Do Vendi Scores Converge with Finite Samples? Truncated Vendi Score for Finite-Sample Convergence Guarantees

Azim Ospanov[1] **Farzan Farnia**[1]

[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong

Abstract

Evaluating the diversity of generative models without reference data poses methodological challenges. The reference-free Vendi [Friedman and Dieng, 2023] and RKE [Jalali et al., 2023] scores address this by quantifying the diversity of generated data using matrix-based entropy measures. Among these two, the Vendi score is typically computed via the eigendecomposition of an $n \times n$ kernel matrix constructed from n generated samples. However, the prohibitive computational cost of eigendecomposition for large n often limits the number of samples used to fewer than 20,000. In this paper, we investigate the statistical convergence of the Vendi and RKE scores under restricted sample sizes. We numerically demonstrate that, in general, the Vendi score computed with standard sample sizes below 20,000 may not converge to its asymptotic value under infinite sampling. To address this, we introduce the *t-truncated Vendi score* by truncating the eigenspectrum of the kernel matrix, which is provably guaranteed to converge to its population limit with $n = \mathcal{O}(t)$ samples. We further show that existing Nyström and FKEA approximation methods converge to the asymptotic limit of the truncated Vendi score. In contrast to the Vendi score, we prove that the RKE score enjoys universal convergence guarantees across all kernel functions. We conduct several numerical experiments to illustrate the concentration of Nyström and FKEA computed Vendi scores around the truncated Vendi score, and we analyze how the truncated Vendi and RKE scores correlate with the diversity of image and text data. The code is available at `https://github.com/aziksh-ospanov/truncated-vendi`.

1 INTRODUCTION

The increasing use of generative artificial intelligence has underscored the need for accurate evaluation of generative models. In practice, users often have access to multiple generative models trained with different training datasets and algorithms, requiring evaluation methods to identify the most suitable model. The feasibility of a model evaluation approach depends on factors such as the required generated sample size, computational cost, and the availability of reference data. Recent studies on evaluating generative models have introduced assessment methods that relax the requirements on data and computational resources.

Specifically, to enable the evaluation of generative models without reference data, the recent literature has focused on reference-free evaluation scores that remain applicable in the absence of reference samples. The Vendi score [Friedman and Dieng, 2023] is one such reference-free metric that quantifies the diversity of generated data using the entropy of a kernel similarity matrix formulated for the generated samples. Given the sorted eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ of the normalized matrix $\frac{1}{n}K$[1] for the kernel similarity matrix $K = \big[k(x_i, x_j)\big]_{1\leq i,j\leq n}$ of n generated samples $x_1, \ldots, x_n$, the definition of (order-1) Vendi score is as:

$$\mathrm{Vendi}(x_1, \ldots, x_n) := \exp\Big(\sum_{i=1}^{n} \lambda_i \log \frac{1}{\lambda_i}\Big) \tag{1}$$

Following conventional definitions in information theory, the Vendi score corresponds to the exponential of the *Von Neumann entropy* of normalized kernel matrix $\frac{1}{n}K$. More generally, Jalali et al. [2023] define the Rényi Kernel Entropy (RKE) score by applying order-2 Rényi entropy to this matrix, which reduces to the inverse-squared Frobenius norm of the normalized kernel matrix:

$$\mathrm{RKE}(x_1, \ldots, x_n) := \frac{1}{\left\|\frac{1}{n}K\right\|_F^2} \tag{2}$$

[1]In general, we consider the trace-normalized kernel matrix $\frac{1}{\mathrm{Tr}(K)}K$, which given $\forall x:\ k(x,x) = 1$, reduces to $\frac{1}{n}K$.

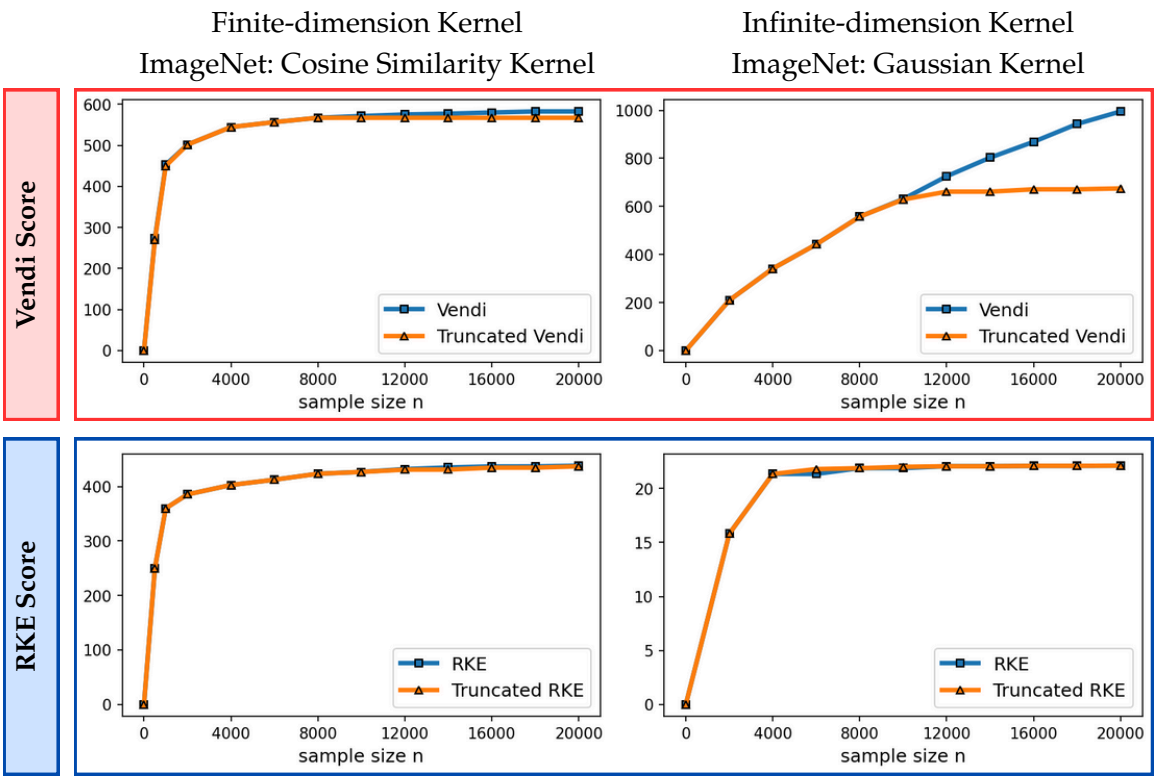

Figure 1: Statistical convergence of Vendi and RKE scores for different sample sizes on ImageNet data: (Left plots) finite-dimension cosine similarity kernel (Right plots) infinite dimension Gaussian kernel with bandwidth $\sigma = 30$. The RKE and truncated Vendi scores converged with below 20000 samples, but the Vendi score with Gaussian kernel did not converge.

Although the Vendi and RKE scores do not require reference samples, their computational cost increases rapidly with the number of generated samples n. Specifically, calculating the Vendi score for the $n \times n$ kernel matrix K generally involves an eigendecomposition of K, requiring $O(n^3)$ computations. Therefore, the computational load of Vendi score becomes substantial for a large sample size n, and the Vendi score is typically evaluated for sample sizes limited to 20,000. In other words, the Vendi score, as defined in Equation (2), would be *computationally infeasible* to compute with standard processors for sample sizes greater than a few tens of thousands.

Following the above discussion, a key question that arises is whether the Vendi score estimated from restricted sample sizes (i.e. $n \leq 20000$) has converged to its asymptotic value with infinite samples, which we call the *population Vendi*. However, the statistical convergence of the Vendi score has not been thoroughly investigated in the literature. In this work, we study the statistical convergence of the Vendi and RKE diversity scores and aim to analyze the concentration of the estimated scores from a limited number of generated samples $n \lessapprox 20000$.

