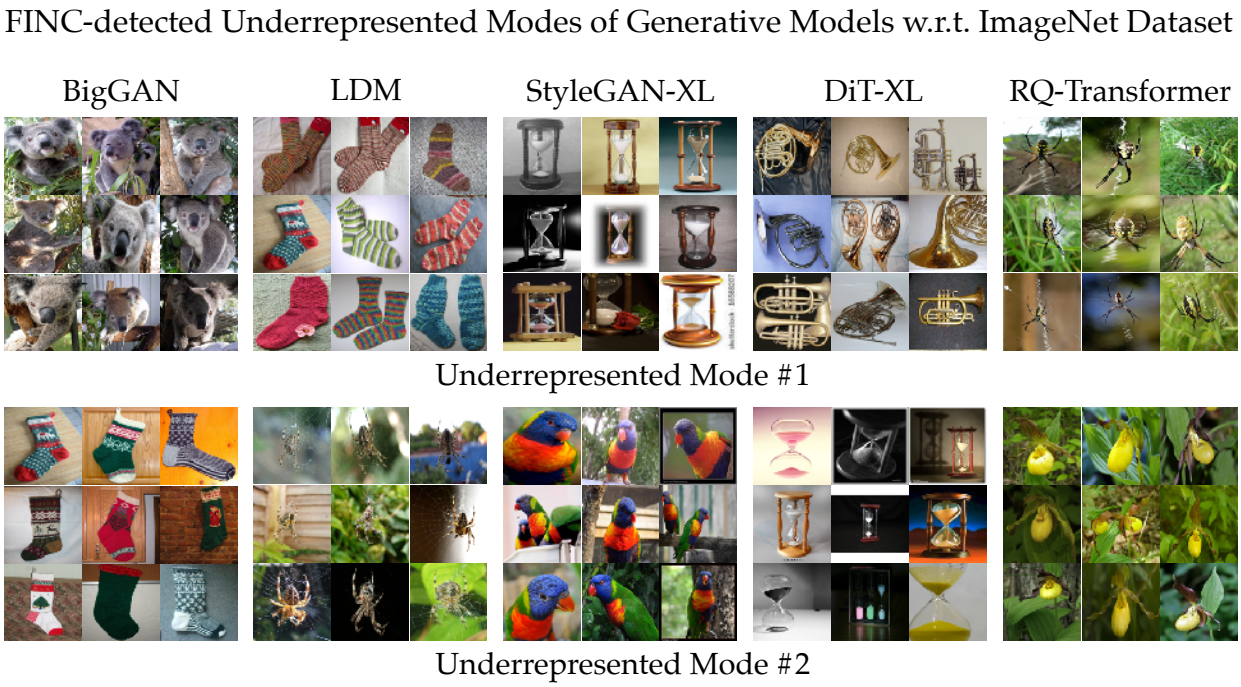

Figure 16. The FINC-identified top two underrepresented modes expressed by test generative models with a lower frequency than in the reference ImageNet dataset. Embedding is from DINOv2.

