Towards a Scalable Reference-Free Evaluation of Generative Models

Azim Ospanov[*]
aospanov9@cse.cuhk.edu.hk

Jingwei Zhang[*]
jwzhang22@cse.cuhk.edu.hk

Mohammad Jalali[*]
mjalali24@cse.cuhk.edu.hk

Xuenan Cao [†]
xuenancao@cuhk.edu.hk

Andrej Bogdanov [‡]
abogdano@uottawa.ca

Farzan Farnia[*]
farnia@cse.cuhk.edu.hk

Abstract

While standard evaluation scores for generative models are mostly reference-based, a reference-dependent assessment of generative models could be generally difficult due to the unavailability of applicable reference datasets. Recently, the reference-free entropy scores, VENDI [1] and RKE [2], have been proposed to evaluate the diversity of generated data. However, estimating these scores from data leads to significant computational costs for large-scale generative models. In this work, we leverage the random Fourier features framework to reduce the computational price and propose the *Fourier-based Kernel Entropy Approximation (FKEA)* method. We utilize FKEA's approximated eigenspectrum of the kernel matrix to efficiently estimate the mentioned entropy scores. Furthermore, we show the application of FKEA's proxy eigenvectors to reveal the method's identified modes in evaluating the diversity of produced samples. We provide a stochastic implementation of the FKEA assessment algorithm with a complexity $O(n)$ linearly growing with sample size n. We extensively evaluate FKEA's numerical performance in application to standard image, text, and video datasets. Our empirical results indicate the method's scalability and interpretability applied to large-scale generative models. The codebase is available at `https://github.com/aziksh-ospanov/FKEA`.

1 Introduction

A quantitative comparison of generative models requires evaluation metrics to measure the quality and diversity of the models' produced data. Since the introduction of variational autoencoders (VAEs) [3], generative adversarial networks (GANs) [4], and diffusion models [5] that led to impressive empirical results in the last decade, several evaluation scores have been proposed to assess generative models learned by different training methods and architectures. Due to the key role of evaluation criteria in comparing generative models, they have been extensively studied in the literature.

While various statistical methods have been applied to measure the fidelity and variety of a generative model's produced data, the standard scores commonly perform a reference-based evaluation of generative models, i.e., they quantify the characteristics of generated samples in comparison to a reference distribution. The reference distribution is usually chosen to be either the distribution of

[*]Department of Computer Science & Engineering, The Chinese University of Hong Kong, Hong Kong
[†]Department of Cultural and Religious Studies, The Chinese University of Hong Kong, Hong Kong
[‡]School of Electrical Engineering and Computer Science, University of Ottawa, Canada

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-3841

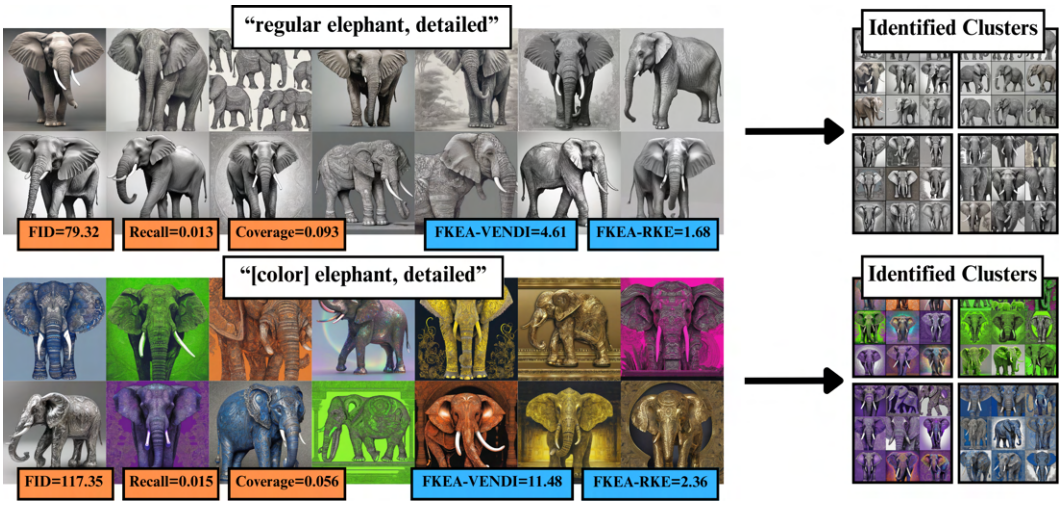


Figure 1: Reference-based vs. reference-free scores on two datasets of Stable Diffusion XL generated elephant images. FID, Recall, and Coverage scores (colored orange) are reference-based, whereas VENDI and RKE scores (colored blue) are reference-free. Inception.V3 is used as the backbone embedding. Reference-based metrics use 'Indian elephant' samples in ImageNet as reference data.

samples in the test data partition or a comprehensive dataset containing a significant fraction of real-world sample types, e.g. ImageNet [6] for evaluating image-based generative models.

To provide well-known examples of reference-dependent metrics, note that the distance scores, Fréchet Inception Distance (FID) [7] and Kernel Inception Distance (KID) [8], are explicitly reference-based, measuring the distance between the generative and reference distributions. Similarly, the standard quality/diversity score pairs, Precision/Recall [9, 10] and Density/Coverage [11], perform the evaluation in comparison to a reference dataset. Even the seemingly reference-free Inception Score (IS) [12] can be viewed as implicitly reference-based, since it quantifies the variety and fidelity of data based on the labels and confidence scores assigned by an ImageNet pre-trained neural net, where ImageNet implicitly plays the role of the reference dataset. The reference-based nature of these evaluation scores is desired in many instances including standard image-based generative models, where either a sufficiently large test set or a comprehensive reference dataset such as ImageNet is available for the reference-based evaluation.

On the other hand, a reference-based assessment of generative models may not always be feasible, because the selection of a reference distribution may be challenging in a general learning scenario. For example, in prompt-based generative models where the data are created in response to a user's input text prompts, the generated data could follow an a priori unknown distribution depending on the specific distribution of the user's input prompts. Figure 1 shows one such example where we compare reference-based diversity scores of regular and colored elephant image samples generated by Stable Diffusion XL [13]. While the diversity of the colored images looks significantly higher to the human eye, the evaluated reference-based FID, Recall, and Coverage metrics do not suggest a higher diversity. As this example suggests, a proper reference-based evaluation of every user's generated data would require a distinct reference dataset, which may not be available to the user during the assessment time. Moreover, finding a comprehensive text or video dataset to choose as the reference set would be more difficult compared to image data, because the higher length of text and video samples could significantly contribute to their variety, requiring an inefficiently large reference set to cover all text or video sample types.

The discussed challenging scenarios of conducting a reference-based evaluation highlight the need for reference-free assessment methods that remain functional in the absence of a reference dataset. Recently, entropy-based diversity evaluation scores, the VENDI metric family [1, 14] and RKE score [2], have been proposed to address the need for reference-free assessment metrics. These scores calculate the entropy of the eigenvalues of a kernel similarity matrix for the generated data. Based on the theoretical results in [2], the evaluation process of these scores can be interpreted as an unsupervised identification of the generative model's produced sample clusters, followed by the entropy calculation for the frequencies of the detected clusters. In Figure 1, we observe that the

reference-free VENDI and RKE scores grow when the generated samples are colored, which is due to the increase in the quantity of identified clusters in the colored case.

While the VENDI and RKE entropy scores provide reference-free assessments of generative models, estimating these scores from generated data could incur significant computational costs. In this work, we show that computing the precise RKE and VENDI scores would require at least $\Omega(n^2)$ and $\Omega(n^{2.373})$[4] computations for a sample size n, respectively. While the randomized projection methods in [15, 1] can reduce the computational costs to $O(n^2)$ for a general VENDI_α score, the quadratic growth would be a barrier to the method's application to large n values. Although the computational expenses could be reduced by limiting the sample size n, an insufficient sample size would lead to significant error in estimating the entropy scores. As an example on the ImageNet dataset, Figure 7 in the Appendix shows the adverse effects of limiting the sample size on the quality of clusters used in the calculation of the VENDI scores.

To overcome the challenges of computing the scores, we leverage the random Fourier features (RFFs) framework [16] and develop a scalable entropy-based evaluation method that can be efficiently applied to large sample sizes. Our proposed method, *Fourier-based Kernel Entropy Approximation (FKEA)*, is designed to approximate the kernel covariance matrix using the RFFs drawn from the Fourier transform-inverse of a target shift-invariant kernel. We prove that using a Fourier feature size $r = \mathcal{O}\big(\frac{\log n}{\epsilon^2}\big)$, FKEA computes the eigenspace of the kernel matrix within an ϵ-bounded error. Furthermore, we demonstrate the application of the eigenvectors of the FKEA's proxy kernel matrix for identifying the sample clusters used in the reference-free evaluation of entropic diversity.

Finally, we present numerical results of the entropy-based evaluation of standard generative models using the baseline eigendecomposition and our proposed FKEA methods. In our experiments, the baseline spectral decomposition algorithm could not efficiently scale to sample sizes above a few ten thousand. On the other hand, our stochastic implementation of the FKEA method could scalably apply to large sample sizes. Utilizing the standard embeddings of image, text, and video data, we tested the FKEA assessment while computing the sample clusters and their frequencies in application to large-scale datasets and generative models. Here is a summary of our work's main contributions:

- Characterizing the computational complexity of the kernel entropy scores of generative models,
- Developing the Fourier-based FKEA method to approximate the kernel covariance eigenspace and entropy of generated data,
- Proving guarantees on FKEA's required size of random Fourier features indicating a complexity logarithmically growing with the dataset size,
- Providing numerical results on FKEA's reference-free assessment of large-scale image,text, video-based datasets and generative models.

2 Related Work

Evaluation of deep generative models. The assessment of generative models has been widely studied in the literature. The existing scores either quantify a distance between the distributions of real and generated data, as in FID [7] and KID [8] scores, or attempt to measure the quality and diversity of the trained generative models, including the Inception Score [12], quality/diversity metric pairs Precision/Recall [9, 10] and Density/Coverage [11]. The mentioned scores are reference-based, while in this work we focus on reference-free metrics. Also, we note that the evaluation of memorization and novelty has received great attention, and several scores including the authenticity score [17], the feature likelihood divergence [18], and the rarity score [19] have been proposed to quantify the generalizability and novelty of generated samples. Note that the evaluation of novelty and generalization is, by nature, reference-based. On the other hand, our study focuses on the diversity of data which can be evaluated in a reference-free way as discussed in [1, 2].

Role of embedding in quantitative evaluation. Following the discussion in [20], we utilize DinoV2 [21] image embeddings in most of our image experiments, as [20]'s results indicate DinoV2 can yield scores more aligned with the human notion of diversity. As noted in [22], it is possible to utilize other non-ImageNet feature spaces such as CLIP [23] and SwAV [24] as opposed to InceptionV3 [25] to

[4]This computation complexity is the minimum known achievable cost for multiplying $n \times n$ matrices which we prove to lower-bound the complexity of computing matrix-based entropy scores.

 https://doi.org/10.52202/079017-3841

further improve metrics such as FID. In this work, we mainly focus on DinoV2 feature space, while we note that other feature spaces are also compatible with entropy-based diversity evaluation.

Diversity assessment for text-based models. To quantify the diversity of text data, the n-gram-based methods are commonly used in the literature. A well-known metric is the BLEU score [26], which is based on the geometric average of n-gram precision scores times the Brevity Penalty. To adapt BLEU score to measure text diversity, [27] proposes the Self-BLEU score, calculating the average BLEU score of various generated samples. To further isolate and measure diversity, N-Gram Diversity scores [28, 29, 30] were proposed and defined by a ratio between the number of unique n-grams and overall number of n-grams in the text. Other prominent metrics include Homogenization (ROUGE-L) [31], FBD [32] and Compression Ratios [33].

Kernel PCA, Spectral Cluttering, and Random Fourier Features. Kernel PCA [34] is a well-studied method of dimensionality reduction that utilizes the eigendecomposition of the kernel matrix, similar to the kernel-based diversity evaluation methods in [1, 2]. The related papers [35, 36] study the connections between kernel PCA and spectral clustering. Also, the analysis of random Fourier features [16] for performing scalable kernel PCA has been studied in [37, 38, 39, 40, 41]. We note that while the mentioned works characterize the complexity of estimating the eigenvectors, our analysis focuses on the complexity of computing the kernel matrix's eigenvalues via Fourier features, as we primarily seek to quantify the diversity of generated data using the kernel matrix's eigenvalues.

3 Preliminaries

Consider a generative model $\mathcal{G}$ generating random samples $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ following the model's probability distribution $P_{\mathcal{G}}$. In our analysis, we assume the n generated samples are independently drawn from $P_{\mathcal{G}}$. Note that in VAEs [3] and GANs [4], the generative model $\mathcal{G}$ is a deterministic function $G : \mathbb{R}^r \to \mathbb{R}^d$ mapping an r-dimensional latent random vector $\mathbf{Z} \sim P_Z$ from a known distribution P_Z to $G(\mathbf{Z})$ distributed according to $P_{\mathcal{G}}$. On the other hand, in diffusion models, $\mathcal{G}$ represents an iterative random process that generates a sample from $P_{\mathcal{G}}$. The goal of a sample-based diversity evaluation of generative model $\mathcal{G}$ is to quantify the variety of its generated data $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

3.1 Kernel Function, Kernel Covariance Matrix, and Matrix-based Rényi Entropy

Following standard definitions, $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is called a kernel function if for every integer $n \in \mathbb{N}$ and set of inputs $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$, the kernel similarity matrix $K = \big[k(\mathbf{x}_i, \mathbf{x}_j)\big]_{n\times n}$ is positive semi-definite. We call a kernel function k normalized if for every input $\mathbf{x}$ we have $k(\mathbf{x}, \mathbf{x}) = 1$. A well-known example of a normalized kernel function is the Gaussian kernel $k_{\text{Gaussian}(\sigma^2)}$ with bandwidth parameter σ^2 defined as

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

For every kernel function k, there exists a feature map $\phi : \mathbb{R}^d \to \mathbb{R}^m$ such that $k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$ is the inner product of the m-dimensional feature maps $\phi(\mathbf{x})$ and $\phi(\mathbf{x}')$. Given a kernel k with feature map ϕ, we define the kernel covariance matrix $C_X \in \mathbb{R}^{m\times m}$ of a distribution P_X as

$$C_X \;:=\; \mathbb{E}_{\mathbf{X}\sim P_X}\Big[\phi(\mathbf{X})\phi(\mathbf{X})^\top\Big] \;=\; \int p_X(\mathbf{x})\phi(\mathbf{x})\phi(\mathbf{x})^\top \mathrm{d}\mathbf{x}$$

The above matrix C_X is positive semi-definite with non-negative values. Furthermore, assuming a normalized kernel k, it can be seen that the eigenvalues of C_X will add up to 1 (i.e., it has unit trace $\mathrm{Tr}(C_X) = 1$), providing a probability model. Therefore, one can consider the entropy of C_X's eigenvalues as a quantification of the diversity of distribution P_X based on the kernel similarity score k. Here, we review the general family of Rényi entropy used to define VENDI and RKE scores.