1.1 OUR RESULTS ON VENDI'S CONVERGENCE

We discuss the answer to the Vendi convergence question for two types of kernel functions: 1) kernel functions with a finite feature dimension, e.g. the cosine similarity and polynomial kernels, 2) kernel functions with an infinite feature map such as Gaussian (RBF) kernels. For kernel functions with a finite feature dimension d, we theoretically and numerically show that a sample size $n = O(d)$ is sufficient to guarantee convergence to the population Vendi (asymptotic value when $n \to \infty$). For example, the left plot in Figure 1 shows that in the case of the cosine similarity kernel, the Vendi score on n randomly selected ImageNet [Deng et al., 2009] samples has almost converged as the sample size reaches 5000, where the dimension d (using standard DINOv2 embedding [Oquab et al., 2023]) is 768.

In contrast, our numerical results for kernel functions with an infinite feature map demonstrate that for standard datasets, a sample size bounded by 20,000 could be insufficient for convergence of the Vendi score. For example, the right plot of Figure 1 shows the evolution of the Vendi score with the Gaussian kernel on ImageNet data, and the score continues to grow at a significant rate with 20,000 samples[2].

Observing the difference between Vendi score convergence for finite and infinite-dimension kernel functions, a natural question is how to extend the definition of Vendi score from finite to infinite dimension case such that the diversity

[2]The heavy computational cost prohibits an empirical evaluation of the sample size required for Vendi's convergence.

score would statistically converge in both scenarios. We attempt to address the question by introducing an alternative Vendi statistic, which we call the *t-truncated Vendi score*. The t-truncated Vendi score is defined using only the top-t $\lambda_1 \geq \cdots \geq \lambda_t$ eigenvalues of the kernel matrix, where t is an integer hyperparameter. This modified score is defined as

$$\text{Truncated-Vendi}^{(t)}(x_1, \ldots, x_n) = \exp\Big(\sum_{i=1}^{t} \lambda_i^{\text{trunc}} \log \frac{1}{\lambda_i^{\text{trunc}}}\Big)$$

where we shift each of the top-t eigenvalue $\lambda_i^{\text{trunc}} = \lambda_i + c$ by the same constant $c = \big(1 - \sum_{i=1}^{t} \lambda_i\big)/t$ to ensure they add up to 1 and provide a valid probability model. Observe that for a finite kernel dimension d satisfying $d \leq t$, the truncated and original Vendi scores take the same value, because the truncation will have no impact on the eigenvalues. On the other hand, under an infinite kernel dimension, the two scores may take different values.

As a main theoretical result, we prove that a sample size $n = O(t)$ is always enough to estimate the *t-truncated population Vendi* from n empirical samples, regardless of the finiteness of the kernel feature dimension. This result shows that the *t-truncated* Vendi score provides a statistically converging extension of the Vendi score from the finite kernel dimension to the infinite dimension case. To connect the defined t-truncated Vendi score to existing computation methods for the original Vendi score, we show that the existing computationally-efficient methods for computing the Vendi score can be viewed as approximations of our defined *t-truncated* Vendi. Specifically, we show that the Nyström method in [Friedman and Dieng, 2023] and the FKEA method proposed by Ospanov et al. [2024b] provide an estimate of the t-truncated Vendi.

1.2 OUR RESULTS ON RKE'S CONVERGENCE

For the RKE score, we prove a universal convergence guarantee that holds for every kernel function. The theoretical guarantee shows that the RKE score, and more generally every order-α entropy score with $\alpha \geq 2$, will converge to its population value within $O(\frac{1}{\sqrt{n}})$ error for n samples. Our theoretical guarantee also transfers to the truncated version of the RKE score. However, note that the truncation of the eigenspectrum becomes unnecessary in the RKE case, since the score enjoys universal convergence guarantees. Figure 1 shows that using both the cosine-similarity and Gaussian kernel functions, the RKE score nearly converges to its limit value with less than 10000 samples.

Finally, we present the findings of several numerical experiments to validate our theoretical results on the convergence of Vendi, truncated Vendi, and RKE scores. Our numerical results on standard image, text, and video datasets and generative models indicate that in the case of a finite-dimension kernel map, the Vendi score can converge to its asymptotic limit, in which case, as we explained earlier, the Vendi score is identical to the truncated Vendi. On the other hand, in the case of infinite-dimension Gaussian kernel functions, we numerically observe the growth of the score beyond n =10,000. Our numerical results further confirm that the scores computed by Nyström method in [Friedman and Dieng, 2023] and the FKEA method [Ospanov et al., 2024b] provide tight estimations of the population truncated Vendi. The following summarizes this work's contributions:

- Analyzing the statistical convergence of Vendi and RKE diversity scores under restricted sample sizes $n \lessapprox 2 \times 10^4$,
- Providing numerical evidence on the Vendi score's lack of convergence for infinite-dimensional kernel functions, e.g. the Gaussian (RBF) kernel,
- Introducing the truncated Vendi score as a statistically converging extension of the Vendi score from finite to infinite dimension kernel functions,
- Demonstrating the universal convergence of the RKE diversity score across all kernel functions.

2 RELATED WORKS

Diversity evaluation for generative models Diversity evaluation in generative models can be categorized into two primary types: reference-based and reference-free methods. Reference-based approaches rely on a predefined dataset to assess the diversity of generated data. Metrics such as FID [Heusel et al., 2018], KID and distributed KID [Bińkowski et al., 2018, Wang et al., 2023] measure the distance between the generated data and the reference, while Recall [Sajjadi et al., 2018, Kynkäänniemi et al., 2019] and Coverage [Naeem et al., 2020] evaluate the extent to which the generative model captures existing modes in the reference dataset. Pillutla et al. [2021, 2023] propose MAUVE metric that uses information divergences in a quantized embedding space to measure the gap between generated data and reference distribution. In contrast, the reference-free metrics, Vendi [Friedman and Dieng, 2023] and RKE [Jalali et al., 2023], assign diversity scores based on the eigenvalues of a kernel similarity matrix of the generated data. Jalali et al. [2023] interpret the approach as identifying modes and their frequencies within the generated data followed by entropy calculation for the frequency parameters. The Vendi and RKE scores have been further extended to quantify the diversity of conditional prompt-based generative AI models [Ospanov et al., 2024a, Jalali et al., 2024] and to select generative models in online settings [Rezaei et al., 2024, Hu et al., 2024, 2025]. Also, [Zhang et al., 2024, 2025, Jalali et al., 2025, Gong et al., 2025, Wu et al., 2025] extend the entropic kernel-based scores to measure novelty and embedding dissimilarity. In our work, we specifically focus on the statistical convergence of the vanilla Vendi and RKE scores.

Statistical convergence analysis of kernel matrices' eigenvalues. The convergence analysis of the eigenvalues of

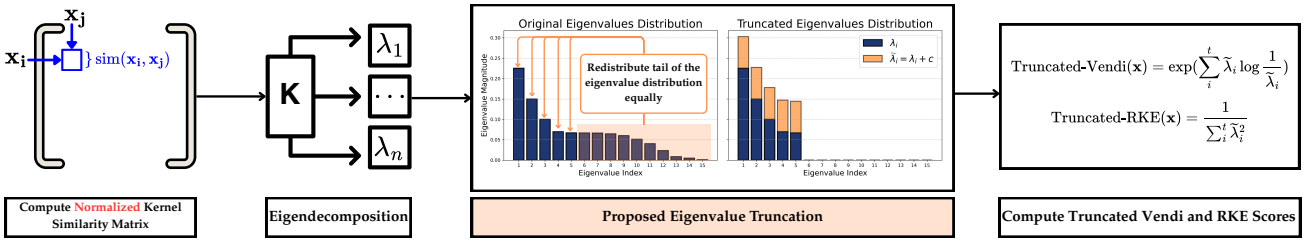

Figure 2: Computation of the proposed t-truncated Vendi score. The kernel similarity matrix eigenspectrum is truncated, and the mass of the truncated tail (excluding the top-t eigenvalues) is uniformly redistributed among the top-t eigenvalues.

kernel matrices has been studied by several related works. Shawe-Taylor et al. [2005] provide a concentration bound for the eigenvalues of a kernel matrix. We note that the bounds in [Shawe-Taylor et al., 2005] use the expectation of eigenvalues $\mathbb{E}_m[\hat{\boldsymbol{\lambda}}(S)]$ for a random dataset $S = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$ of fixed size m as the center vector in the concentration analysis. However, since eigenvalues are nonlinear functions of a matrix, this concentration center vector $\mathbb{E}_m[\hat{\boldsymbol{\lambda}}(S)]$ does not match the eigenvalues of the asymptotic kernel matrix as the sample size approaches to infinity. On the other hand, our convergence analysis focuses on the asymptotic eigenvalues with an infinite sample size, which determines the limit value of Vendi scores. In another related work, Bach [2022] discusses a convergence result for the Von-Neumann entropy of kernel matrix. While this result proves a non-asymptotic guarantee on the convergence of the entropy function, the bound may not guarantee convergence at standard sample sizes for computing Vendi scores (less than 10000 in practice). In our work, we aim to provide convergence guarantees for the finite-dimension and generally truncated Vendi scores with restricted sample sizes.