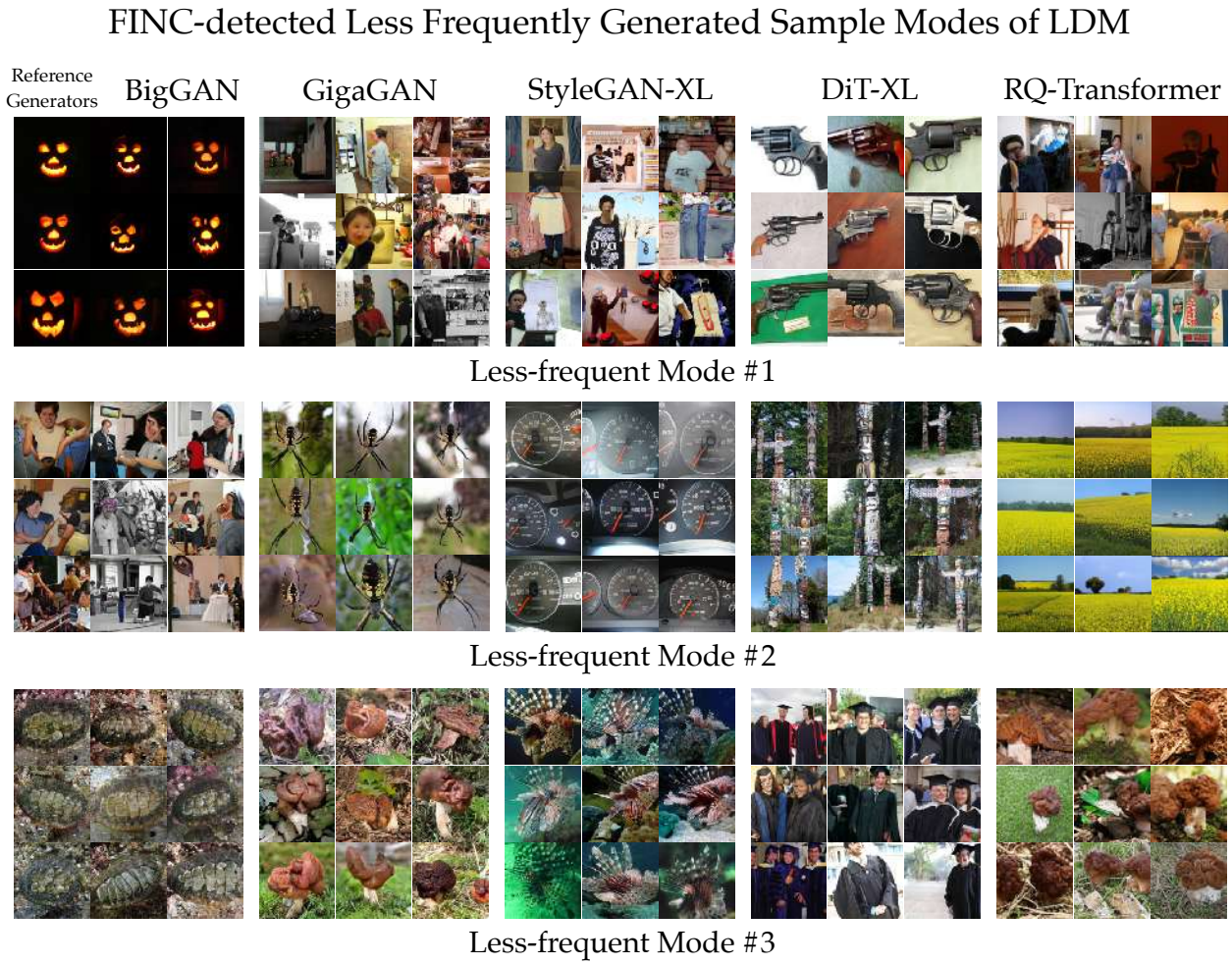

Figure 17. Representative samples from FINC-detected top-3 modes with the maximum frequency gap between LDM (lower frequency) and reference generative models. Embedding is from DINOv2.