Definition 1. *For a positive semi-definite matrix $C_X \in \mathbb{R}^{m\times m}$ with eigenvalues $\lambda_1, \ldots, \lambda_m$, the order-$\alpha$ Rényi entropy $H_\alpha(C_X)$ for $\alpha > 0$ is defined as*

$$H_\alpha(C_X) \;:=\; \frac{1}{1-\alpha}\log\Big(\sum_{i=1}^m \lambda_i^\alpha\Big)$$

To estimate the entropy scores from finite empirical samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$, we consider the empirical kernel covariance matrix $\widehat{C}_X$ defined as $\widehat{C}_X := \frac{1}{n}\sum_{i=1}^{n} \phi\big(\mathbf{x}_i\big)\phi\big(\mathbf{x}_i\big)^\top$. This matrix provides an empirical estimation of the population kernel covariance matrix C_X.

It can be seen that the $m \times m$ empirical matrix $\widehat{C}_X$ and normalized kernel matrix $\frac{1}{n}K = \frac{1}{n}\big[k(\mathbf{x}_i, \mathbf{x_j})\big]_{n\times n}$ share the same non-zero eigenvalues. Therefore, to compute the matrix-based entropy of the empirical covariance matrix $\widehat{C}_X$, one can equivalently compute the entropy of the eigenvalues of the kernel similarity matrix K. This approach results in the definition of the VENDI and RKE diversity scores: [1] defines the family of VENDI scores as

$$\mathrm{VENDI}_\alpha(\mathbf{x}_1, \ldots, \mathbf{x}_n) := \exp\Big(H_\alpha\big(\frac{1}{n}K\big)\Big) = \Big(\sum_{i=1}^{n} \lambda_i^\alpha\Big)^{\frac{1}{1-\alpha}},$$

where $\lambda_1, \ldots, \lambda_n$ denote the eigenvalues of the kernel matrix $\frac{1}{n}K$. Also, [2] proposes the RKE score, which is the special order-2 Renyi entropy, $\mathrm{RKE}(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \exp(H_2(\frac{1}{n}K))$. To compute RKE without computing the eigenvalues, [2] points out the RKE score reduces to the Frobenius norm $\|\cdot\|_F$ of the kernel matrix as follows:

$$\mathrm{RKE}(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \Big\|\frac{1}{n}K\Big\|_F^{-2} = \Big(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k\big(\mathbf{x}_i, \mathbf{x}_j\big)^2\Big)^{-1}$$

3.2 Shift-Invariant Kernels and Random Fourier Features

A kernel function k is called shift-invariant, if there exists a function $\kappa : \mathbb{R}^d \to \mathbb{R}$ such that $k(\mathbf{x}, \mathbf{x}') = \kappa(\mathbf{x} - \mathbf{x}')$ for every $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$. Bochner's theorem proves that a function $\kappa : \mathbb{R}^d \to \mathbb{R}$ will lead to a shift-invariant kernel similarity score $\kappa(\mathbf{x} - \mathbf{x}')$ between $\mathbf{x}, \mathbf{x}'$ *if and only if* its Fourier transform $\widehat{\kappa} : \mathbb{R}^d \to \mathbb{R}$ is non-negative everywhere (i.e, $\widehat{\kappa}(\boldsymbol{\omega}) \geq 0$ for every $\boldsymbol{\omega}$). Note that the Fourier transform $\widehat{\kappa}$ is defined as

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

Specifically, Bochner's theorem shows the Fourier transform $\widehat{\kappa}$ of a normalized shift-invariant kernel $k(\mathbf{x}, \mathbf{x}') = \kappa(\mathbf{x}-\mathbf{x}')$, where $\kappa(0) = 1$, will be a probability density function (PDF). The framework of random Fourier features (RFFs) [16] utilizes independent samples drawn from PDF $\widehat{\kappa}$ to approximate the kernel function. Here, given independent samples $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_r \sim \widehat{\kappa}$, we form the following proxy feature map $\widetilde{\phi}_r : \mathbb{R}^d \to \mathbb{R}^{2r}$

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

As demonstrated in [16, 42], the $2r$-dimensional proxy map $\widetilde{\phi}_r$ can approximate the kernel function as $k(\mathbf{x}, \mathbf{x}') = \mathbb{E}_{\boldsymbol{\omega}\sim\widehat{\kappa}}\Big[\cos(\boldsymbol{\omega}^\top \mathbf{x})\cos(\boldsymbol{\omega}^\top \mathbf{x}') + \sin(\boldsymbol{\omega}^\top \mathbf{x})\sin(\boldsymbol{\omega}^\top \mathbf{x}')\Big] \approx \widetilde{\phi}_r(\mathbf{x})^\top \widetilde{\phi}_r(\mathbf{x}')$.

4 Computational Complexity of VENDI & RKE Scores

As discussed, computing RKE and general VENDI_α scores requires computing the order-α entropy of kernel matrix $\frac{1}{n}K$. Using the standard definition of α-norm $\|\mathbf{v}\|_\alpha = \big(\sum_{i=1}^{n} |v_i|^\alpha\big)^{1/\alpha}$, we observe that the computation of VENDI_α score is equivalent to computing the α-norm $\|\boldsymbol{\lambda}\|_\alpha$ of the n-dimensional eigenvalue vector $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_n]$ where $\lambda_1 \geq \cdots \geq \lambda_n$ are the sorted eigenvalues of the normalized kernel matrix $\frac{1}{n}K$.

In the following theorem, we prove that except order $\alpha = 2$, which is the RKE score, computing any other VENDI_α score is at least as expensive as computing the product of two $n \times n$ matrices. Therefore, the theorem suggests that the computational complexity of every member of the VENDI family is lower-bounded by $\Omega(n^{2.372})$ which is the least known cost of multiplying $n \times n$ matrices.

In the theorem, we suppose $\mathcal{B}$ is any fixed set of "basis" functions. A circuit $\mathcal{C}$ is a directed acyclic graph each of whose internal nodes is labeled by a *gate* coming from a set $\mathcal{B}$. A subset of gates are

https://doi.org/10.52202/079017-3841

designated as outputs of $\mathcal{C}$. A circuit with n source nodes and m outputs computes a function from $\mathbb{R}^n$ to $\mathbb{R}^m$ by evaluating the gate at each internal gate in topological order. The size of a circuit is the number of gates. Also, $\nabla\mathcal{B}$ is the basis consisting of the gradients of all functions in $\mathcal{B}$. We will provide the proof of the theorems in the Appendix.

Theorem 1. *If* $\mathrm{VENDI}_\alpha(K)$ *for* $\alpha \neq 2$ *is computable by a circuit* $\mathcal{C}$ *of size* $s(n)$ *over basis* $\mathcal{B}$*, then* $n \times n$ *matrices can be multiplied by a circuit* $\mathcal{C}$ *of size* $O(s(n))$ *over basis* $\mathcal{B} \cup \nabla\mathcal{B} \cup \{+, \times\}$.

Remark 1. *The smallest known circuits for multiplying* $n \times n$ *matrices have size* $\Theta(n^\omega)$*, where* $\omega \approx 2.372$*. Despite tremendous research efforts only minor improvements have been obtained in recent years. There is evidence that* ω *is bounded away from 2 for certain classes of circuits [43, 44]. In contrast,* $\mathcal{S}_2$ *is computable in quadratic time* $\Theta(n^2)$ *in the basis* $B = \{\times, +, \log\}$.

The above discussion indicates that except the $\mathrm{RKE}(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, i.e. order-2 Renyi entropy, whose computational complexity is quadratically growing with sample size $\Theta(n^2)$, the other members of the VENDI family VENDI_α would have a super-quadratic complexity on the order of $\mathcal{O}(n^{2.372})$. In practice, the computation of VENDI_α scores is performed by the eigendecomposition of the $n \times n$ kernel matrix that requires $O(n^3)$ computations for precise computation and $O(n^2 M)$ computations using a randomized projection onto an M-dimensional space [15, 1].

5 A Scalable Fourier-based Method for Computing Kernel Entropy Scores

As we showed earlier, the complexity of computing RKE and VENDI scores are at least quadratically growing with the sample size n. The super-linear growth of the scores' complexity with sample size n can hinder their application to large-scale datasets and generative models with potentially hundreds of sample types. In such cases, a proper entropy estimation should be performed over potentially hundreds of thousands of data, where the quadratic complexity of the scores would be a significant barrier toward their accurate estimation.

Here, we consider a shift-invariant kernel matrix $k(\mathbf{x}, \mathbf{x}') = \kappa(\mathbf{x} - \mathbf{x}')$ where $\kappa(\mathbf{0}) = 1$ and propose applying the random Fourier features (RFF) framework [16] to perform an efficient approximation of the RKE and VENDI scores. To do this, we utilize the Fourier transform $\widehat{\kappa}$ that, according to Bochner's theorem, is a valid PDF, and we independently generate $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_r \overset{\text{iid}}{\sim} \widehat{\kappa}$. Note that in the case of the Gaussian kernel $k_{\text{Gaussian}(\sigma^2)}$, the corresponding PDF will be an isotropic Gaussian $\mathcal{N}(\mathbf{0}, \frac{1}{\sigma^2} I_d)$ with zero mean and covariance matrix $\frac{1}{\sigma^2} I_d$. Then, we consider the RFF proxy feature map $\widetilde{\phi}_r : \mathbb{R}^d \to \mathbb{R}^{2r}$ as defined in (1) and define the proxy kernel covariance matrix $\widetilde{C}_{X,r} \in \mathbb{R}^{2r \times 2r}$:

$$\widetilde{C}_{X,r} \;=\; \frac{1}{n} \sum_{i=1}^{n} \widetilde{\phi}_r\big(\mathbf{x}_i\big)\, \widetilde{\phi}_r\big(\mathbf{x}_i\big)^\top \tag{2}$$

Note that the $2r \times 2r$ matrix $\widehat{C}_{X,r}$ has the same non-zero eigenvalues as the $n \times n$ RFF proxy kernel matrix $\frac{1}{n}\widetilde{K}_r$, and therefore can be utilized to approximate the eigenvalues of the original $n \times n$ kernel matrix $\frac{1}{n}K$. Therefore, we propose the *Fourier-based Kernel Entropy Approximation (FKEA)* method to approximate the RKE and VENDI_α scores as follows:

$$\text{FKEA-RKE}\big(\mathbf{x}_1, \ldots, \mathbf{x}_n\big) \;=\; \exp\big(H_2(\widetilde{C}_{X,r})\big) \;=\; \big\|\widetilde{C}_{X,r}\big\|_F^{-2}, \tag{3}$$

$$\text{FKEA-VENDI}_\alpha\big(\mathbf{x}_1, \ldots, \mathbf{x}_n\big) \;=\; \exp\big(H_\alpha(\widehat{C}_{X,r})\big) \;=\; \Big(\sum_{i=1}^{2r} \widetilde{\lambda}_{r,i}^{\alpha}\Big)^{\frac{1}{1-\alpha}} \tag{4}$$

Note that in the above, $\widetilde{\lambda}_{r,i}^{\alpha}$ denotes the ith eigenvalue of the $2r \times 2r$ matrix $\widehat{C}_{X,r}$. We remark that the computation of both FKEA-RKE and FKEA-VENDI_α can be done by a stochastic algorithm which computes the proxy covariance matrix (2) by summing the sample-based $2r \times 2r$ matrix terms, and then computing the resulting matrix's Frobenius norm for RKE score or all the $2r$ matrix's eigenvalues for a general VENDI_α with $\alpha \neq 2$. Algorithm 1 presents the steps of the FKEA method where the computation needed for the proxy kernel covariance matrix is $O(n)$ and grows only linearly with sample size n.

Therefore, to show the FKEA method's scalability, we need to bound the required RFF size $2r$ for an accurate approximation of the original $n \times n$ kernel matrix. The following theorem proves that the needed feature size will be $\mathcal{O}\big(\frac{\log n}{\epsilon^2}\big)$ for an ϵ-accurate approximations of the matrix's eigenspace.