Efficient computation of matrix-based entropy. Several strategies have been proposed in the literature to reduce the computational complexity of matrix-based entropy calculations, which involve the computation of matrix eigenvalues—a process that scales cubically with the size of the dataset. Dong et al. [2023] propose an efficient algorithm for approximating matrix-based Renyi's entropy of arbitrary order α, which achieves a reduction in computational complexity down to $O(n^2sm)$ with $s, m \ll n$. Additionally, kernel matrices can be approximated using low-rank techniques such as incomplete Cholesky decomposition [Fine and Scheinberg, 2001, Bach and Jordan, 2002] or CUR matrix decompositions [Mahoney and Drineas, 2009], which provide substantial computational savings. Pasarkar and Dieng [2024] suggest to leverage Nyström method [Williams and Seeger, 2000] with m components, which results in $O(nm^2)$ computational complexity. Further reduction in complexity is possible using Random Fourier Features, as suggested by Ospanov et al. [2024b], which allows the computation to scale linearly with $O(n)$ as a function of the dataset size. This work focuses on the latter two methods and the population quantities estimated by them.

Impact of embedding spaces on diversity evaluation. In our image-related experiments, we used the DinoV2 embedding [Oquab et al., 2023], as Stein et al. [2023] demonstrate the alignment of this embedding with human evaluations. We note that the kernel function in the Vendi score can be similarly applied to other embeddings, including the standard InceptionV3[Szegedy et al., 2016] and CLIP embeddings [Radford et al., 2021] as suggested by Kynkäänniemi et al. [2022].

3 PRELIMINARIES

Consider a generative model $\mathcal{G}$ that generates samples from a probability distribution P_X. To conduct a reference-free evaluation of the model, we suppose the evaluator has access to n independently generated samples from P_X, denoted by $x_1, \ldots, x_n \in \mathcal{X}$. The assessment task is to estimate the diversity of generative model $\mathcal{G}$ by measuring the variety of the observed generated data, $x_1, \ldots x_n$. In the following subsections, we will discuss kernel functions and their application to define the Vendi and RKE diversity scores.

3.1 KERNEL FUNCTIONS AND MATRICES

Following the standard definition, $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a kernel function if for every integer $n \in \mathbb{N}$ and inputs $x_1, \ldots, x_n \in \mathcal{X}$, the following kernel similarity matrix $K \in \mathbb{R}^{n \times n}$ is positive semi-definite (PSD):

$$K = \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{bmatrix} \tag{3}$$

Aronszajn's Theorem [Aronszajn, 1950] shows that this definition is equivalent to the existence of a feature map $\phi : \mathcal{X} \to \mathbb{R}^d$ such that for every $x, x' \in \mathcal{X}$ we have the following where $\langle \cdot, \cdot \rangle$ denotes the standard inner product in the $\mathbb{R}^d$ space:

$$k(x, x') = \langle \phi(x), \phi(x') \rangle \tag{4}$$

In this work, we study the evaluation using two types of kernel functions: 1) finite-dimension kernels where dimension d is finite, 2) infinite-dimension kernels where there is no

feature map satisfying (4) with a finite d value. A standard example of a finite-dimension kernel is the cosine similarity function where $\phi_{\text{cosine}}(x) = x/\|x\|_2$. Also, a widely-used infinite-dimension kernel is the Gaussian (RBF) kernel with bandwidth parameter $\sigma > 0$ defined as

$$k_{\text{Gaussian}(\sigma)}(x, x') := \exp\Big(-\frac{\|x - x'\|_2^2}{2\sigma^2}\Big) \tag{5}$$

Both the mentioned kernel examples belong to normalized kernels which require $k(x, x) = 1$ for every x, i.e., the feature map $\phi(x)$ has unit Euclidean norm for every x. Given a normalized kernel function, the non-negative eigenvalues of the normalized kernel matrix $\frac{1}{n}K$ for n points $x_1, \ldots x_n$ will sum up to 1, i.e., they form a probability model.

3.2 MATRIX-BASED ENTROPY FUNCTIONS AND VENDI SCORE

For a PSD matrix $A \in \mathbb{R}^{d\times d}$ with unit trace $\mathrm{Tr}(A) = 1$, A's eigenvalues form a probability model. The order-α Renyi entropy of matrix A is defined using the order-α entropy of its eigenvalues as

$$H_\alpha(A) := \frac{1}{1-\alpha}\log\Big(\sum_{i=1}^{d}\lambda_i^\alpha\Big) \tag{6}$$

For the special case $\alpha = 2$, one can consider the Frobenius norm $\|\cdot\|_F$ and apply the identity $\|A\|_F^2 = \sum_{i=1}^d \lambda_i^2$ to show $H_2(A) = \log\big(1/\|A\|_F^2\big)$. Moreover, for $\alpha = 1$, the above definition reduces to the Shannon entropy of the eigenvalues as $H_1(A) := \sum_{i=1}^d \lambda_i \log(1/\lambda_i)$ [Rényi, 1961].

Jalali et al. [2023] applies the above definition for order $\alpha = 2$ to the normalized kernel similarity matrix $\frac{1}{n}K$ to define the RKE diversity score (called RKE mode count), which reduces to

$$\mathrm{RKE}(x_1, \ldots, x_n) := \exp\Big(H_2\big(\frac{1}{n}K\big)\Big) = \Big\|\frac{1}{n}K\Big\|_F^{-2} \tag{7}$$

For a general entropy order α, [Friedman and Dieng, 2023, Pasarkar and Dieng, 2024] apply the matrix-based entropy definition to the normalized kernel matrix $\frac{1}{n}K$ and define the order-α Vendi score for samples $x_1, \ldots, x_n$ as

$$\mathrm{Vendi}_\alpha(x_1, \ldots, x_n) := \exp\Big(H_\alpha\big(\frac{1}{n}K\big)\Big) \tag{8}$$

Specifically, for order $\alpha = 1$, the above definition results in the standard (order-1) Vendi score in Equation (2).

3.3 STATISTICAL ANALYSIS OF VENDI SCORE

To derive the population limits of Vendi and RKE scores under infinite sampling, which we call *population Vendi* and *population RKE*, respectively, we review the following discussion from [Bach, 2022, Jalali et al., 2023]. First, note that the normalized kernel matrix $\frac{1}{n}K$, whose eigenvalues are used in the definition of Vendi score, can be written as:

$$\frac{1}{n}K = \frac{1}{n}\Phi\Phi^\top \tag{9}$$

where $\Phi \in \mathbb{R}^{n\times d}$ is an $n \times d$ matrix whose rows are the feature presentations of samples, i.e., $\phi(x_1), \ldots, \phi(x_n)$. Therefore, the normalized kernel matrix $\frac{1}{n}K$ shares the same non-zero eigenvalues with $\frac{1}{n}\Phi^\top\Phi$, where the multiplication order is flipped. Note that $\frac{1}{n}\Phi^\top\Phi$ is equal to the empirical kernel covariance matrix $\widehat{C}_X$ defined as:

$$\widehat{C}_X := \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)\phi(x_i)^\top = \frac{1}{n}\Phi^\top\Phi.$$

As a result, the empirical covariance matrix $\widehat{C}_X = \frac{1}{n}\Phi^\top\Phi$ and kernel matrix $\frac{1}{n}K = \frac{1}{n}\Phi\Phi^\top$ share the same non-zero eigenvalues and therefore have the same matrix-based entropy value for every order α: $H_\alpha(\frac{1}{n}K) = H_\alpha(\widehat{C}_X)$. Therefore, if we consider the population kernel covariance matrix $\widetilde{C}_X = \mathbb{E}_{x\sim P_X}\big[\phi(x)\phi(x)^\top\big]$, we can define the population Vendi score as follows.