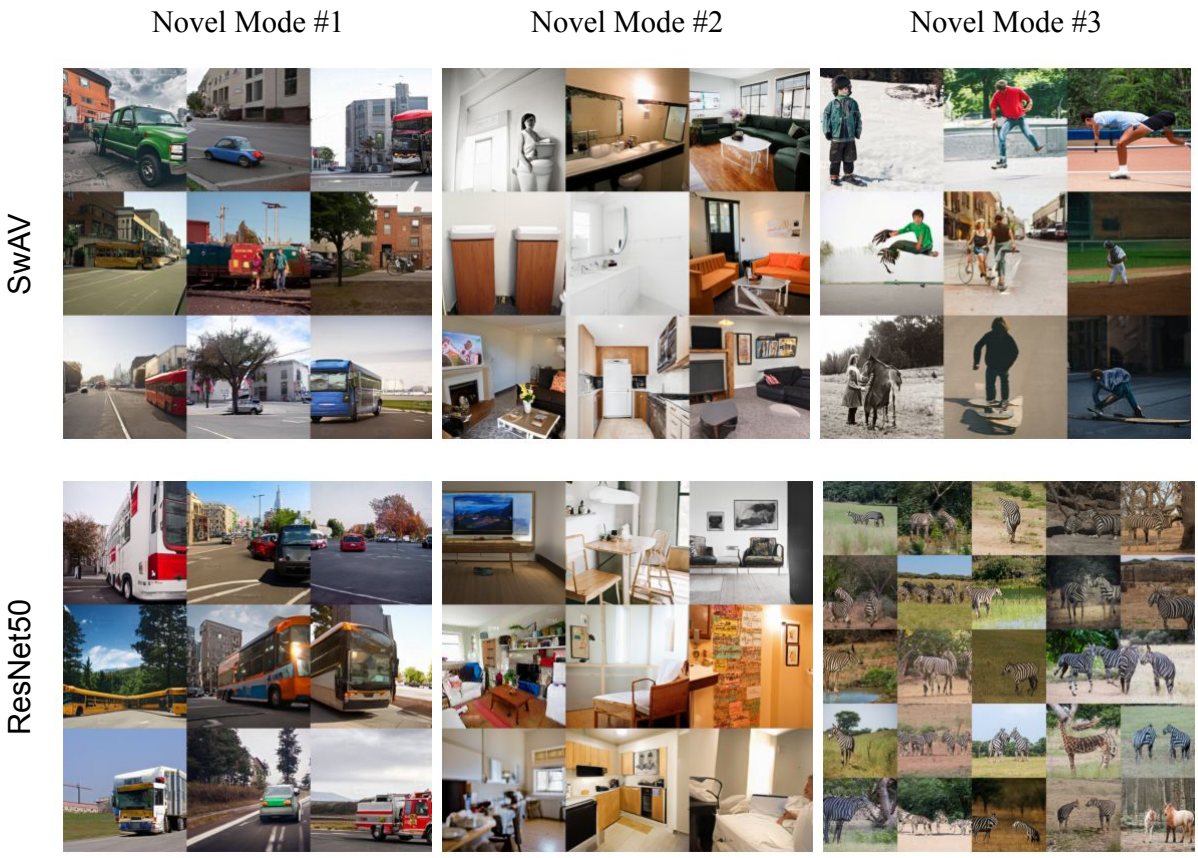

Figure 18. Identified top-3 novel modes on the GigaGAN text-to-image model generated on the MS-COCO dataset using SwAV and ResNet50 pre-trained models as feature extractors.