Algorithm 1 FKEA Algorithm for Computing VENDI and RKE reference-free scores

Input: n datapoints $\mathbf{x} = \{x_1, \ldots, x_n\}$, kernel bandwidth σ^2, RFF dimension r
Draw i.i.d. samples $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_r \sim \hat{\kappa}$ ▷ For Gaussian Kernel $\hat{\kappa} \sim \mathcal{N}(0, \frac{1}{\sigma^2} I_d)$
Initialize the covariance matrix $\widetilde{C} \leftarrow \mathbf{0}$
Compute the covariance matrix:
for $i = 1$ to n **do**
    Compute the RFF feature for x_i:

$$\widetilde{\phi}_r(x_i) = \frac{1}{\sqrt{r}} \left[\cos(\boldsymbol{\omega}_1^\top x_i), \quad \sin(\boldsymbol{\omega}_1^\top x_i), \quad \cdots, \quad \cos(\boldsymbol{\omega}_r^\top x_i), \quad \sin(\boldsymbol{\omega}_r^\top x_i)\right]^\top$$

    Update $\widetilde{C}$:

$$\widetilde{C} \leftarrow \widetilde{C} + \frac{1}{n} \widetilde{\phi}_r(x_i)\, \widetilde{\phi}_r(x_i)^\top$$

end for
Perform eigendecomposition on the covariance matrix:

$$\{\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_{2r}\} \leftarrow \text{Eigendecomposition}(\widetilde{C})$$

Compute VENDI and RKE metrics using the eigenvalues $\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_{2r}$

Theorem 2. *Consider a shift-invariant kernel $k(\mathbf{x}, \mathbf{x}') = \kappa(\mathbf{x} - \mathbf{x}')$ where $\kappa(\mathbf{0}) = 1$. Suppose $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_r \sim \widehat{\kappa}$ are independently drawn from PDF $\widehat{\kappa}$. Let $\lambda_1 \geq \ldots \geq \lambda_n$ be the sorted eigenvalues of the normalized kernel matrix $\frac{1}{n}K = \frac{1}{n}\left[k(\mathbf{x}_i, \mathbf{x}_j)\right]_{n\times n}$. Also, consider the eigenvalues of $\widetilde{\lambda}_1 \geq \ldots \geq \widetilde{\lambda}_{2r}$ of random matrix $\widetilde{C}_{X,r}$ with their corresponding eigenvectors $\widetilde{\mathbf{v}}_1, \ldots, \widetilde{\mathbf{v}}_{2r}$. Let $\widetilde{\lambda}_j = 0$ for every $j > 2r$. Then, for every $\delta > 0$, the following holds with probability at least $1 - \delta$:*

$$\sqrt{\sum_{i=1}^{n} \left(\widetilde{\lambda}_i - \lambda_i\right)^2} \leq \sqrt{\frac{8\log(n/2\delta)}{r}} \qquad \text{and} \qquad \sqrt{\sum_{i=1}^{n} \left\| \frac{1}{n} K \widehat{\mathbf{v}}_i - \lambda_i \widehat{\mathbf{v}}_i \right\|_2^2} \leq \sqrt{\frac{32\log(n/2\delta)}{r}},$$

where $\widehat{\mathbf{v}}_i := \sum_{j=1}^{r} \sin\left(\widetilde{\mathbf{v}}_{2j}^\top \mathbf{x}_i\right) \widetilde{\mathbf{v}}_{2j} + \cos\left(\widetilde{\mathbf{v}}_{2j-1}^\top \mathbf{x}_i\right) \widetilde{\mathbf{v}}_{2j-1}$ is the ith proxy eigenvector for $\frac{1}{n}K$.

Corollary 1. *In the setting of Theorem 2, the following approximation guarantees hold for* RKE *and* VENDI_α *scores*

- *For every VENDI_α with $\alpha \geq 2$, including RKE for $\alpha = 2$, the following dimension-independent bound holds with probability at least $1 - \delta$:*

$$\left| \text{FKEA-VENDI}_\alpha\left(\mathbf{x}_1, \ldots, \mathbf{x}_n\right)^{\frac{1-\alpha}{\alpha}} - \mathrm{VENDI}_\alpha\left(\mathbf{x}_1, \ldots, \mathbf{x}_n\right)^{\frac{1-\alpha}{\alpha}} \right| \leq \sqrt{\frac{8\log(n/2\delta)}{r}}$$

- *For every VENDI_α with $1 \leq \alpha < 2$, assuming a finite dimension for the kernel feature map $\dim(\phi) = m$, the following bound holds with probability at least $1 - \delta$:*

$$\left| \text{FKEA-VENDI}_\alpha\left(\mathbf{x}_1, \ldots, \mathbf{x}_n\right)^{\frac{1-\alpha}{\alpha}} - \mathrm{VENDI}_\alpha\left(\mathbf{x}_1, \ldots, \mathbf{x}_n\right)^{\frac{1-\alpha}{\alpha}} \right| \leq m^{\frac{1}{\alpha}-\frac{1}{2}} \sqrt{\frac{8\log(n/2\delta)}{r}}$$

Remark 2. *According to the theoretical results in [45], the top-t eigenvectors of kernel covariance matrix C_X will correspond to the mean of the modes of a mixture distribution with t well-separable modes. Theorem 2 shows for every $1 \leq i \leq 2r$, FKEA provides the proxy score function $\widetilde{u}_i : \mathbb{R}^d \to \mathbb{R}$ that assigns a likelihood score for an input $\mathbf{x}$ to belong to the ith identified mode:*

$$\widetilde{u}_i(\mathbf{x}) = \sum_{j=1}^{r} \sin\left(\widetilde{\mathbf{v}}_{2j}^\top \mathbf{x}\right) \widetilde{\mathbf{v}}_{2j,i} + \cos\left(\widetilde{\mathbf{v}}_{2j-1}^\top \mathbf{x}\right) \widetilde{\mathbf{v}}_{2j-1,i} \tag{5}$$

Therefore, one can compute the above FKEA-based score for each of the $2r$ eigenvectors over a sample set, and use the samples with the highest scores according to every $\widetilde{u}_i$ to characterize the i sample cluster captured by the FKEA method.

https://doi.org/10.52202/079017-3841

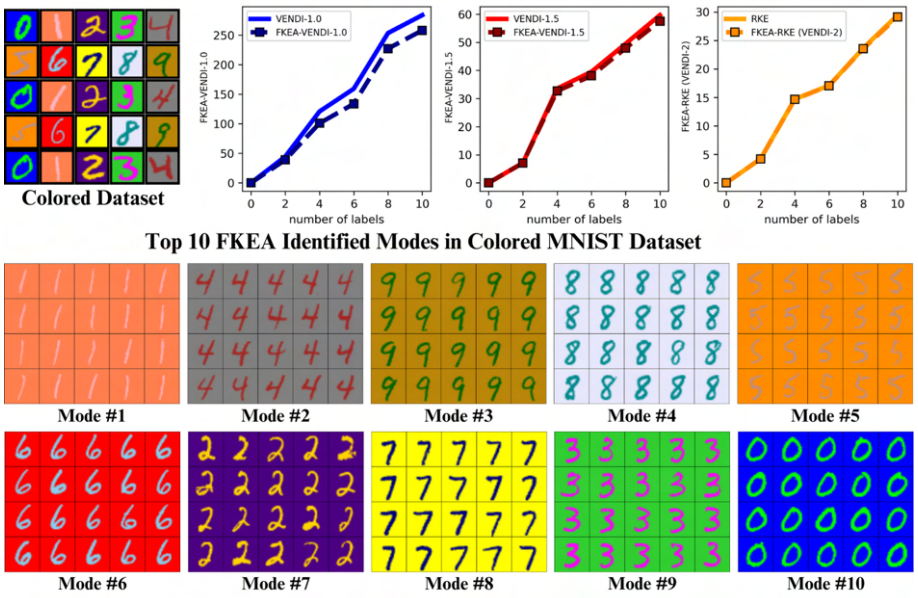


Figure 2: RFF-based identified clusters used in FKEA Evaluation in single-colored MNIST [46] dataset with *pixel* embedding, Fourier feature dimension $2r = 4000$ and bandwidth $\sigma = 7$. The graphs indicate increase in FKEA RKE/VENDI diversity metrics with increasing number of labels.

Table 1: Time complexity for FKEA and non-FKEA based metrics (RKE and VENDI) on ImageNet dataset with *DinoV2* embedding. Computation of VENDI and RKE on 40k+ samples are omitted due to memory overflow during metric computation.

	Time (sec)													
	$2r = 8000$							$2r = 16000$						
Metric	n=10k	n=20k	n=30k	n=40k	n=50k	n=100k	n=250k	n=10k	n=20k	n=30k	n=40k	n=50k	n=100k	n=250k
FKEA-RKE	7	16	25	34	43	87	238	37	78	120	162	203	433	1138
FKEA-VENDI	11	19	27	37	45	104	267	48	89	130	173	213	459	1236
RKE	217	1324	4007	-	-	-	-	218	1330	4021	-	-	-	-
VENDI	286	1774	5488	-	-	-	-	287	1780	5502	-	-	-	-

6 Numerical Results

We evaluated the FKEA method on several image, text, and video datasets to assess its performance in quantifying diversity in different domains. In the experiments, we computed the empirical covariance matrix of $2r$-dimensional Fourier features with a Gaussian kernel with bandwidth parameter σ tuned for each dataset, and then applied FKEA approximation for the VENDI_1, $\mathrm{VENDI}_{1.5}$, and the RKE (same as VENDI_2) scores. An algorithm to compute these scores is presented in Algorithm 1. Experiments were conducted on RTX3090 GPUs. We interpreted the modes identified by FKEA entropy-based diversity evaluation using the eigenvectors of the proxy covariance matrix as discussed in Remark 2. For each eigenvector, we presented the training data with maximum eigenfunction values corresponding to the eigenvector.

Time Complexity of FKEA metrics. To highlight the computational advantages of transitioning to FKEA, Table 1 presents a comparison of the metric computations for VENDI and RKE on the ImageNet dataset, with sample sizes ranging from 10k to 250k. Our results show that VENDI and RKE become computationally intractable due to memory overflow. In contrast, the FKEA method efficiently scales up to $n = 250k$ samples, maintaining optimal computational time.

Experimental Results on Image Data. To investigate the FKEA method's diversity evaluation in settings where we know the ground-truth clusters and their quantity, we simulated an experiment on the colored MNIST [46] data with the images of 10 colored digits as shown in Figure 2. We evaluated the FKEA approximations of the diversity scores while including samples from t digits for $t \in \{1, \ldots, 10\}$. The plots in Figure 2 indicate the increasing trend of the scores and FKEA's tight

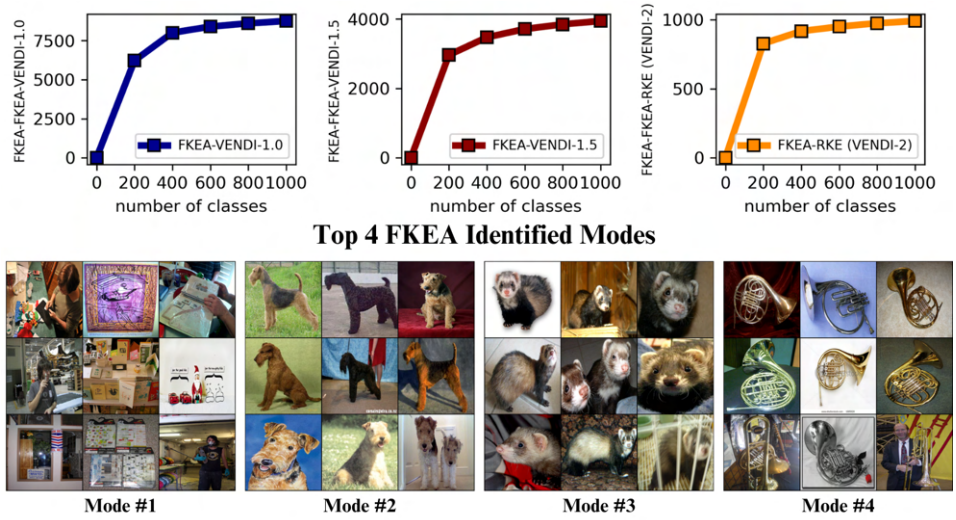


Figure 3: RFF-based identified clusters used in FKEA Evaluation in ImageNet dataset with *DinoV2* embedding, Fourier feature dimension $2r = 16k$ and Gaussian Kernel bandwidth $\sigma = 25$. The graphs indicate increase in FKEA diversity metrics with increasing number of labels per 50k samples.

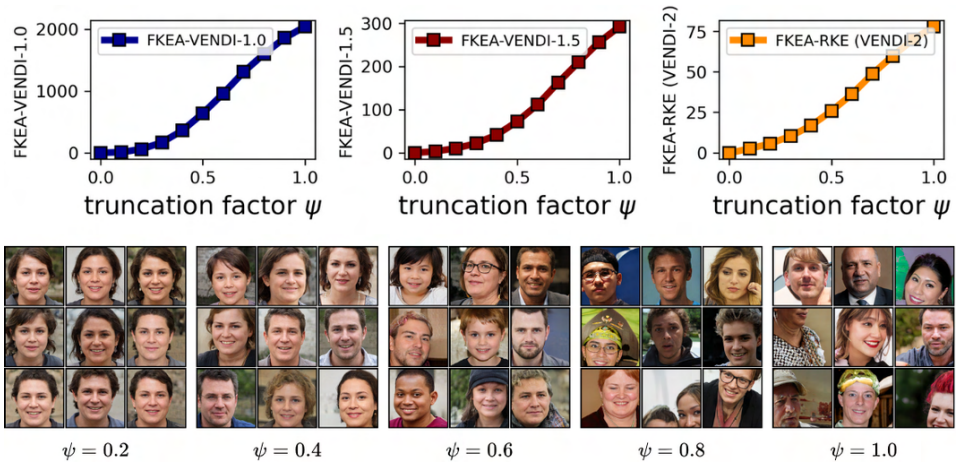


Figure 4: FKEA metrics behavior under different truncation factor ψ of StyleGAN3 [47] generated FFHQ samples.

approximations of the scores. Also, we show the top-20 training samples with the highest scores according to the top-10 FKEA eigenvectors, showing the method captured the ground-truth clusters.

We conducted an experiment on the ImageNet data to monitor the scores' behavior evaluated for 50k samples from an increasing number of ImageNet labels. Figure 3 shows the increasing trend of the scores as well as the top-9 samples representing the top-4 identified clusters used for the entropy calculation. Also, Figure 4 presents the FKEA approximated entropy scores with different truncation factors in StyleGAN3 [47] on 30k generated data for each truncation factor, showing the increasing diversity scores with the truncation factor. We defer discussing the results on AFHQ [48], MS-COCO [49], F-MNIST [50] datasets to the Appendix.

Table 2: Top 5 synthetic countries dataset modes with *text-embedding-3-large* embedding, Fourier features dim $2r = 8000$ and Gaussian Kernel bandwidth $\sigma = 0.9$. The table summarises the mentions of each country in the top 100 paragraphs identified for the eigenvectors corresponding to each mode.