Definition 1. *Given data distribution P_X, we define the order-α population Vendi, $\mathrm{Vendi}_\alpha(P_X)$, using the matrix-based entropy of the population kernel covariance matrix $\widetilde{C}_X = \mathbb{E}_{x\sim P_X}\big[\phi(x)\phi(x)^\top\big]$ as*

$$\mathrm{Vendi}_\alpha(P_X) := \exp\Big(H_\alpha(\widetilde{C}_X)\Big) \tag{10}$$

Note that the population RKE score is identical to the population Vendi_2, since RKE and Vendi_2 are the same.

4 STATISTICAL CONVERGENCE OF VENDI AND RKE SCORES

Given the definitions of the Vendi score and the population Vendi, a relevant question is how many samples are required to accurately estimate the population Vendi using the Vendi score. To address this question, we first prove the following concentration bound on the vector of ordered eigenvalues $[\lambda_1, \ldots, \lambda_n]$ of the kernel matrix for a normalized kernel function.

Theorem 1. *Consider a normalized kernel function k satisfying $k(x, x) = 1$ for every $x \in \mathcal{X}$. Let $\widehat{\boldsymbol{\lambda}}_n$ be the vector of sorted eigenvalues of the normalized kernel matrix $\frac{1}{n}K$ for n independent samples $x_1, \ldots, x_n \sim P_X$. If we define $\widetilde{\boldsymbol{\lambda}}$ as the vector of sorted eigenvalues of underlying covariance matrix $\widetilde{C}_X$, then if $n \geq 2 + 8\log(1/\delta)$, the following inequality holds with probability at least $1 - \delta$:*

$$\big\|\widehat{\boldsymbol{\lambda}}_n - \widetilde{\boldsymbol{\lambda}}\big\|_2 \leq \sqrt{\frac{32\log\big(2/\delta\big)}{n}}$$

Note that in calculating the subtraction $\widehat{\boldsymbol{\lambda}}_n - \widetilde{\boldsymbol{\lambda}}$, we add $|d-n|$ zero entries to the lower-dimension vector, if the dimension of vectors $\widehat{\boldsymbol{\lambda}}_n$ and $\widetilde{\boldsymbol{\lambda}}$ do not match.

Proof. We defer the proof to the Appendix.

Theorem 1 results in the following corollary on a *dimension-free convergence guarantee* for every Vendi_α score with order $\alpha \geq 2$, including the RKE score (i.e. Vendi_2).

Corollary 1. *In the setting of Theorem 1, for every $\alpha \geq 2$ and $n \geq 2 + 8\log(1/\delta)$, the following bound holds with probability at least $1-\delta$:*

$$\left|\mathrm{Vendi}_\alpha(x_1,\ldots,x_n)^{\frac{1-\alpha}{\alpha}} - \mathrm{Vendi}_\alpha(P_X)^{\frac{1-\alpha}{\alpha}}\right| \leq \sqrt{\frac{32\log\frac{2}{\delta}}{n}}$$

Notably, for $\alpha = 2$, we arrive at the following bound on the gap between the empirical and population RKE scores:

$$\left|\mathrm{RKE}(x_1,\ldots,x_n)^{-1/2} - \mathrm{RKE}(P_X)^{-1/2}\right| \leq \sqrt{\frac{32\log\frac{2}{\delta}}{n}}$$

Proof. We defer the proof to the Appendix.

Therefore, the bound in Corollary 1 holds regardless of the dimension of kernel feature map, indicating that the RKE score enjoys a universal convergence guarantee across all kernel functions. Next, we show that Theorem 1 implies the following corollary on a dimension-dependent convergence guarantee for order-α Vendi score with $1 \leq \alpha < 2$, including standard (order-1) Vendi score.

Corollary 2. *In the setting of Theorem 1, consider a finite dimension kernel map where we suppose $\dim(\phi) = d < \infty$.*
(a) For $\alpha = 1$, assuming $n \geq 32e^2\log(2/\delta)$, the following bound holds with probability at least $1-\delta$:

$$\left|\log\big(\mathrm{Vendi}_1(x_1,\ldots,x_n)\big) - \log\big(\mathrm{Vendi}_1(P_X)\big)\right| \leq \sqrt{\frac{8d\log(2/\delta)}{n}}\log\Big(\frac{nd}{32\log(2/\delta)}\Big).$$

(b) For every $1 < \alpha < 2$ and $n \geq 2 + 8\log(1/\delta)$, the following bound holds with probability at least $1-\delta$:

$$\left|\mathrm{Vendi}_\alpha(x_1,\ldots,x_n)^{\frac{1-\alpha}{\alpha}} - \mathrm{Vendi}_\alpha(P_X)^{\frac{1-\alpha}{\alpha}}\right| \leq \sqrt{\frac{32d^{2-\alpha}\log(2/\delta)}{n}}$$

Proof. We defer the proof to the Appendix.

Therefore, assuming a finite feature map $d < \infty$ and given an entropy order $1 \leq \alpha < 2$, the above results indicate the convergence of the Vendi score to the underlying population Vendi given $n = O(d^{2-\alpha})$ samples. Observe that this result is consistent with our numerical observations of the convergence of Vendi score using the finite-dimension cosine similarity kernel in Figure 1.

5 TRUNCATED VENDI SCORE AND ITS ESTIMATION VIA PROXY KERNELS

Corollaries 1, 2 demonstrate that if the Vendi score order α is greater than 2 or the kernel feature map dimension d is finite, then the Vendi score can converge to the population Vendi with $n = O(d)$ samples. However, the theoretical results do not apply to an order $1 \leq \alpha < 2$ when the kernel map dimension is infinite, e.g. the original order-1 Vendi score [Friedman and Dieng, 2023] with a Gaussian kernel. Our numerical observations indicate that a standard sample size below 20000 could be insufficient for the convergence of order-1 Vendi score (Figure 1). To address this gap, here we define the truncated Vendi score by truncating the eigenspectrum of the kernel matrix, and then show that the existing kernel approximation algorithms for Vendi score concentrate around this modified Vendi score.

Definition 2. *Consider the normalized kernel matrix $\frac{1}{n}K$ of samples $x_1,\ldots,x_n$. Then, for an integer parameter $t \geq 1$, consider the top-t eigenvalues of $\frac{1}{n}K$: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_t$. Define $S_t = \sum_{i=1}^t \lambda_i$ and consider the truncated probability sequence $[\lambda_1^{trunc},\ldots,\lambda_t^{trunc}]$:*

$$\lambda_i^{trunc} = \lambda_i + \frac{1-S_t}{t} \quad \textit{for } i = 1,\ldots,t$$

We define the order-α t-truncated Vendi score as

$$\mathrm{Vendi}_\alpha^{(t)}(x_1,\ldots,x_n) := \exp\Big(\frac{1}{1-\alpha}\log\Big(\sum_{i=1}^t \lambda_i^{trunc\,\alpha}\Big)\Big)$$

Notably, for order $\alpha = 1$, the t-truncated Vendi score is:

$$\mathrm{Vendi}_1^{(t)}(x_1,\ldots,x_n) := \exp\Big(\sum_{i=1}^t \lambda_i^{trunc}\log\frac{1}{\lambda_i^{trunc}}\Big)$$

Remark 1. *The above definition of t-truncated Vendi score leads to the definition of t-truncated population Vendi $\mathrm{Vendi}_\alpha^{(t)}(P_X)$, where the mentioned truncation process is applied to the eigenspectrum of the population kernel covariance matrix $\widetilde{C}_X$. Note that the truncated Vendi score is a statistic and function of random samples $x_1,\ldots,x_n$, whereas the truncated population Vendi is deterministic and a characteristic of the population distribution P_X.*

According to Definition 2, we find the probability model with the minimum ℓ_2-norm difference from the t-dimensional vector $[\lambda_1,\ldots,\lambda_t]$ including only the top-t eigenvalues. Then, we use the order-α entropy of the probability model to define the order-α t-truncated population Vendi. Our next result shows that this population quantity can be estimated using $n = O(t)$ samples by t-truncated Vendi score for every kernel function.