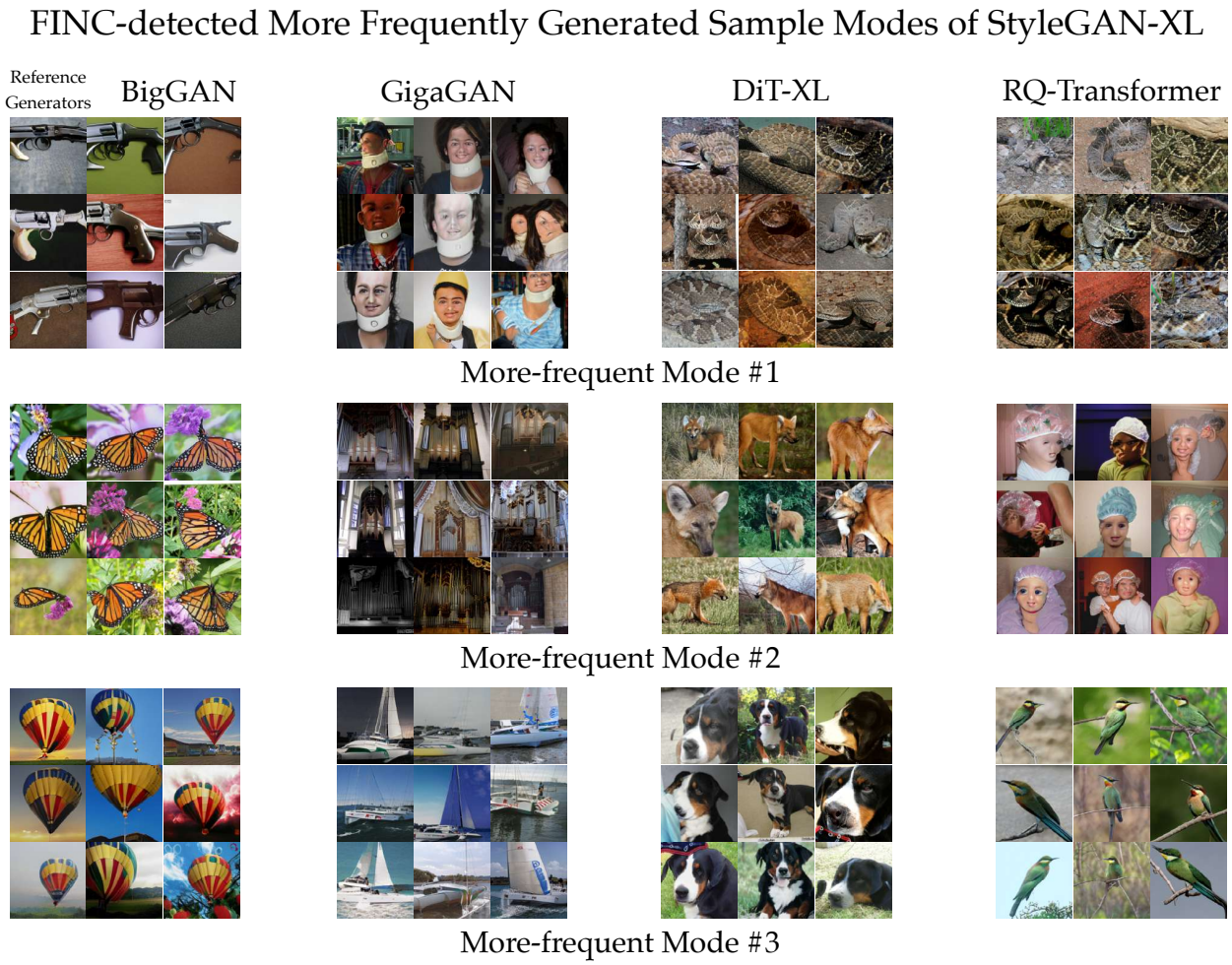

Figure 19. Representative samples from FINC-detected top-3 modes with the maximum frequency gap between StyleGAN-XL (higher frequency) and reference generative models.