Mode #1		**Mode #2**		**Mode #3**		**Mode #4**		**Mode #5**		**Mode #6**	
Burkina Faso	34%	Argentina	77%	Azerbaijan	100%	Cambodia	94%	Belarus	100%	Bolivia	97%
Benin	23%	Chile	23%			Afghanistan	6%			Ecuador	3%
Chad	22%										
Burundi	13%										
Cameroon	8%										

https://doi.org/10.52202/079017-3841

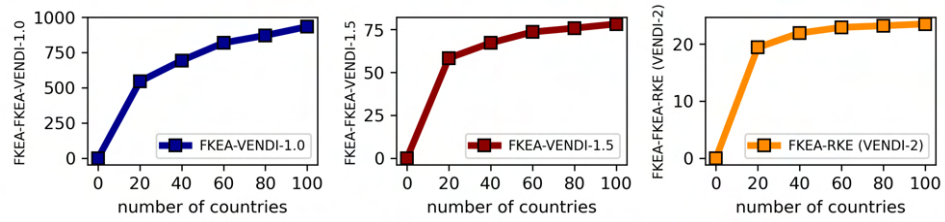
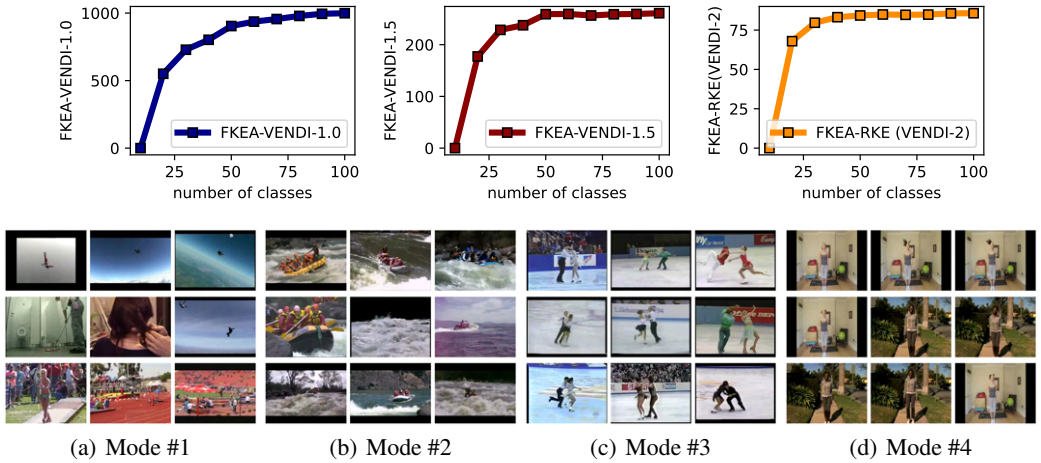


Figure 5: FKEA diversity metrics with the increasing number of countries in the synthetic dataset.



(a) Mode #1 (b) Mode #2 (c) Mode #3 (d) Mode #4

Figure 6: RFF-based identified clusters used in FKEA evaluation in UCF101 dataset with *I3D* embedding. The graphs indicate an increase in FKEA diversity metrics with more classes.

Experimental Results on Text and Video Data. To perform experiments on the text data with known clustering ground-truth, we generated 500,000 paragraphs using GPT-3.5 [51] about 100 randomly selected countries (5k samples per country). In the experiments, the text embedding used was *text-embedding-3-large* [52, 53, 51]. We evaluated the diversity scores over data subsets of size 50k with different numbers of mentioned countries. Figure 5 shows the growing trend of the diversity scores when including more countries. The figure also shows the countries mentioned in the top-6 modes provided by FKEA-based principal eigenvectors, which shows the RFF-based clustering of countries correlates with their continent and geographical location. We discuss the numerical results on non-synthetic text datasets, Wikipedia, CNN/Dailymail [54][55], CMU Movie Corpus [56], in the Appendix.

For video data experiments, we considered two standard video datasets, UCF101 [57] and Kinetics-400 [58]. Following the video evaluation literature [59, 60], we used the I3D pre-trained model [61] as embedding, which maps a video sample to a 1024-dimensional vector. As shown in Figure 6, increasing the number of video classes of test samples led to an increase in the FKEA approximated diversity metrics. Also, while the samples identified for the first identified cluster look broad, the next modes seemed more specific. We discuss the results of the Kinetics dataset in the Appendix.

7 Conclusion

In this work, we proposed the Fourier-based FKEA method to efficiently approximate the kernel-based entropy scores VENDI_α and RKE scores. The application of FKEA results in a scalable reference-free evaluation of generative models, which could be utilized in applications where no reference data is available for evaluation. A future direction to our work is to study the sample complexity of the matrix-based entropy scores and the FKEA's approximation under high-dimensional kernel feature maps, e.g. the Gaussian kernel. Also, analyzing the role of feature embedding in the method's application to text and video data would be interesting for future exploration.

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 a CUHK Direct Research Grant with CUHK Project No. 4055164. Xuenan Cao's work is supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China, Project 14602223. Andrej Bogdanov's work is supported by an NSERC Discovery Grant.

References

[1] Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. In *Transactions on Machine Learning Research*, 2023.

[2] Mohammad Jalali, Cheuk Ting Li, and Farzan Farnia. An information-theoretic evaluation of generative models in learning multi-modal distributions. In *Advances in Neural Information Processing Systems*, volume 36, pages 9931–9943, 2023.

[3] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2013.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

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

[6] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. In *International Journal of Computer Vision (IJCV)*, number 3, pages 211–252, 2015.

[7] 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. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2018.

[8] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018.

[9] Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

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

[11] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *ICML'20*, pages 7176–7185. JMLR.org, 2020.

[12] 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*, volume 29. Curran Associates, Inc., 2016.

[13] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024.

[14] Amey Pasarkar and Adji Bousso Dieng. Cousins of the vendi score: A family of similarity-based diversity metrics for science and machine learning. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2024.

 https://doi.org/10.52202/079017-3841

[15] Yuxin Dong, Tieliang Gong, Shujian Yu, and Chen Li. Optimal randomized approximations for matrix-based rényi's entropy. *IEEE Transactions on Information Theory*, 2023.

[16] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

[17] Ahmed M. Alaa, Boris van Breugel, Evgeny Saveliev, and Mihaela van der Schaar. How Faithful is your Synthetic Data? Sample-level Metrics for Evaluating and Auditing Generative Models, July 2022. arXiv:2102.08921 [cs, stat].

[18] Marco Jiralerspong, Joey Bose, Ian Gemp, Chongli Qin, Yoram Bachrach, and Gauthier Gidel. Feature likelihood score: Evaluating the generalization of generative models using samples. *Advances in Neural Information Processing Systems*, 36, 2024.

[19] 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. *arXiv preprint arXiv:2206.08549*, 2022.

[20] George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 3732–3784. Curran Associates, Inc., 2023.

[21] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. In *Transactions on Machine Learning Research*, 2023.

[22] Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The Role of ImageNet Classes in Fréchet Inception Distance. September 2022.

[23] 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. In *International Conference on Machine Learning*, pages 8748–8763. arXiv, February 2021. arXiv:2103.00020 [cs].

[24] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In *Advances in Neural Information Processing Systems*, volume 33, pages 9912–9924. Curran Associates, Inc., 2020.

[25] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826. ISSN: 1063-6919.

[26] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318. Association for Computational Linguistics, 2002.

[27] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 1097–1100. Association for Computing Machinery, 2018.

[28] Vishakh Padmakumar and He He. Does Writing with Language Models Reduce Content Diversity?, March 2024. arXiv:2309.05196 [cs].

[29] Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive Decoding: Open-ended Text Generation as Optimization. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada, July 2023. Association for Computational Linguistics.

[30] Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. Locally Typical Sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121, 2023. Place: Cambridge, MA Publisher: MIT Press.

[31] Chin-Yew Lin and Franz Josef Och. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, Barcelona, Spain, July 2004.

[32] Ehsan Montahaei, Danial Alihosseini, and Mahdieh Soleymani Baghshah. Jointly Measuring Diversity and Quality in Text Generation Models. arXiv, May 2019. arXiv:1904.03971 [cs, stat].

[33] Chantal Shaib, Joe Barrow, Jiuding Sun, Alexa F. Siu, Byron C. Wallace, and Ani Nenkova. Standardizing the measurement of text diversity: A tool and a comparative analysis of scores.

[34] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, 10(5):1299–1319, July 1998. Conference Name: Neural Computation.

[35] Yoshua Bengio, Pascal Vincent, Jean-François Paiement, O Delalleau, M Ouimet, and N LeRoux. Learning eigenfunctions of similarity: linking spectral clustering and kernel pca. Technical report, Technical Report 1232, Departement d'Informatique et Recherche Oprationnelle . . . , 2003.

[36] Yoshua Bengio, Pascal Vincent, Jean-François Paiement, Olivier Delalleau, Marie Ouimet, and Nicolas Le Roux. *Spectral clustering and kernel PCA are learning eigenfunctions*, volume 1239. Citeseer, 2003.

[37] 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.

[38] Mina Ghashami, Daniel J Perry, and Jeff Phillips. Streaming kernel principal component analysis. In *Artificial intelligence and statistics*, pages 1365–1374. PMLR, 2016.

[39] Enayat Ullah, Poorya Mianjy, Teodor Vanislavov Marinov, and Raman Arora. Streaming kernel pca with $o(\sqrt{n})$ random features. *Advances in Neural Information Processing Systems*, 31, 2018.

[40] Bharath K Sriperumbudur and Nicholas Sterge. Approximate kernel pca: Computational versus statistical trade-off. *The Annals of Statistics*, 50(5):2713–2736, 2022.

[41] Daniel Gedon, Antônio H Ribeiro, Niklas Wahlström, and Thomas B Schön. Invertible kernel pca with random fourier features. *IEEE Signal Processing Letters*, 30:563–567, 2023.

[42] Danica J. Sutherland and Jeff Schneider. On the Error of Random Fourier Features. In *Uncertainty in Artificial Intelligence (UAI) 2015*. arXiv, June 2015. arXiv:1506.02785 [cs, stat].

[43] Josh Alman. Limits on the universal method for matrix multiplication. *Theory of Computing*, 17(1):1–30, 2021.

[44] Matthias Christandl, Péter Vrana, and Jeroen Zuiddam. Barriers for fast matrix multiplication from irreversibility. *Theory of Computing*, 17(2):1–32, 2021.

[45] Jingwei Zhang, Cheuk Ting Li, and Farzan Farnia. An interpretable evaluation of entropy-based novelty of generative models. In *International Conference on Machine Learning (ICML 2024)*.

 https://doi.org/10.52202/079017-3841

[46] Li Deng. The MNIST database of handwritten digit images for machine learning research [best of the web]. In *IEEE Signal Processing Magazine*, volume 29, pages 141–142, 2012. Conference Name: IEEE Signal Processing Magazine.

[47] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 852–863. Curran Associates, Inc.

[48] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[49] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *European Conference on Computer Vision (ECCV)*, volume 8693, pages 740–755. Springer International Publishing, 2014. Book Title: Computer Vision – ECCV 2014 Series Title: Lecture Notes in Computer Science.

[50] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

[51] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

[52] OpenAI. text-embedding-3-large. `https://platform.openai.com/docs/models/embeddings`, 2024.

[53] OpenAI. GPT-4 Technical Report, March 2024. arXiv:2303.08774 [cs].

[54] Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701, 2015.

[55] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[56] Brendan O'Connor David Bamman and Noah A. Smith. Learning latent personas of film characters. In *ACL 2013*, 2013.

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

[58] 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.

[59] 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, Nov 2020.

[60] 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.

[61] 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.

[62] S. Linnainmaa. The representation of the cumulative rounding error of an algorithm as a taylor expansion of the local rounding errors. Master's thesis, University of Helsinki, 1970.

[63] W. Baur and V. Strassen. The complexity of partial derivatives. *Theoretical Computer Science*, 22:317 – 330, 1983.

[64] 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.

[65] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc.

[66] Andrew Brock, Jeff Donahue, and K. Simonyan. Large scale GAN training for high fidelity natural image synthesis.

[67] William Peebles and Saining Xie. Scalable diffusion models with transformers.

[68] 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.

[69] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685. IEEE Computer Society.

[70] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

[71] 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.

[72] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 Conference Proceedings*, volume abs/2201.00273, 2022.

[73] Rewon Child. Very deep VAEs generalize autoregressive models and can outperform them on images.

[74] Ceyuan Yang, Yujun Shen, Yinghao Xu, and Bolei Zhou. Data-efficient instance generation from instance discrimination. In *Advances in Neural Information Processing Systems*, volume 34, pages 9378–9390. Curran Associates, Inc.

[75] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. YAKE! keyword extraction from single documents using multiple local features. In *Information Sciences*, volume 509, pages 257–289, 2020.

 https://doi.org/10.52202/079017-3841

A Proofs

A.1 Proof of Theorem 1

The proof of Theorem 1 combines third ingredients. The first is the relation between the circuit size of a function C and of its partial derivatives $\nabla C = (\partial C/\partial x_1, \ldots, \partial C/\partial x_n)$.

Lemma 1. *The function ∇C has a circuit over basis $\nabla B \cup \{+, \times\}$ whose size is within a constant factor of the size of C.*

Lemma 1 is a feature of the backpropagation algorithm [62, 63]. This is a linear-time algorithm for constructing a circuit for ∇C given the circuit C as input. In contrast, the forward propagation algorithm allows efficient computation of a single (partial) derivative even for circuits with multivalued outputs, giving the second ingredient:

Lemma 2. *Let C be a circuit over basis B and t be an input to C. There exists a circuit that computes the derivative $\partial g/\partial t$ for every gate g of C over basis $\nabla B \cup \{+, \times\}$ whose size is within a constant factor of the size of C.*

The last ingredient is the following identity. For a scalar function f over the complex numbers and matrix X diagonalizable as $U\Lambda U^T$, we define $f(X)$ to be the function $Uf(\Lambda)U^T$ where f is applied entry-wise to the diagonal matrix Λ.

Lemma 3. *For every f that is analytic over an open domain Ω containing all sufficiently large complex numbers and every matrix X whose spectrum is contained in Ω, $\nabla \mathrm{Tr}(f(X)) = f'(X)$.*

We first illustrate the proof in the special case when $\|X\|$ is within the radius of convergence of f. Namely, $f(x)$ is represented by the absolutely convergent series $\sum f^{(k)}(0)x^k/k!$ for all $|x| \leq \rho$. Then $f(X) = \sum f^{(k)}(0)X^k/k!$ assuming $\|X\| \leq \rho$. By linearity (and using the fact that derivatives preserve radius of convergence) it is sufficient to show that

$$\nabla \mathrm{Tr} X^k = \frac{dX^k}{dX}, \tag{6}$$

which can be verified by explicit calculation: Both sides equal kX^{k-1}. This is sufficient to establish Theorem 1 for all integer $\alpha > 2$.