Theorem 2. *Consider the setting in Theorem 1. Then, for every $n \geq 2 + 8\log(1/\delta)$, the difference between the t-truncated population Vendi and the empirical t-truncated*

Vendi score of samples $x_1, \ldots, x_n$ is bounded with probability at least $1-\delta$:

$$\left| \mathrm{Vendi}_\alpha^{(t)}(x_1, \ldots, x_n)^{\frac{1-\alpha}{\alpha}} - \mathrm{Vendi}_\alpha^{(t)}(P_X)^{\frac{1-\alpha}{\alpha}} \right| \leq \sqrt{\frac{32 \max\{1, t^{2-\alpha}\} \log(2/\delta)}{n}}$$

Proof. We defer the proof to the Appendix.

As implied by Theorem 2, the t-truncated population Vendi can be estimated using $O(t)$ samples, i.e. the truncation parameter t plays the role of the bounded dimension of a finite-dimension kernel map. Our next theorem shows that the Nyström method [Friedman and Dieng, 2023] and the FKEA method [Ospanov et al., 2024b] for reducing the computational costs of Vendi scores have a bounded difference with the truncated population Vendi.

Theorem 3. *Consider the setting of Theorem 1. (a) Assume that the kernel function is shift-invariant and the FKEA method with t random Fourier features is used to approximate the Vendi score. Then, for every δ satisfying $n \geq 2 + 8\log(1/\delta)$, with probability at least $1-\delta$:*

$$\left| \text{FKEA-Vendi}_\alpha^{(t)}(x_1, \ldots, x_n)^{\frac{1-\alpha}{\alpha}} - \mathrm{Vendi}_\alpha^{(t)}(P_X)^{\frac{1-\alpha}{\alpha}} \right| \leq \sqrt{\frac{128 \max\{1, t^{2-\alpha}\} \log(3/\delta)}{\min\{n, t\}}}$$

(b) Assume that the Nyström method is applied with parameter t for approximating the kernel function. Then, if for some $r \geq 1$, the kernel matrix K's rth-largest eigenvalue satisfies $\lambda_r \leq \tau$ and $t \geq r\tau \log(n)$, the following holds with probability at least $1 - \delta - 2n^{-3}$:

$$\left| \text{Nystrom-Vendi}_\alpha^{(t)}(x_1, \ldots, x_n)^{\frac{1-\alpha}{\alpha}} - \mathrm{Vendi}_\alpha^{(t)}(P_X)^{\frac{1-\alpha}{\alpha}} \right| \leq \mathcal{O}\Big(\sqrt{\frac{\max\{1, t^{2-\alpha}\} \log(2/\delta) t\tau^2 \log(n)^2}{n}}\Big)$$

Proof. We defer the proof to the Appendix.

6 NUMERICAL RESULTS

We evaluated the convergence of the Vendi score, the truncated Vendi score, and the proxy Vendi scores using the Nyström method and FKEA in our numerical experiments. We provide a comparative analysis of these scores across different data types and models, including image, text, and video. In our experiments, we considered the cosine similarity kernel as a standard kernel function with a finite-dimension map and the Gaussian (RBF) kernel as a kernel function with an infinite-dimension feature map. In the experiments with Gaussian kernels, we matched the kernel bandwidth parameter with those chosen by [Jalali et al., 2023, Ospanov et al., 2024b] for the same datasets. We used 20,000 number of samples per score computation, consistent with standard practice in the literature. To investigate how computation-cutting methods compare to each other, in the experiments we matched the truncation parameter t of our defined t-truncated Vendi score with the Nyström method's hyperparameter on the number of randomly selected rows of kernel matrix and the FKEA's hyperparameter of the number of random Fourier features. The Vendi and FKEA implementations were adopted from the corresponding references' GitHub webpages, while the Nyström method was adopted from the `scikit-learn` Python package.

6.1 CONVERGENCE ANALYSIS OF VENDI SCORES

To assess the convergence of the discussed Vendi scores, we conducted experiments on four datasets including ImageNet and FFHQ [Karras et al., 2019] image datasets, a synthetic text dataset with 400k paragraphs generated by GPT-4 about 100 randomly selected countries, and the Kinetics video dataset [Kay et al., 2017]. Our results, presented in Figures 3, 4, and 5, show that for the finite-dimension cosine similarity kernel the Vendi score converges rapidly to the underlying value and the proxy versions including truncated and Nyström Vendi scores were almost identical to the original Vendi score. This observation is consistent with our theoretical results on the convergence of Vendi scores under finite-dimension kernel maps. On the other hand, in the case of infinite dimension Gaussian kernel, we observed that the Vendi_1 score did not converge using 20k samples and the score value kept growing with a considerable rate. However, the t-truncated Vendi score with $t = 10000$ converged to its underlying statistic shortly after 10000 samples were used. Consistent with our theoretical result, the proxy Nyström and FKEA estimated scores with their rank hyperparameter matched with t also converged to the limit of the truncated Vendi scores. The numerical results show the connection between the truncated Vendi score and the existing kernel methods for approximating the Vendi score.

6.2 CORRELATION BETWEEN THE TRUNCATED VENDI SCORE AND DIVERSITY OF DATA

We performed experiments to test the correlation between the truncated Vendi score and the ground-truth diversity of data. To do this, we applied the truncation technique to the FFHQ-based StyleGAN3 [Karras et al., 2021] model and the ImageNet-based StyleGAN-XL [Sauer et al., 2022] model and simulated generative models with different underlying diversity by varying the truncation technique. Considering the Gaussian kernel, we estimated the t-truncated Vendi score with $t = 10000$ by averaging the estimated t-truncated Vendi scores over 5 independent datasets of size 20k where the score seemed to converge to its under-

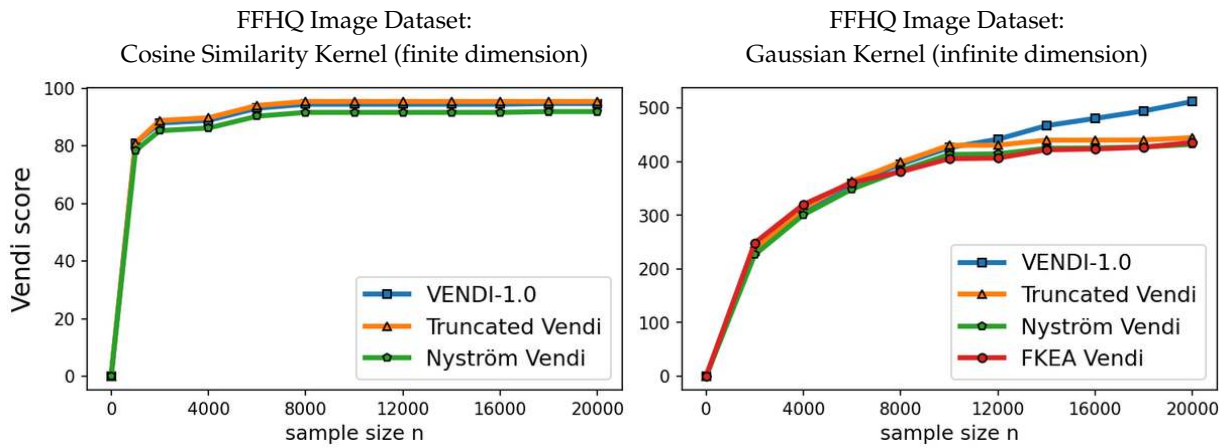

Figure 3: Statistical convergence of Vendi score for different sample sizes on FFHQ[Karras et al., 2019] data: (Left plot) finite-dimension cosine similarity kernel (Right plot) infinite dimension Gaussian kernel with bandwidth $\sigma = 35$. *DINOv2* embedding (dimension 768) is used in computing the scores.