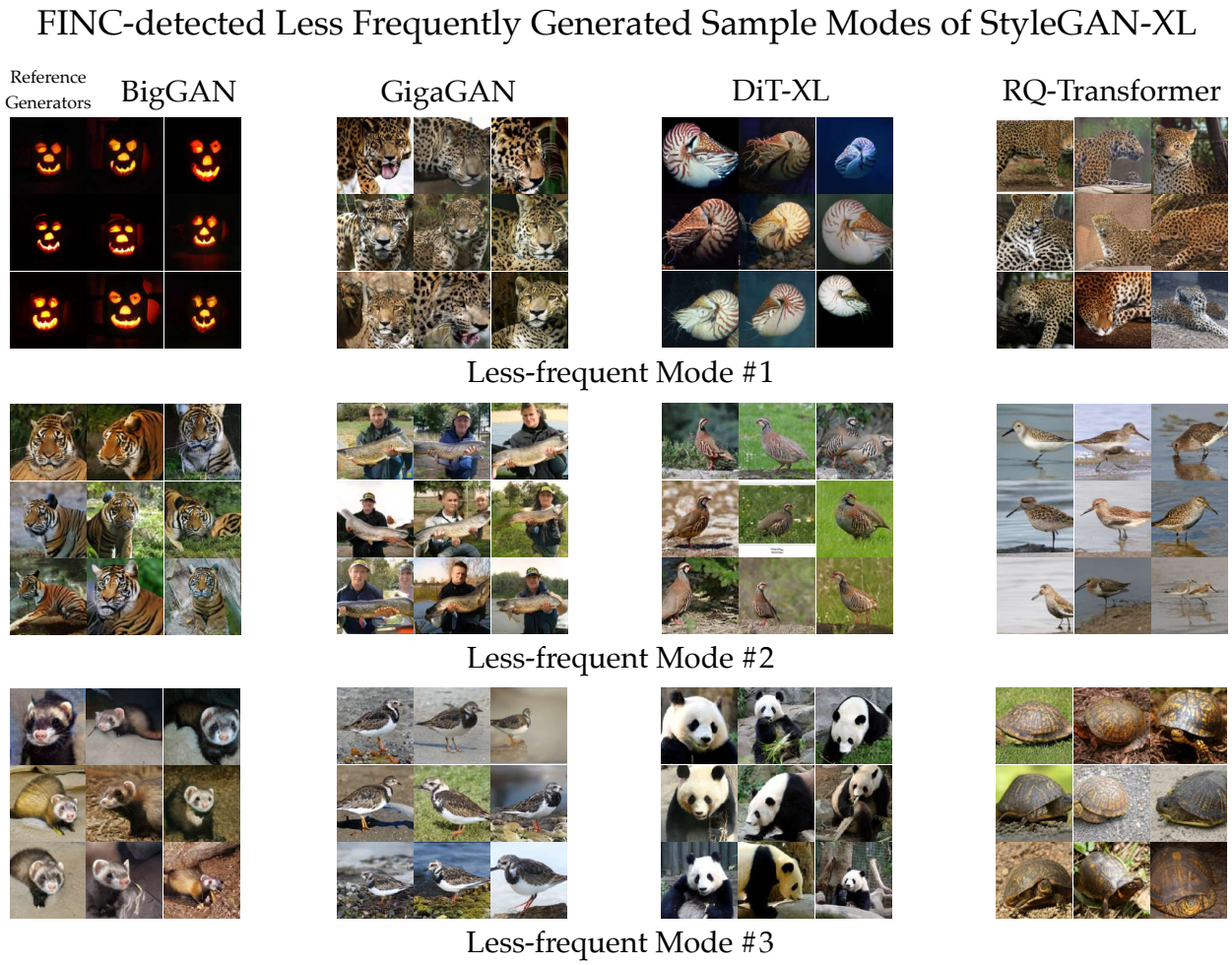

Figure 20. Representative samples from FINC-detected top-3 modes with the maximum frequency gap between StyleGAN-XL (lower frequency) and reference generative models.