Proof of Lemma 3. The Cauchy integral formula for matrices yields the representation

$$f(X) = \frac{1}{2\pi i} \int_C f(z)(zI - X)^{-1} dz,$$

for any closed curve C whose interior contains the spectrum of X. As $(zI - X)^{-1}$ is continuous along C, we can write

$$\nabla \mathrm{Tr} f(X) = \frac{1}{2\pi i} \int_C f(z) \nabla \mathrm{Tr}(zI - X)^{-1} dz. \tag{7}$$

Choosing C to be a circle of radius ρ larger than the spectral norm of X, for all z of magnitude ρ we have the identity

$$(zI - X)^{-1} = z^{-1}(I - z^{-1}X)^{-1} = z^{-1} \sum_{k=0}^{\infty} z^{-k} X^k$$

As the series $\sum z^{-k} \nabla \mathrm{Tr} X^k = \sum k z^{-k} X^{k-1}$ converges absolutely in spectral norm, using (6) we obtain the identity $\nabla \mathrm{Tr}(zI - X)^{-1} = d(zI - X)^{-1}/dX$, namely the lemma holds for the function $f(X) = (zI - X)^{-1}$. Plugging into (7) and exchanging the order of integration and derivation proves the lemma. □

Proof of Theorem 1. Assume $\mathrm{Tr}\rho^\alpha$ (resp., $-\mathrm{Tr}\rho \log \rho$) has circuit size $s(d)$. By Lemma 1 and Lemma 3, $\nabla \mathrm{Tr}\rho^\alpha = \alpha \rho^{\alpha-1}$ (resp., $-\nabla \mathrm{Tr}\rho \log \rho = \log \rho - 1/\ln 2$) has circuit size $O(s(d))$. For every symmetric matrix X and sufficiently small t, the matrix $\rho = I + tX$ is positive semi-definite. By Lemma 2 the $\mathbb{R}^{d^2}$-valued function $\partial^2 \rho/\partial t^2$ has circuit size $O(s^d)$. The value of this function at

$t = 0$ is $\alpha(\alpha-1)(\alpha-2)X^2$ (resp., X^2), namely the square of the input matrix X up to constant. Finally, computing the product AB reduces to squaring the symmetric matrix

$$\begin{pmatrix} & A^T & B \\ A & & \\ B^T & & \end{pmatrix}. \qquad \square$$

A.2 Proof of Theorem 2

Assuming that the shift-invariant kernel $k(\mathbf{x}, \mathbf{x}') = \kappa(\mathbf{x} - \mathbf{x}')$ is normalized (i.e. $\kappa(\mathbf{0}) = 1$), then the Fourier transform $\widehat{\kappa}$ is a valid PDF according to Bochner's theorem and also an even function because κ takes real values. Then, we have

$$\begin{aligned} k\big(\mathbf{x}, \mathbf{x}'\big) \;&=\; \kappa_\sigma\big(\mathbf{x} - \mathbf{x}'\big) \\ &\overset{(a)}{=} \int \widehat{\kappa_\sigma}(\boldsymbol{\omega}) \exp\big(i\boldsymbol{\omega}^\top(\mathbf{x} - \mathbf{x}')\big) \mathrm{d}\boldsymbol{\omega} \\ &\overset{(b)}{=} \int \widehat{\kappa_\sigma}(\boldsymbol{\omega}) \cos\big(\boldsymbol{\omega}^\top(\mathbf{x} - \mathbf{x}')\big) \mathrm{d}\boldsymbol{\omega} \\ &= \mathbb{E}_{\boldsymbol{\omega}\sim\widehat{\kappa}}\Big[\cos\big(\boldsymbol{\omega}^\top(\mathbf{x} - \mathbf{x}')\big)\Big] \\ &= \mathbb{E}_{\boldsymbol{\omega}\sim\widehat{\kappa}}\Big[\cos\big(\boldsymbol{\omega}^\top\mathbf{x}\big)\cos\big(\boldsymbol{\omega}^\top\mathbf{x}'\big) + \sin\big(\boldsymbol{\omega}^\top\mathbf{x}\big)\sin\big(\boldsymbol{\omega}^\top\mathbf{x}'\big)\Big] \end{aligned}$$

Here, (a) comes from the synthesis property of the Fourier transform. (b) holds since $\widehat{\kappa_\sigma}$ is an even function, resulting in a zero imaginary term in the Fourier synthesis.

Therefore, since $\big|\cos\big(\boldsymbol{\omega}^\top\mathbf{y}\big)\big| \leq 1$ for all $\boldsymbol{\omega}$ and $\mathbf{y}$, one can apply Hoeffding's inequality to show for independently drawn $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_r \overset{\text{iid}}{\sim} \widehat{\kappa}$ the following probably correct bound holds:

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

Therefore, as the identity $\cos(a-b) = \cos(a)\cos(b) + \sin(a)\sin(b)$ reveals $\frac{1}{r}\sum_{i=1}^{r}\cos\big(\boldsymbol{\omega}_i^\top(\mathbf{x} - \mathbf{x}')\big) = \widetilde{\phi}_r(\mathbf{x})^\top\widetilde{\phi}_r(\mathbf{x}')$, the above bound can be rewritten as

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

Also, $\widetilde{k}_r(\mathbf{x}, \mathbf{x}') = \widetilde{\phi}_r(\mathbf{x})^\top\widetilde{\phi}_r(\mathbf{x}')$ is by definition a normalized kernel, implying that

$$\forall \mathbf{x} \in \mathbb{R}^d : \quad \widetilde{\phi}_r(\mathbf{x})^\top\widetilde{\phi}_r(\mathbf{x}) - k(\mathbf{x}, \mathbf{x}) \;=\; 0.$$

As a result, one can apply the union bound to combine the above inequalities and show for every sample set $\mathbf{x}_1, \ldots, \mathbf{x}_n$:

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

Considering the normalized kernel matrix $\frac{1}{n}K = \frac{1}{n}\big[k(\mathbf{x}_i, \mathbf{x}_j)\big]_{1\leq i,j\leq n}$ and the proxy normalized kernel matrix $\frac{1}{n}\widetilde{K} = \frac{1}{n}\big[\widetilde{\phi}_r(\mathbf{x}_i)^\top\widetilde{\phi}_r(\mathbf{x}_j)\big]_{1\leq i,j\leq n}$, the above inequality implies that

$$\begin{aligned} &\mathbb{P}\Big(\big\|\frac{1}{n}\widetilde{K} - \frac{1}{n}K\big\|_F^2 \;\geq\; n^2\frac{\epsilon^2}{n^2}\Big) \;\leq\; \binom{n}{2}\exp\Big(-\frac{r\epsilon^2}{2}\Big). \\ \implies \quad &\mathbb{P}\Big(\big\|\frac{1}{n}\widetilde{K} - \frac{1}{n}K\big\|_F \;\geq\; \epsilon\Big) \;<\; \frac{n^2}{2}\exp\Big(-\frac{r\epsilon^2}{2}\Big). \end{aligned} \tag{8}$$

Leveraging the eigenvalue-perturbation bound in [64], we can translate the above bound to the following for the sorted eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ of $\frac{1}{n}K$ and the sorted eigenvalues $\widetilde{\lambda}_1 \geq \cdots \geq \widetilde{\lambda}_n$ of $\frac{1}{n}\widetilde{K}$

$$\sqrt{\sum_{i=1}^{n}(\widetilde{\lambda}_i - \lambda_i)^2} \;\leq\; \big\|\frac{1}{n}\widetilde{K} - \frac{1}{n}K\big\|_F$$

 https://doi.org/10.52202/079017-3841

which shows

$$\mathbb{P}\Big(\sqrt{\sum_{i=1}^{r'}(\widetilde{\lambda}_i-\lambda_i)^2}\ \geq\ \epsilon\Big)\ \leq\ \frac{n^2}{2}\exp\Big(-\frac{r\epsilon^2}{2}\Big) \tag{9}$$

Defining $\delta = \frac{n^2}{2}\exp\Big(-\frac{r\epsilon^2}{2}\Big)$, i.e., $\epsilon = \sqrt{\frac{8\log\big(n/2\delta\big)}{r}}$, leads to

$$\mathbb{P}\Big(\sqrt{\sum_{i=1}^{n}(\widetilde{\lambda}_i-\lambda_i)^2}\ \leq\ \epsilon\Big)\ \geq\ 1-\delta. \tag{10}$$

Noting that the normalized proxy kernel matrix $\widetilde{K}$ and the proxy kernel covariance matrix $\widetilde{C}_X$ share identical non-zero eigenvalues together with the above bound finish the proof of Theorem 2's first part.

Concerning Theorem 2's approximation guarantee for the eigenvectors, note that for each eigenvectors $\widehat{\mathbf{v}}_i$ of the proxy kernel matrix $\frac{1}{n}\widetilde{K}$, the following holds:

$$\begin{aligned}\Big\|\frac{1}{n}K\widehat{\mathbf{v}}_i-\lambda_i\widehat{\mathbf{v}}_i\Big\|_2\ &\leq\ \Big\|\frac{1}{n}K\widehat{\mathbf{v}}_i-\widetilde{\lambda}_i\widehat{\mathbf{v}}_i\Big\|_2+\Big\|\widetilde{\lambda}_i\widehat{\mathbf{v}}_i-\lambda_i\widehat{\mathbf{v}}_i\Big\|_2\\ &=\ \Big\|\big(\frac{1}{n}K-\frac{1}{n}\widetilde{K}\big)\widehat{\mathbf{v}}_i\Big\|_2+\big|\tilde{\lambda}_i-\lambda_i\big|\end{aligned}$$

Therefore, applying Young's inequality shows that

$$\begin{aligned}\Big\|\frac{1}{n}K\widehat{\mathbf{v}}_i-\lambda_i\widehat{\mathbf{v}}_i\Big\|_2^2\ &\leq\ 2\Big\|\big(\frac{1}{n}K-\frac{1}{n}\widetilde{K}\big)\widehat{\mathbf{v}}_i\Big\|_2^2+2\big(\tilde{\lambda}_i-\lambda_i\big)^2\\ &=\ 2\mathrm{Tr}\Big(\widehat{\mathbf{v}}_i^\top\big(\frac{1}{n}K-\frac{1}{n}\widetilde{K}\big)^2\widehat{\mathbf{v}}_i\Big)+2\big(\tilde{\lambda}_i-\lambda_i\big)^2\\ &=\ 2\mathrm{Tr}\Big(\widehat{\mathbf{v}}_i\widehat{\mathbf{v}}_i^\top\big(\frac{1}{n}K-\frac{1}{n}\widetilde{K}\big)^2\Big)+2\big(\tilde{\lambda}_i-\lambda_i\big)^2,\end{aligned}$$

which implies that

$$\begin{aligned}\sum_{i=1}^{n}\Big\|\frac{1}{n}K\widehat{\mathbf{v}}_i-\lambda_i\widehat{\mathbf{v}}_i\Big\|_2^2\ &\leq 2\mathrm{Tr}\Big(\big(\sum_{i=1}^{n}\widehat{\mathbf{v}}_i\widehat{\mathbf{v}}_i^\top\big)\big(\frac{1}{n}K-\frac{1}{n}\widetilde{K}\big)^2\Big)+2\sum_{i=1}^{n}\big(\tilde{\lambda}_i-\lambda_i\big)^2\\ &=2\mathrm{Tr}\Big(\big(\frac{1}{n}K-\frac{1}{n}\widetilde{K}\big)^2\Big)+2\sum_{i=1}^{n}\big(\tilde{\lambda}_i-\lambda_i\big)^2\\ &=2\Big\|\frac{1}{n}K-\frac{1}{n}\widetilde{K}\Big\|_F^2+2\sum_{i=1}^{n}\big(\tilde{\lambda}_i-\lambda_i\big)^2\\ &\leq 4\Big\|\frac{1}{n}K-\frac{1}{n}\widetilde{K}\Big\|_F^2.\end{aligned}$$

The above proves that

$$\begin{aligned}\mathbb{P}\Big(\sqrt{\sum_{i=1}^{n}\Big\|\frac{1}{n}K\widehat{\mathbf{v}}_i-\lambda_i\widehat{\mathbf{v}}_i\Big\|_2^2}\geq\epsilon\Big)\ &\leq\ \mathbb{P}\Big(\big\|\frac{1}{n}\widetilde{K}-\frac{1}{n}K\big\|_F\ \geq\ \frac{\epsilon}{2}\Big)\\ &<\ \frac{n^2}{2}\exp\Big(-\frac{r\epsilon^2}{8}\Big).\end{aligned}$$

Therefore, considering the provided definition $\delta = \frac{n^2}{2}\exp\Big(-\frac{r\epsilon^2}{2}\Big)$, i.e., $2\epsilon = \sqrt{\frac{32\log\big(n/2\delta\big)}{r}}$, we will have the following which completes the proof:

$$\mathbb{P}\Big(\sqrt{\sum_{i=1}^{n}\Big\|\frac{1}{n}K\widehat{\mathbf{v}}_i-\lambda_i\widehat{\mathbf{v}}_i\Big\|_2^2}\leq 2\epsilon\Big)\ \geq\ 1-\delta.$$

Figure 7: RFF-based identified clusters used in FKEA Evaluation in ImageNet dataset with *DinoV2* embeddings and bandwidth $\sigma = 25$ at varying number of samples n

https://doi.org/10.52202/079017-3841

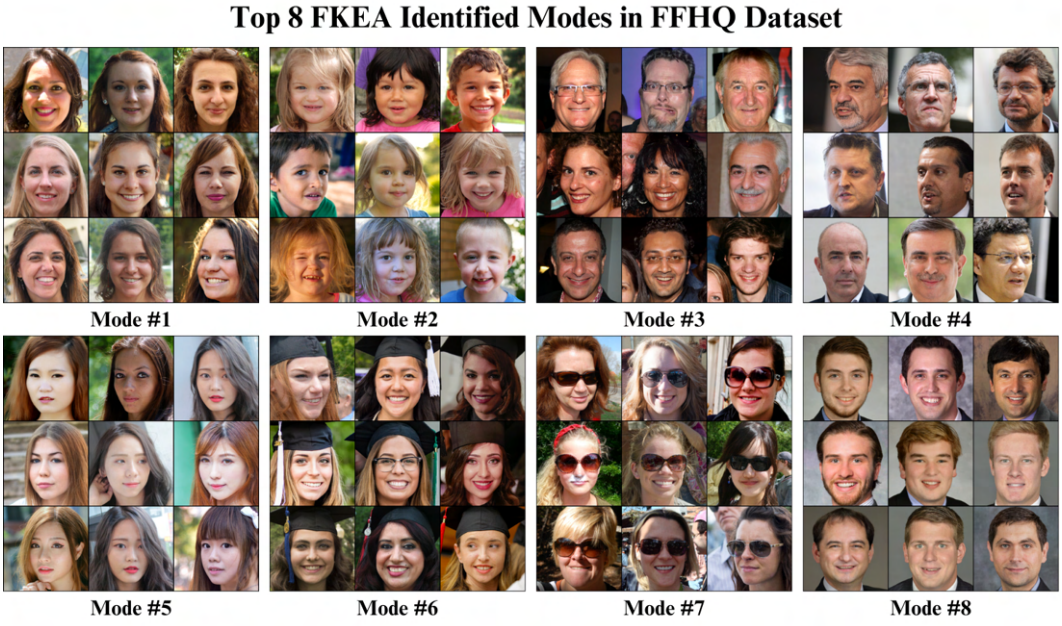


Figure 8: RFF-based identified clusters used in FKEA Evaluation in FFHQ dataset with *DinoV2* embeddings and bandwidth $\sigma = 20$

B Limitations

Incompatibility with non shift-invariant kernels. Our analysis targets a shift-invariant kernel, which does not apply to a general kernel function, such as polynomial kernels. In practice, many ML algorithms rely on simpler kernels that may not have the shift-invariant property. Due to to specifics of FKEA framework, we cannot directly extend the work to such kernels. We leave the framework's extension to other kernel functions for future studies.