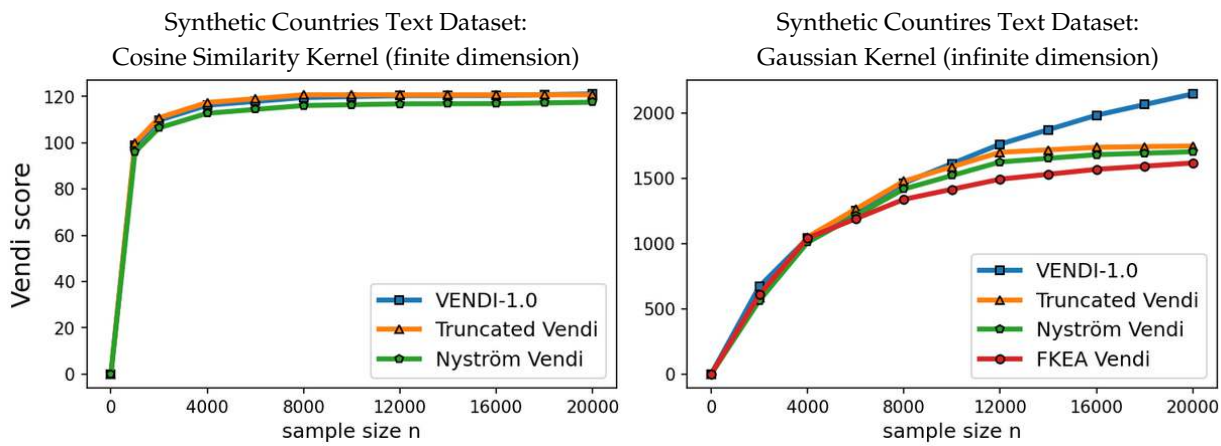

Figure 4: Statistical convergence of Vendi score for different sample sizes on Synthetic Countries data: (Left plot) finite-dimension cosine similarity kernel (Right plot) infinite dimension Gaussian kernel with bandwidth $\sigma = 0.6$. *text-embedding-3-large* embedding (dimension 3072) is used in computing the scores.

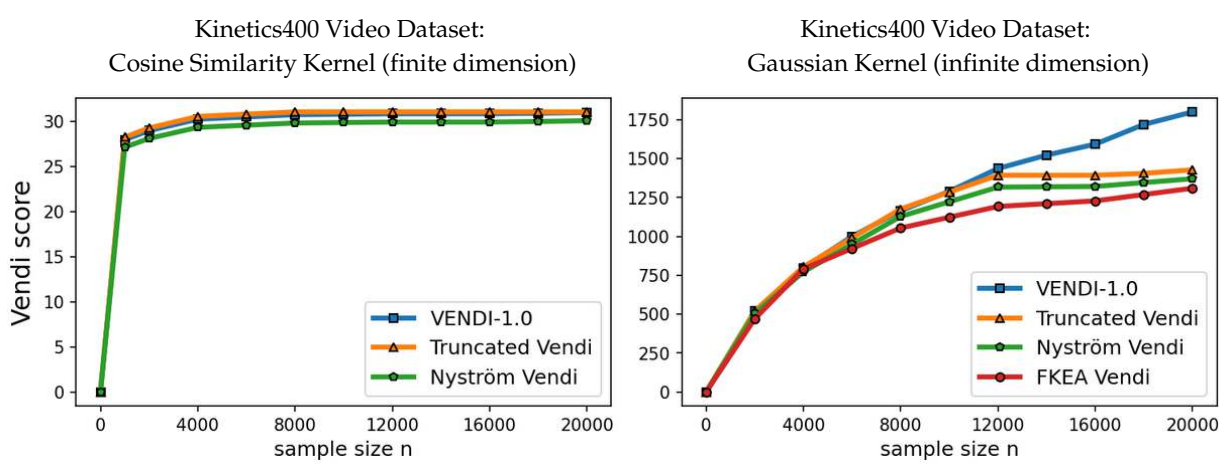

Figure 5: Statistical convergence of Vendi score for different sample sizes on Kinetics400[Kay et al., 2017] data: (Left plot) finite-dimension cosine similarity kernel (Right plot) infinite dimension Gaussian kernel with bandwidth $\sigma = 4.0$. *I3D* embedding (dimension 1024) is used in computing the scores.

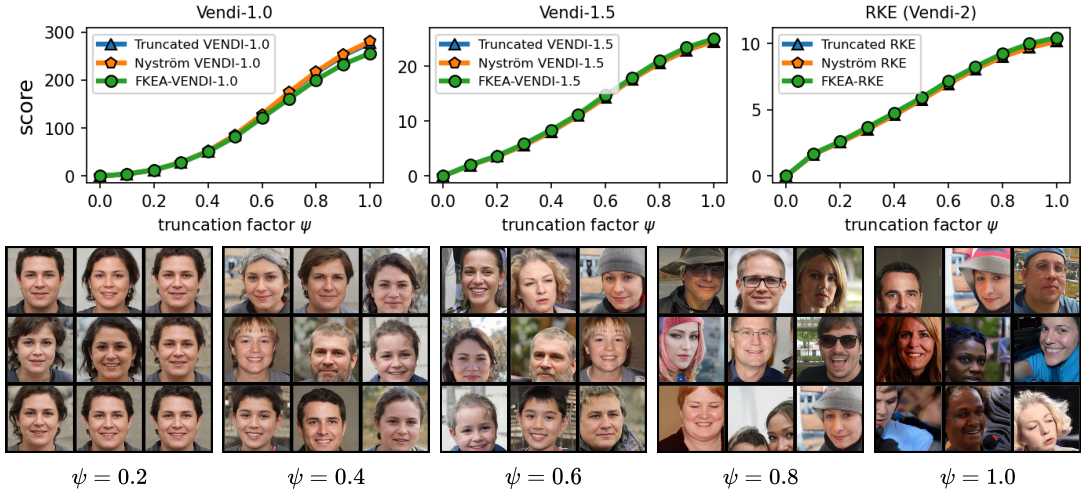

Figure 6: Diversity evaluation of Vendi scores on truncated StyleGAN3 generated FFHQ dataset with varying truncation coefficient ψ. Fixed sample size n =20k is used for estimating the scores.

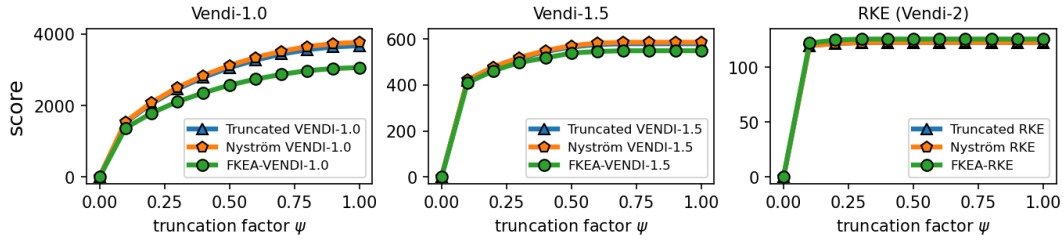

Figure 7: Diversity evaluation of Vendi scores on truncated StyleGAN-XL generated ImageNet dataset with varying truncation coefficient ψ. Fixed sample size n =20k is used for estimating the scores.

lying value. Figures 6, 7 show how the estimated statistic correlates with the truncation parameter for order-α Vendi scores with $\alpha = 1,\ 1.5, 2$. In all these experiments, the estimated truncated Vendi score correlated with the underlying diversity of the models. In addition, we plot the proxy Nyström and FKEA proxy Vendi values computed using 20000 samples which remain close to the estimated t-truncated statistic. These empirical results suggest that the estimated t-truncated Vendi score with Gaussian kernel can be used to evaluate the diversity of generated data. Also, the Nyström and FKEA methods were both computationally efficient in estimating the truncated Vendi score from limited generated data. We defer the presentation of the additional numerical results on the convergence of Vendi scores with different orders, kernel functions and embedding spaces to the Appendix.