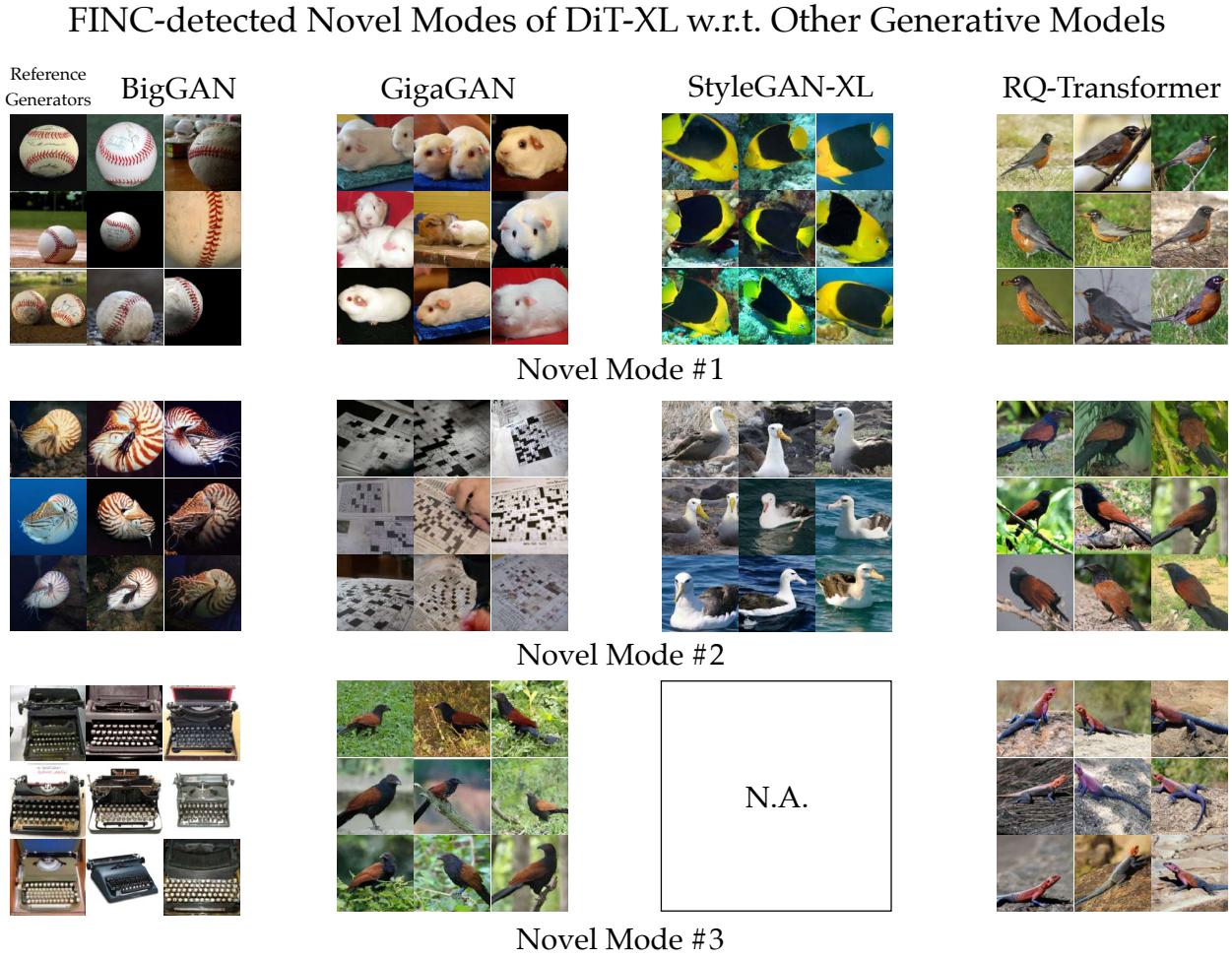

Figure 21. Representative samples from FINC-detected top-3 novel modes (novelty threshold $\rho = 5$) of DiT-XL compared to reference generative models. "N.A." indicates that DiT-XL has fewer than three modes with a frequency $\rho = 5$ times higher than the reference model. Additionally, it is worth noting that FINC does not detect any modes of DiT-XL with a frequency $\rho = 5$ times higher than that of LDM.

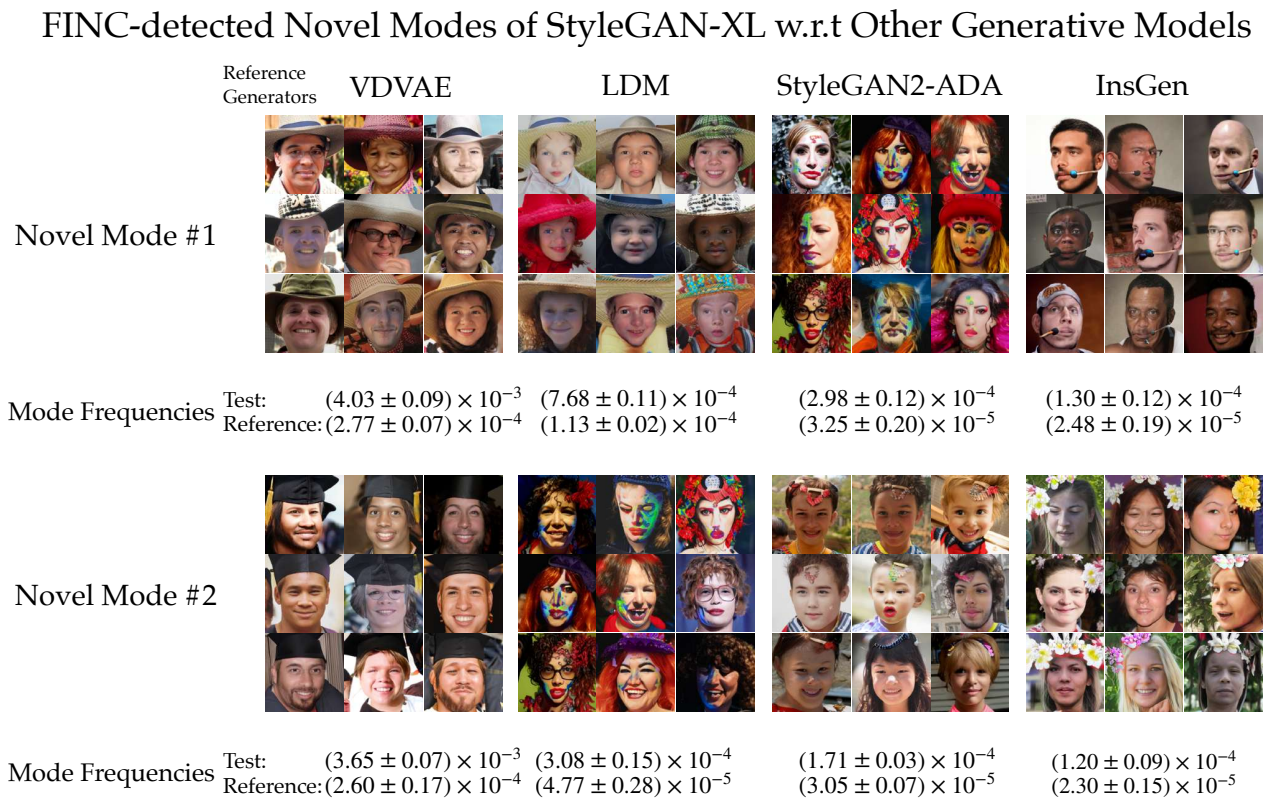

Figure 22. Representative samples from FINC-detected top-3 novel modes (novelty threshold $\rho = 5$) of StyleGAN-XL compared to reference generative models.