Reliance on Embeddings. FKEA clustering and diversity assessment metrics rely on the quality of the underlying embedding space. Depending on the training and pre-training datasets, the semantic clustering properties may change. We leave in-depth study of embedding space behavior for future research.

C Additional Numerical Results

C.1 Real Image Dataset Modes

This section details the results of cluster analyses conducted on various real-world datasets, including FFHQ, AFHQ, MSCOCO, and Fashion-MNIST. Each dataset's results are organized into clusters identified by RFF method in FKEA evaluation.

C.2 The effect of number of datapoints on clustering results with FKEA

In this section, we evaluate the quality of clusters obtained from the ImageNet dataset as the number of samples n varies. Specifically, we compare clustering results for 10k, 50k, 100k, and 250k samples. Figure 7 illustrates the first four modes derived from the FKEA framework.

At $n = 10k$, the clusters exhibit noise and often merge unrelated modes, as seen in Mode 2, where elephants and foxes appear within the same cluster. As n increases, the clustering quality improves, becoming more coherent and meaningful. This trend is particularly evident in Modes 1 and 2, where the clusters more accurately reflect distinct semantic groups.

These findings highlight the importance of scaling VENDI and RKE scores, as computational overhead becomes a critical factor in assessing the diversity of generative models. Scaling these

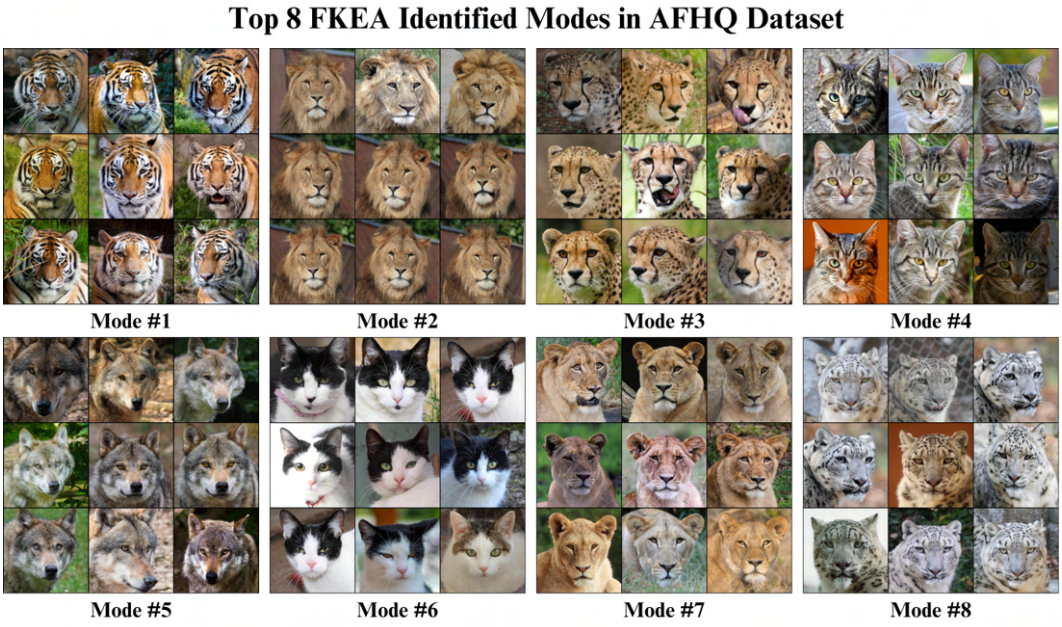


Figure 9: RFF-based identified clusters used in FKEA Evaluation in AFHQ dataset with *DinoV2* embeddings and bandwidth $\sigma = 20$

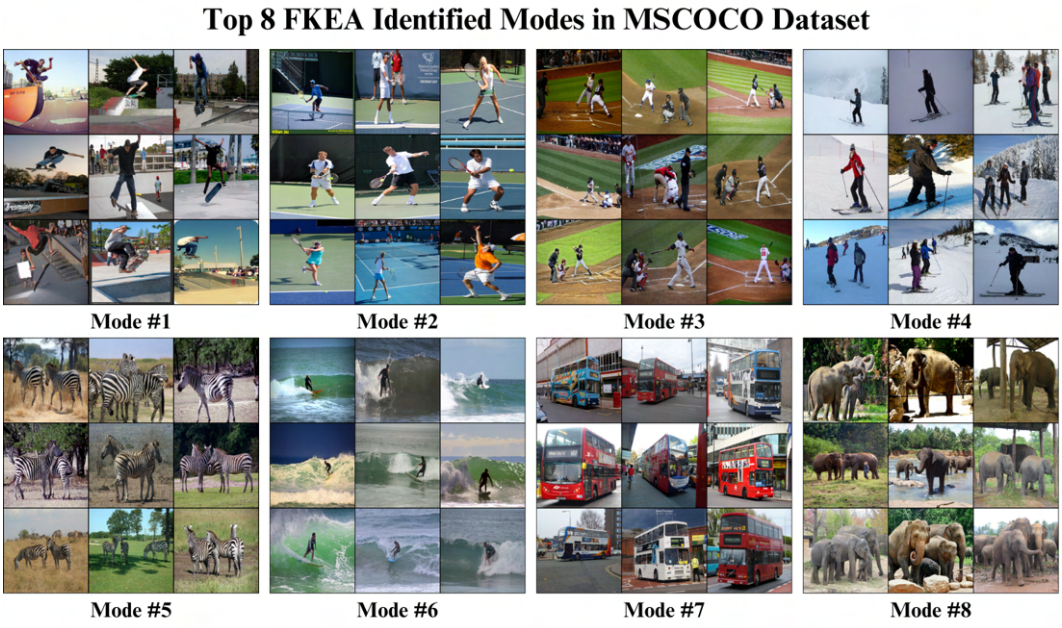


Figure 10: RFF-based identified clusters used in FKEA Evaluation in Microsoft COCO dataset with *DinoV2* embeddings and bandwidth $\sigma = 22$

https://doi.org/10.52202/079017-3841

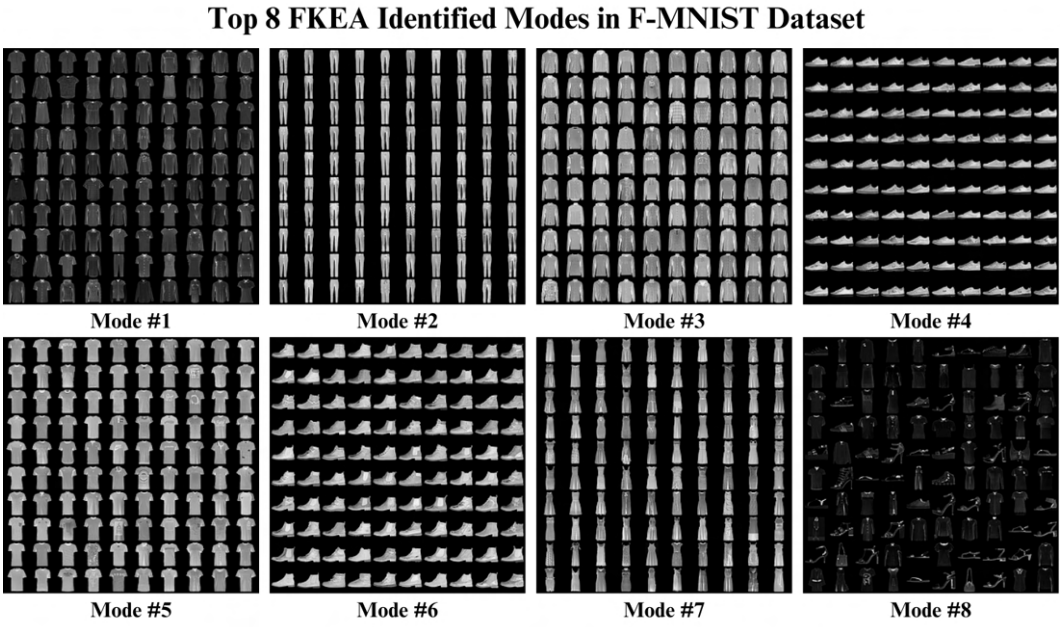


Figure 11: RFF-based identified clusters used in FKEA Evaluation in FASHION-MNIST [50] dataset with *pixel* embeddings and bandwidth $\sigma = 15$

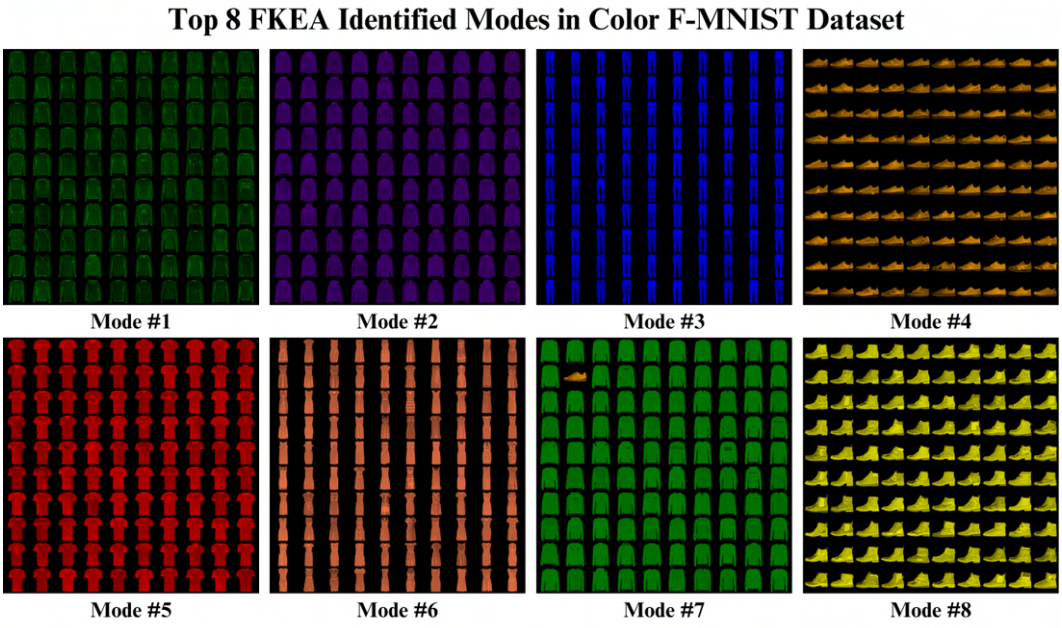


Figure 12: RFF-based identified clusters used in FKEA Evaluation in colored FASHION-MNIST [50] dataset with *pixel* embeddings and bandwidth $\sigma = 4.5$

Table 3: Evaluated scores for ImageNet generative models. The Gaussian kernel bandwidth parameter chosen for RKE, VENDI, FKEA-VENDI and FKEA-RKE is $\sigma = 25$ and Fourier features dimension $2r = 16k$. The scores were obtained by running the GitHub of [20] on pre-generated 50k samples.

Method	IS ↑	FID ↓	Precision ↑	Recall ↑	Density ↑	Coverage ↑	FKEA VENDI-1 ↑	FKEA RKE ↑
Dataset (100k)	-	-	-	-	-	-	9176.9	996.7
ADM [65]	542.6	11.12	0.78	0.79	0.88	0.89	8360.3	633.4
ADMG [65]	659.3	5.63	0.87	0.84	0.80	0.85	8524.2	811.5
ADMG-ADMU [65]	701.6	4.78	0.90	0.73	1.20	0.96	8577.6	839.8
BigGAN [66]	696.4	7.91	0.81	0.44	0.99	0.57	7120.5	492.4
DiT-XL-2 [67]	743.2	3.56	0.92	0.84	1.16	0.97	8626.5	855.8
GigaGAN [68]	678.8	4.29	0.89	0.74	0.74	0.70	8432.5	671.6
LDM [69]	734.4	4.75	0.93	0.76	1.04	0.93	8573.7	811.9
Mask-GIT [70]	717.4	5.66	0.91	0.72	1.01	0.82	8557.4	759.5
RQ-Transformer [71]	558.3	9.57	0.80	0.76	0.77	0.59	8078.4	512.1
StyleGAN-XL[72]	675.4	4.34	0.89	0.74	1.18	0.96	8171.9	703.5

metrics allows for a more efficient evaluation, especially when dealing with large datasets and high sample counts.

C.3 Comparison between Generative Models on ImageNet dataset

In this section we report the FKEA scores for various generative models on ImageNet dataset. Table 3 evaluates the diversity scores of various ImageNet GAN models using the FKEA method applied to VENDI-1 and RKE, with potential extension to the entire VENDI family. The comparison includes baseline diversity metrics such as Inception Score [12], FID [7], Improved Precision/Recall [10], and Density/Coverage [11].