7 CONCLUSION

In this work, we investigated the statistical convergence behavior of Vendi diversity scores estimated from empirical samples. We highlighted that, due to the high computational complexity of the score for datasets larger than a few tens of thousands of generated data points, the score is often calculated using sample sizes below 10,000. We demonstrated that such restricted sample sizes do not pose a problem for statistical convergence as long as the kernel feature dimension is bounded. However, our numerical results showed a lack of convergence to the population Vendi when using an infinite-dimensional kernel map, such as the Gaussian kernel. To address this gap, we introduced the truncated population Vendi as an alternative target quantity for diversity evaluation. We showed that existing Nyström and FKEA methods for approximating Vendi scores concentrate around this truncated population Vendi. An interesting future direction is to explore the relationship between other kernel approximation techniques and the truncated population Vendi. Also, a comprehensive analysis of the computational-statistical trade-offs involved in estimating the Vendi score is another relevant future direction.

Acknowledgements

This work is partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China, Project 14209920, and is partially supported by CUHK Direct Research Grants with CUHK Project No. 4055164 and 4937054. Finally, the authors thank the anonymous reviewers for their thoughtful feedback and constructive suggestions.

References

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950. doi: 10.2307/1990404.

Francis Bach. Information Theory with Kernel Methods, August 2022. URL http://arxiv.org/abs/2202.08545. arXiv:2202.08545 [cs, math, stat].

Francis R Bach and Michael I Jordan. Kernel independent component analysis. In *Journal of Machine Learning Research*, volume 3, pages 1–48, 2002.

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

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. URL https://proceedings.neurips.cc/paper/2020/hash/70feb62b69f16e0238f741fab228fec2-Abstract.html.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition (CVPR)*, pages 248–255. IEEE, 2009.

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

Shai Fine and Katya Scheinberg. Efficient svm training using low-rank kernel representations. In *Journal of Machine Learning Research (JMLR)*, pages 243–250, 2001.

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

Shizhan Gong, Yankai Jiang, Qi Dou, and Farzan Farnia. Kernel-based unsupervised embedding alignment for enhanced visual representation in vision-language models. *arXiv preprint arXiv:2506.02557*, 2025.

David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

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.

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

Xiaoyan Hu, Ho-fung Leung, and Farzan Farnia. A multi-armed bandit approach to online selection and evaluation of generative models. In *International Conference on Artificial Intelligence and Statistics*, pages 1864–1872. PMLR, 2025.

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.

Mohammad Jalali, Azim Ospanov, Amin Gohari, and Farzan Farnia. Conditional vendi score: An information-theoretic approach to diversity evaluation of prompt-based generative models. *arXiv preprint arXiv:2411.02817*, 2024.

Mohammad Jalali, Bahar Dibaei Nia, and Farzan Farnia. Towards an explainable comparison and alignment of feature embeddings. *arXiv preprint arXiv:2506.06231*, 2025.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019.

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., 2021.

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.

Jonas Moritz Kohler and Aurelien Lucchi. Sub-sampled cubic regularization for non-convex optimization. In *International Conference on Machine Learning*, pages 1895–1904. PMLR, 2017.

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.

Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The Role of ImageNet Classes in Fréchet Inception Distance. September 2022. URL https://openreview.net/forum?id=4oXTQ6m_ws8.

Michael W. Mahoney and Petros Drineas. Cur matrix decompositions for improved data analysis. In *Proceedings of the National Academy of Sciences*, volume 106, pages 697–702, 2009.

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.

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. URL https://openreview.net/forum?id=a68SUt6zFt.

Azim Ospanov, Mohammad Jalali, and Farzan Farnia. Dissecting clip: Decomposition with a schur complement-based approach. *arXiv preprint arXiv:2412.18645*, 2024a.

Azim Ospanov, Jingwei Zhang, Mohammad Jalali, Xuenan Cao, Andrej Bogdanov, and Farzan Farnia. Towards a scalable reference-free evaluation of generative models. In *Advances in Neural Information Processing Systems*, volume 38, 2024b.

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.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. MAUVE: Measuring the gap between neural text and human text using divergence frontiers. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=Tqx7nJp7PR.

Krishna Pillutla, Lang Liu, John Thickstun, Sean Welleck, Swabha Swayamdipta, Rowan Zellers, Sewoong Oh, Yejin Choi, and Zaid Harchaoui. MAUVE Scores for Generative Models: Theory and Practice. *JMLR*, 2023.

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. doi: 10.48550/arXiv.2103.00020. URL http://arxiv.org/abs/2103.00020. arXiv:2103.00020 [cs].

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

Alfréd Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 547–561. University of California Press, 1961.

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.

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. URL https://arxiv.org/abs/2201.00273.

J. Shawe-Taylor, C.K.I. Williams, N. Cristianini, and J. Kandola. On the eigenspectrum of the gram matrix and the generalization error of kernel-pca. *IEEE Transactions on Information Theory*, 51(7):2510–2522, 2005. doi: 10.1109/TIT.2005.850052.

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.

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)*, 2016.

Zixiao Wang, Farzan Farnia, Zhenghao Lin, Yunheng Shen, and Bei Yu. On the distributed evaluation of generative models. *arXiv preprint arXiv:2310.11714*, 2023.

Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In *Advances in neural information processing systems*, pages 682–688, 2000.

Youqi Wu, Jingwei Zhang, and Farzan Farnia. Fusing cross-modal and uni-modal representations: A kronecker product approach, 2025. URL https://arxiv.org/abs/2506.08645.

Zenglin Xu, Rong Jin, Bin Shen, and Shenghuo Zhu. Nystrom approximation for sparse kernel methods: Theoretical analysis and empirical evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Jingwei Zhang, Cheuk Ting Li, and Farzan Farnia. An interpretable evaluation of entropy-based novelty of generative models. *arXiv preprint arXiv:2402.17287*, 2024.

Jingwei Zhang, Mohammad Jalali, Cheuk Ting Li, and Farzan Farnia. Unveiling differences in generative models: A scalable differential clustering approach. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8269–8278, 2025.

Supplementary Material

Azim Ospanov[1] **Farzan Farnia**[1]

[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong

A PROOFS

A.1 PROOF OF THEOREM 1

To prove the theorem, we will use the following lemma followed from [Gross, 2011, Kohler and Lucchi, 2017].