C.4 Synthetic Image Dataset Modes

In addition to running clustering on ImageNet dataset, we also applied FKEA with varying Gaussian Kernel bandwidth parameter σ to other datasets. The results are presented for FFHQ (Figure 8), AFHQ (Figure 9) Microsoft COCO (Figure 10) and Mono/Color versions of F-MNIST[50] (Figures 11 and 12) up to top 8 modes.

The experimental setup is similar to figure 3 with the only change is optimised bandwidth for each dataset, since datasets differ in number and typicality of the samples.

C.5 Effect of other embeddings on FKEA clustering

Even though DinoV2 is a primary embedding in our experimental settings, we acknowldge the use of other embedding models such as SwAV[24] and CLIP[23]. The resulting clusters differ from original DinoV2 clusters and require separate bandwidth parameter finetuning. In our experiments, SwAV embedding emphasizes object placement, such as animal in grass or white backgrounds, as seen in Figure 17. CLIP on the other hand clusters by objects, such as birds/dogs/bugs, as seen in Figure 18. These results indicate that FKEA powered by other embeddings will slightly change the clustering features; however, it does not hinder the clustering performance of RFF based clustering with FKEA method.

C.6 Effect of embeddings on score convergence

To highlight the compatibility of FKEA across diverse embedding spaces, we conducted convergence experiments on various text and image embeddings. Figure 19 presents the convergence results of the VENDI and RKE scores, comparing both FKEA and non-FKEA counterparts. Our findings show that the convergence remains consistent across different embedding spaces, demonstrating the robustness of the proposed method.

C.7 Text Dataset Modes

To understand the applicability and effectiveness of the FKEA method beyond images, we extended our study to text datasets. We observed that clustering text data poses a more challenging task

 https://doi.org/10.52202/079017-3841

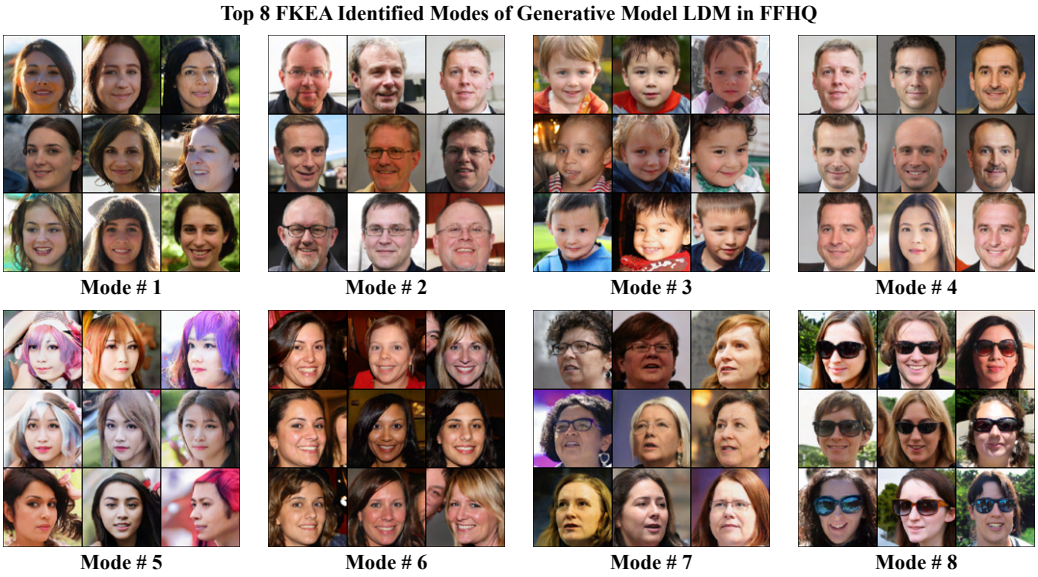


Figure 13: RFF-based identified clusters used in FKEA Evaluation of LDM [69] generative model in FFHQ with *DINOv2* embeddings and bandwidth $\sigma = 20$

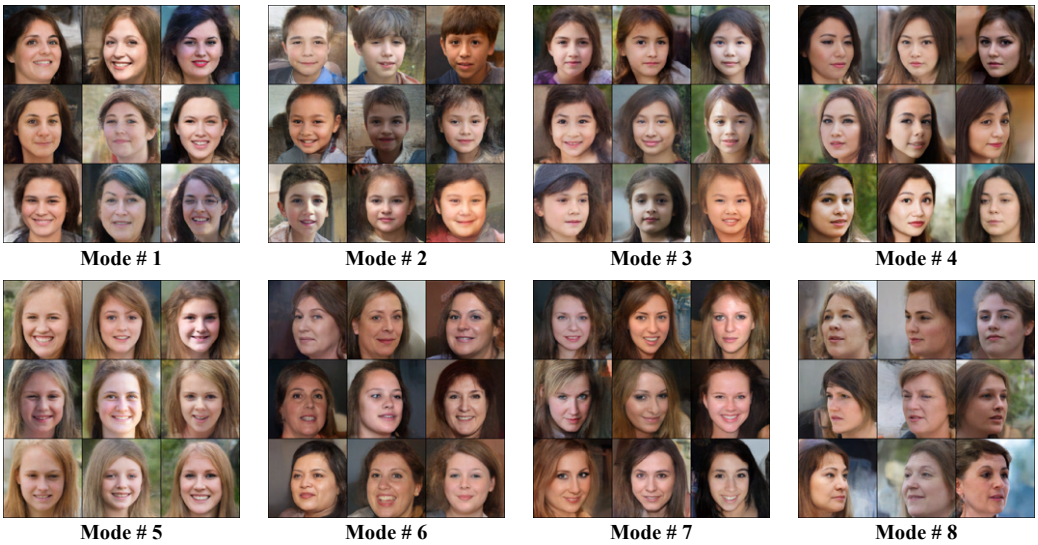


Figure 14: RFF-based identified clusters used in FKEA Evaluation of VDVAE [73] generative model in FFHQ with *DINOv2* embeddings and bandwidth $\sigma = 20$

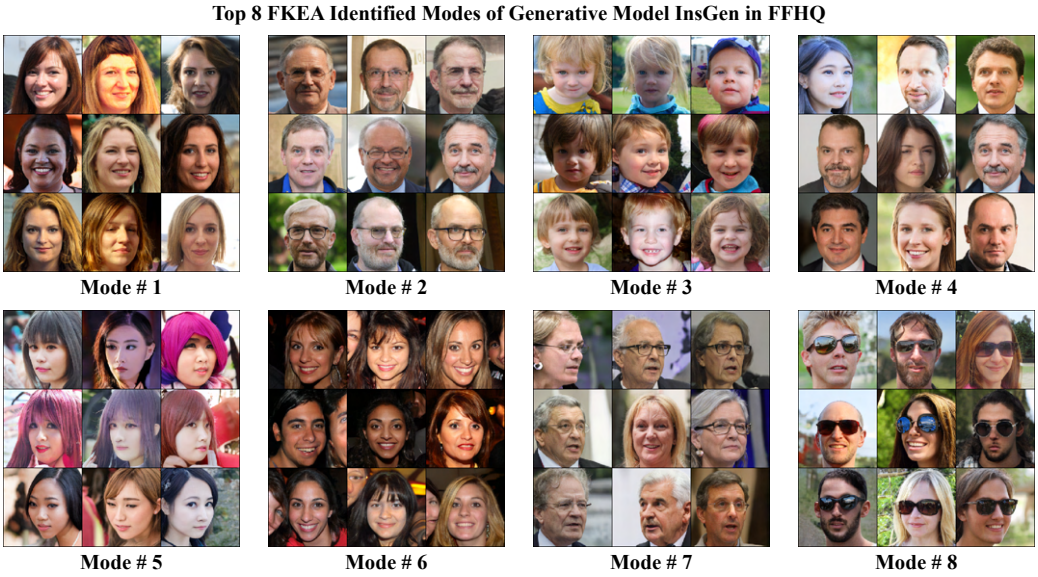


Figure 15: RFF-based identified clusters used in FKEA Evaluation of InsGen [74] generative model in FFHQ with *DINOv2* embeddings and bandwidth $\sigma = 20$

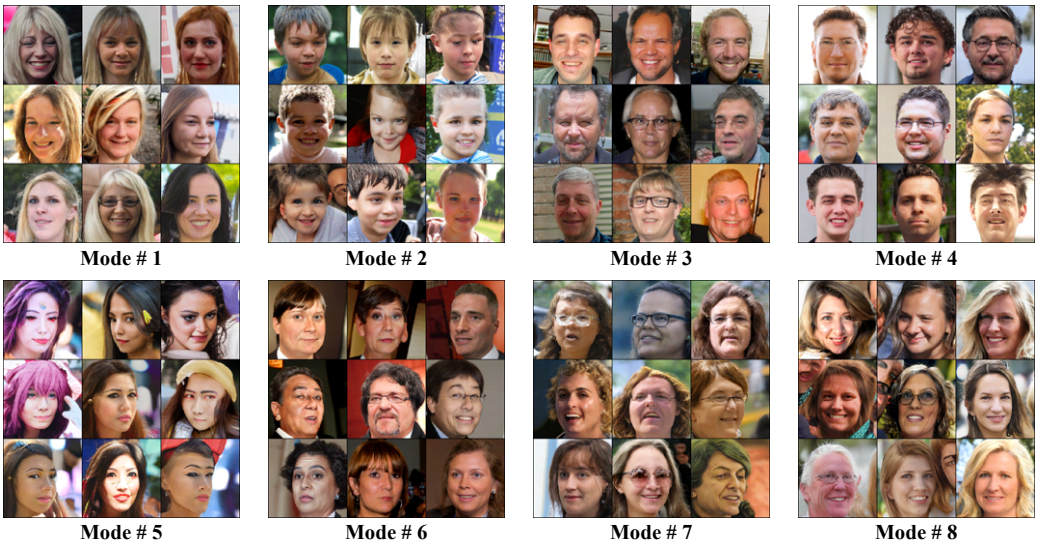


Figure 16: RFF-based identified clusters used in FKEA Evaluation of StyleGAN-XL[72] generative model in FFHQ with *DINOv2* embeddings and bandwidth $\sigma = 20$

https://doi.org/10.52202/079017-3841

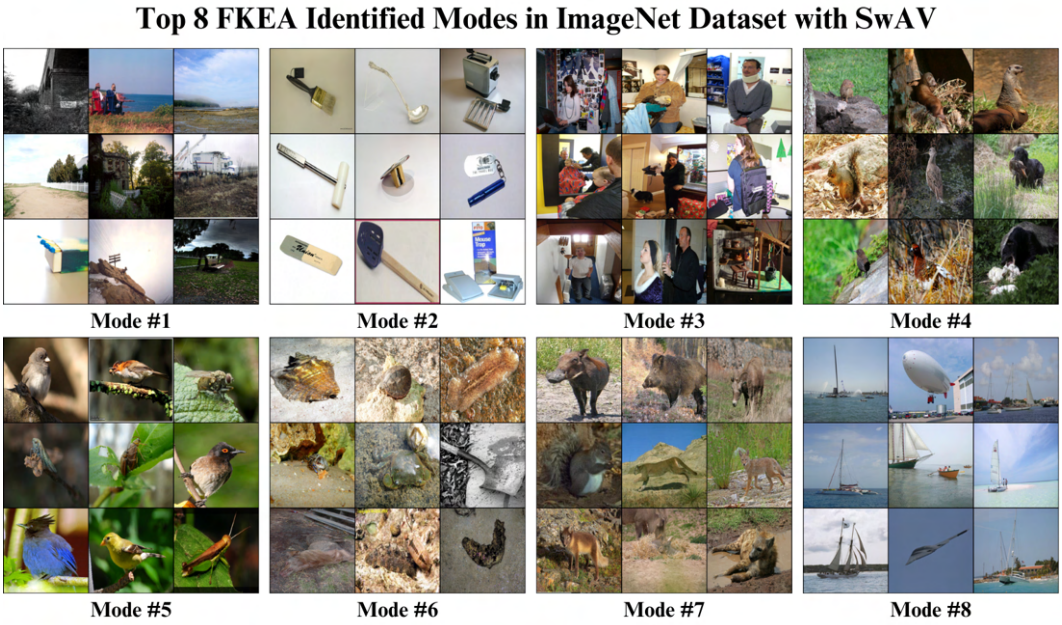


Figure 17: RFF-based identified clusters used in FKEA Evaluation of *SwAV* embedding on ImageNet with bandwidth $\sigma = 0.8$

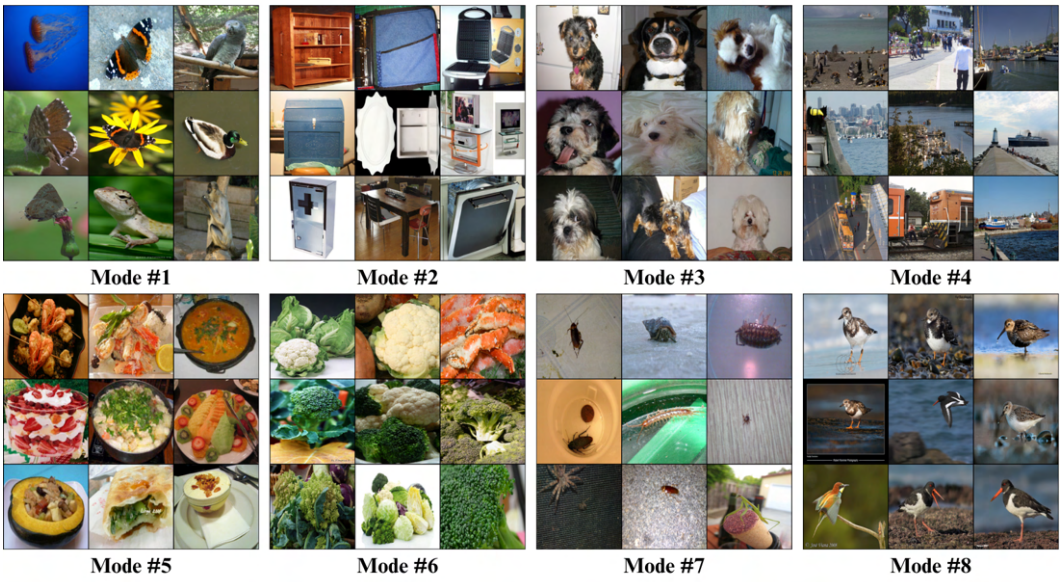
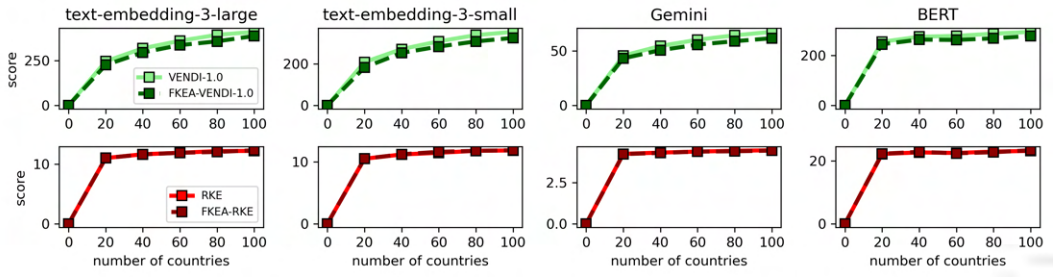
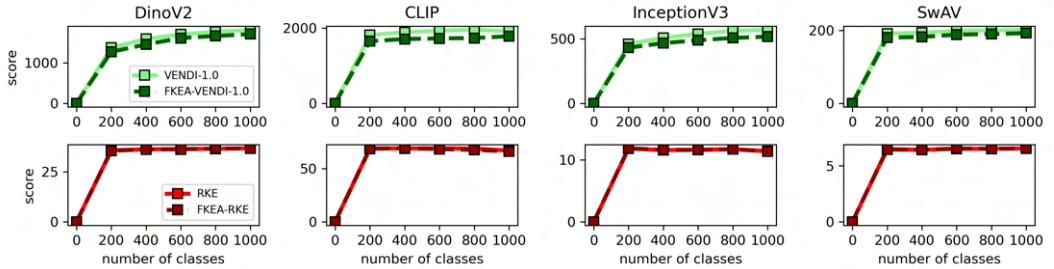


Figure 18: RFF-based identified clusters used in FKEA Evaluation of *CLIP* embedding on ImageNet with bandwidth $\sigma = 7.0$



(a) Diversity convergence on synthetic countries dataset across various text embeddings



(b) Diversity convergence on ImageNet dataset across various image embeddings

Figure 19: Summary of diversity convergence with $r = 12000$ and sample size $n = 20000$.

Mode #1	**Mode #2**	**Mode #3**	**Mode #4**	**Mode #5**
Grosse Pointe	Ishkli	Gerson Garca	Girugamesh (album)	2009 WPSL season
Mark Scharf	Khazora	Valentin Mogilny	Japonesque (album)	2012 Milwaukee...
Alexander McKee	Sis, Azerbaijan	Gerald Lehner (referee)	Documentaly	2020 San Antonio FC...
Clay Huffman	Zasun	Dmitri Nezhelev	EX-Girl	2020 HFX Wanderers FC...
Ravenna, Ohio	Zaravat	Grigori Ivanov	Indie 2000	2020 Sporting Kansas City...
C. M. Eddy Jr.	Bogat	Leonidas Morakis	Triangle (Perfume album)	FC Tucson
Hornell, New York	Yakkakhona	Jos Luis Alonso Ber...	Waste Management (album)	2008 K League
Larchmont, New York	Yava, Tajikistan	Giovanni Gasperini	Fush Yu Mang	201112 New Zealand Football...
Robert Hague	Ikizyak	Mohamed Chab	Fantastipo (song)	200809 Melbourne Victory FC...
General Hershy Bar	Khushikat	Louis Darmanin	Xtort	2012 Pittsburgh Power season
		Keywords		
London	populated places	players category	music video	American football
American History	Maplandia.com Category	Association football	album	players Category
University Press	municipality	FIFA World	studio album	Football League
United States	village	World Cup	Records albums	League
World War	Osh Region	Summer Olympics	Singles Chart	League Soccer

Table 4: Top 5 Wikipedia Dataset Modes with corresponding eigenvalues with *text-embedding-3-large* embeddings and bandwidth $\sigma = 1.0$

compared to image data. This increased difficulty arises from the ambiguity in defining clear separability factors within text, a contrast to the more visually distinguishable criteria in images. The process of evaluating text clusters is not straightforward and often varies significantly based on human judgment and perception.

To visualise the results, we use YAKE [75] algorithm to extract the keywords in each text mode and present the identified unigram and bigram keywords. We demonstrate that the results hold for text datasets and identified clusters are meaningful.

Table 4 displays the identified clusters associated with Wikipedia article titles and keywords analyzed using the FKEA method. Identified mode 1 correlates most with historical figures/events/places. Mode 2 clusters smaller villages and rural regions together. Mode 3 is exclusively about people in sports, such as athletes and referees. Mode 4 visualises various music bands and albums. Lastly, mode 5 presents the articles about sports events, such as football leagues.

Table 5 outlines the largest modes identified within a news dataset analyzed using the FKEA method, with a detailed focus on the content themes of each mode. The most dominant mode is associated with topics related to crime and police activities, indicating a frequent coverage area in the dataset.

https://doi.org/10.52202/079017-3841

Mode #1	Mode #2	Mode #3	Mode #4	Mode #5
police	President Obama	size	people	died
British police	Obama	year	severe weather	family
police officer	Barack Obama	weight	Death toll	plane crash
Police found	President	dress size	heavy rain	mother
family	White House	stone	Environment Agency	plane
found	Obama administration	Slimming World	million people	found
told police	Obama calls	lost	rain	people
court	House	lose weight	flood warnings	children
arrested	United States	diet	people dead	hospital
home	Obama plan	model	people killed	found dead

Table 5: Top 5 CNN/Dailymail 3.0.0 [54][55] Dataset Modes with corresponding eigenvalues with *text-embedding-3-large* embeddings and bandwidth $\sigma = 0.8$.

Mode #1	Mode #2	Mode #3	Mode #4	Mode #5
The House on Tele..	Anand	Bring Your Smile Along	Chhota Bheem...	Walk a Crooked Mile
Seems Like Old Times	I Love You	The Girl Most Likely	Duck Amuck	Assignment to Kill
Shadows and Fog	Toh Baat Pakki	Hips, Hips, Hooray!	Hare-Abian Nights	The Crime of the Century
Obsession	Abodh	Lady Be Good	Porky's Five and Ten	Murder at Glen Athol
Milk Money	Khulay Aasman...	The Courtship of Eddie's...	Sock-a-Doodle-Do	Guns
Very Bad Things	Kasthuri Maan	You Live and Learn	Buccaneer Bunny	Because of the Cats
Blame It on the Bell...	Chhaya	Dames	Hare Lift	The House of Hate
The Miracle Man	Yeh Dillagi	Painting the Clouds...	Scrap Happy Daffy	The Ace of Scotland Yard
The Sleeping Tiger	Deva	Pin Up Girl	Hic-cup Pup	The World Gone Mad
The Scapegoat	Bhalobasa Bha...	Too Young to Kiss	The Goofy Gophers	Firepower
		Keywords		
mystery	hindi film	musical	animation	crime
noir	romance	theme songs	Tom & Jerry	murder
kidnap	love	city	Spike	detective
crime	marriage	romance	adventure	investigation
police	daughter		comedy	killer

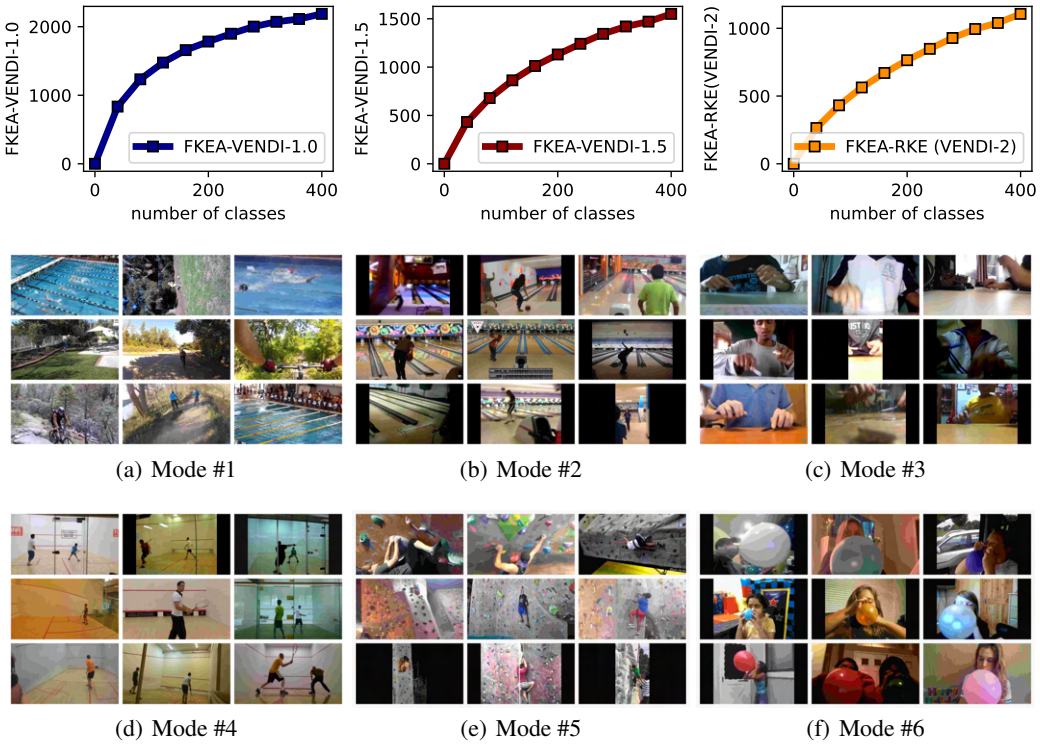
Table 6: Top 5 CMU Movie Summary Corpus [56] Dataset Modes with corresponding eigenvalues with *text-embedding-3-large* embeddings and bandwidth $\sigma = 0.8$. The table summarises the assigned genres to each movie in the first 100 paragraphs in each mode.

Mode 2 is closely correlated with President Obama, reflecting a significant focus on political coverage. Mode 3 pertains to dieting, which suggests a presence of health and lifestyle topics. Mode 4 is linked to environmental disasters, highlighting the dataset's attention to ecological and crisis-related news. Finally, Mode 5 deals with plane crash accidents, underscoring the coverage of major transportation incidents.

Table 6 delineates the distribution of genres and production types within a dataset of movie summaries analyzed using the FKEA method. The first mode predominantly covers drama TV shows without focusing on any specific subtopic, indicating a broad categorization within this genre. From mode 2 onwards, the features become more distinct and defined. Mode 2 specifically represents Bollywood movies, with a significant emphasis on the Romance genre. Mode 3 is dedicated to clustering comedy shows. Mode 4 is exclusively associated with cartoons, evidenced by keywords such as "Tom & Jerry". Lastly, mode 5 clusters together detective and crime fiction shows.

C.8 Video Dataset Modes

In this section, we present additional experiments on the Kinetics-400[58] video dataset. This dataset comprises 400 human action categories, each with a minimum of 400 video clips depicting the action. Similar to the video evaluation metrics, we used the I3D pre-trained model[61] which maps each video to a 1024-vector feature. Figure 20, the first mode captured broader concepts while the other models focused on specific ones. Also, the plots indicate that increasing the number of classes from 40 to 400 results in an increase in the FKEA metrics.



(a) Mode #1 (b) Mode #2 (c) Mode #3

(d) Mode #4 (e) Mode #5 (f) Mode #6

Figure 20: RFF-based identified clusters used in FKEA Evaluation in Kinetics-400 dataset with *I3D* embeddings. Plots indicate that increasing the number of classes from 40 to 400 results in an increase in the FKEA metrics.

https://doi.org/10.52202/079017-3841

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The paper discusses and further expands the ideas of generative model entropic diversity evaluation presented in the abstract and introduction.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: In the paper we discuss the limitations of the reference-free metrics (RKE, VENDI) approximation via the Random Fourier Features within Fourier-based Kernel Entropy Approximation (FKEA) method.

   Guidelines:

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

3. **Theory Assumptions and Proofs**

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

Answer: [Yes]

Justification: The paper provides preliminary information, assumptions and definitions of the theoretical results along with relevant proofs.

Guidelines:

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

4. **Experimental Result Reproducibility**

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

   Answer: [Yes]

   Justification: The paper outlines all necessary parameters in experimental settings for reproducibility purposes.

   Guidelines:

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

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

https://doi.org/10.52202/079017-3841

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

Answer: [Yes]

Justification: The submission comes along with demo code to generate clusters and compute the scores.

Guidelines:

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

6. **Experimental Setting/Details**

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

Answer: [Yes]

Justification: Each experimental result is accompanied by relevant parameters that aid in understanding of presented results.

Guidelines:

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

7. **Experiment Statistical Significance**

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

Answer: [No]

Justification: Error bars are not reported due to limited amount of ground truth samples in the datasets.

Guidelines:

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

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

8. **Experiments Compute Resources**

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

   Answer: [Yes]

   Justification: The paper states the computational resources used in the experimental setups.

   Guidelines:

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

9. **Code Of Ethics**

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

   Answer: [Yes]

   Justification: The paper closely follows the code of ethics and all generated/downloaded data has through sanity checks.

   Guidelines:

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

10. **Broader Impacts**

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

    Answer: [NA]

    Justification: The paper discuss computational complexity and diversity evaluation of existing metrics (RKE, VENDI) and does not directly impact training of models.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

https://doi.org/10.52202/079017-3841

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

11. **Safeguards**

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

    Answer: [NA]

    Justification: In this paper we discuss the evaluation of existing datasets and generative models. We do not release any custom datasets to the public.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All relevant datasets and models are referenced in the paper and supplemental materials.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

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

13. **New Assets**

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

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

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

14. **Crowdsourcing and Research with Human Subjects**

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

    Answer: [NA]

    Justification: The paper does not involve human evaluation and crowdsourcing.

    Guidelines:

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

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

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

    Answer: [NA]

    Justification: The paper does not involve human evaluation and crowdsourcing.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

 https://doi.org/10.52202/079017-3841

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.