Lemma 1 (Vector Bernstein Inequality [Gross, 2011, Kohler and Lucchi, 2017])**.** *Suppose that $z_1, \ldots, z_n$ are independent and identically distributed random vectors with zero mean $\mathbb{E}[z_i] = \mathbf{0}$ and bounded ℓ_2-norm $\|z_i\|_2 \leq c$. Then, for every $0 \leq \epsilon \leq c$, the following holds*

$$\mathbb{P}\Big(\Big\| \frac{1}{n} \sum_{i=1}^{n} z_i \Big\|_2 \geq \epsilon \Big) \leq \exp\Big(-\frac{n\epsilon^2}{8c^2} + \frac{1}{4} \Big)$$

We apply the above Vector Bernstein Inequality to the random vectors $\phi(x_1) \otimes \phi(x_1), \ldots, \phi(x_1) \otimes \phi(x_1)$ where $\otimes$ denotes the Kronecker product. To do this, we define vector $v_i = \phi(x_i) \otimes \phi(x_i) - \mathbb{E}_{x \sim P}\big[\phi(x) \otimes \phi(x)\big]$ for every i. Note that v_i is, by definition, a zero-mean vector and also for every x we have the following for the normalized kernel function k:

$$\big\|\phi(x) \otimes \phi(x)\big\|_2^2 = \big\|\phi(x)\big\|_2^2 \cdot \big\|\phi(x)\big\|_2^2 = k(x,x) \cdot k(x,x) = 1$$

Then, the triangle inequality implies that

$$\big\|v_i\big\|_2 \leq \big\|\phi(x_i) \otimes \phi(x_i)\big\|_2 + \big\|\mathbb{E}_{x \sim P}\big[\phi(x) \otimes \phi(x)\big]\big\|_2 \leq \big\|\phi(x_i) \otimes \phi(x_i)\big\|_2 + \mathbb{E}_{x \sim P}\big[\big\|\phi(x) \otimes \phi(x)\big\|_2\big] = 2$$

As a result, the Vector Bernstein Inequality leads to the following for every $0 \leq \epsilon \leq 2$:

$$\mathbb{P}\Big(\Big\| \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) \otimes \phi(x_i) - \mathbb{E}_{x \sim P}\big[\phi(x) \otimes \phi(x)\big] \Big\|_2 \geq \epsilon \Big) \leq \exp\Big(\frac{8 - n\epsilon^2}{32} \Big)$$

On the other hand, note that $\phi(x) \otimes \phi(x)$ is the vectorized version of rank-1 $\phi(x)\phi(x)^\top$, which shows that the above inequality is equivalent to the following where $\|\cdot\|_{\mathrm{HS}}$ denotes the Hilbert-Schmidt norm, which will simplify to the Frobenius norm in the finite dimension case,

$$\begin{aligned} &\mathbb{P}\Big(\Big\| \frac{1}{n} \sum_{i=1}^{n} \big[\phi(x_i)\phi(x_i)^\top\big] - \mathbb{E}_{x \sim P}\big[\phi(x)\phi(x)^\top\big] \Big\|_{\mathrm{HS}} \geq \epsilon \Big) \leq \exp\Big(\frac{8 - n\epsilon^2}{32} \Big) \\ \implies &\mathbb{P}\Big(\Big\| C_X - \widetilde{C}_X \Big\|_{\mathrm{HS}} \geq \epsilon \Big) \leq \exp\Big(\frac{8 - n\epsilon^2}{32} \Big) \end{aligned}$$

Subsequently, we can apply the Hoffman-Wielandt inequality which shows that for the sorted eigenvalue vectors of C_X (denoted by $\widehat{\boldsymbol{\lambda}}_n$ in the theorem) and $\widetilde{C}_X$ (denoted by $\widetilde{\boldsymbol{\lambda}}$ in the theorem) we will have $\|\widehat{\boldsymbol{\lambda}}_n - \widetilde{\boldsymbol{\lambda}}\|_2 \leq \|C_X - \widetilde{C}_X\|_{\mathrm{HS}}$, which together with the previous inequality leads to

$$\mathbb{P}\Big(\big\|\widehat{\boldsymbol{\lambda}}_n - \widetilde{\boldsymbol{\lambda}}\big\|_2 \geq \epsilon\Big) \;\leq\; \exp\Big(\frac{8 - n\epsilon^2}{32}\Big)$$

If we define $\delta = \exp\big((8 - n\epsilon^2)/32\big)$ that implies $\epsilon \leq \sqrt{\frac{32\log(2/\delta)}{n}}$, we obtain the following for every $\delta \geq \exp\big((2-n)/8\big)$ (since we suppose $0 \leq \epsilon \leq 2$)

$$\mathbb{P}\Big(\big\|\widehat{\boldsymbol{\lambda}}_n - \widetilde{\boldsymbol{\lambda}}\big\|_2 \geq \sqrt{\frac{32\log(2/\delta)}{n}}\Big) \;\leq\; \delta$$
$$\implies \mathbb{P}\Big(\big\|\widehat{\boldsymbol{\lambda}}_n - \widetilde{\boldsymbol{\lambda}}\big\|_2 \leq \sqrt{\frac{32\log(2/\delta)}{n}}\Big) \;\geq\; 1 - \delta$$

which completes the proof.

A.2 PROOF OF COROLLARY 2

The case of $\alpha = 1$. We show that Theorem 1 on the concentration of the eigenvalues $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_d]$ will further imply a concentration bound for the logarithm of Vendi-1 score. In the case of Vendi_1 (when $\alpha \to 1^+$), the concentration bound will be formed for the logarithm of the Vendi score, i.e. the Von-Neumann entropy (denoted as H_α):

$$H_1(C_X) := H_1(\boldsymbol{\lambda}) = \sum_{i=1}^{d} \widetilde{\lambda}_i \log \frac{1}{\widetilde{\lambda}_i}$$

Theorem 1 shows that $\|\widehat{\boldsymbol{\lambda}}_n - \widetilde{\boldsymbol{\lambda}}\|_2 \leq \sqrt{\frac{32\log(2/\delta)}{n}}$ with probability $1-\delta$. To convert this concentration bound to a bound on the order-1 entropy (for Vendi-1 score) difference $H_1(\widehat{C}_n) - H_1(C_X)$, we leverage the following two lemmas:

Lemma 2. *For every* $0 \leq \alpha, \beta \leq 1$ *such that* $|\beta - \alpha| \leq \frac{1}{e}$, *we have*

$$\Big|\alpha \log \frac{1}{\alpha} - \beta \log \frac{1}{\beta}\Big| \;\leq\; |\beta - \alpha| \log \frac{1}{|\beta - \alpha|}$$

Proof. Let $c = |\alpha - \beta|$, where $c \in [0, \frac{1}{e}]$. Defining $g(z) = z\log(\frac{1}{z})$, the first-order optimality condition $g'(z) = -\log(z) - 1 = 0$ yields $\frac{1}{e}$ as the local maximum of $g(z)$. Therefore, there are three cases of placement of α and β on the interval $[0, 1]$: α and β appear before maximum point, after maximum point or maximum point is between α and β. We show that regardless of the placement of α and β, the above inequality remains true.

- **Case 1:** $\alpha, \beta \in [0, \frac{1}{e}]$. Note that $g''(z) = -\frac{1}{z}$. Since the second-order derivative is negative and the function g is monotonically increasing within the interval $[0, \frac{1}{e}]$, the gap between $g(\alpha)$ and $g(\beta)$ is maximized when $\alpha^* = 0$ and $\beta^* = c - \alpha^* = c$. This directly leads to the desired bound as follows:

  $$\Big|\alpha \log \frac{1}{\alpha} - \beta \log \frac{1}{\beta}\Big| \;\leq\; \Big|\alpha^* \log \frac{1}{\alpha^*} - \beta^* \log \frac{1}{\beta^*}\Big| \;=\; \Big|0 \log 0 - c \log \frac{1}{c}\Big| \;\leq\; c \log \frac{1}{c}$$

  Here, we use the standard limit $0 \log 0 = 0$.

- **Case 2:** $\alpha, \beta \in [\frac{1}{e}, 1]$. In this case, we note that g is concave yet decreasing over $[\frac{1}{e}, 1]$, and so the gap between $g(\alpha)$ and $g(\beta)$ will be maximized when $\alpha^* = 1 - c$ and $\beta^* = 1$. This leads to:

  $$\Big|\alpha \log \frac{1}{\alpha} - \beta \log \frac{1}{\beta}\Big| \;\leq\; \Big|\alpha^* \log \frac{1}{\alpha^*} - \beta^* \log \frac{1}{\beta^*}\Big| \;=\; (1-c) \log \frac{1}{(1-c)} \leq c \log \frac{1}{c}$$

  where the last inequality holds because $c \in [0, \frac{1}{e}]$, and if we define the function $h(c) = c \log \frac{1}{c} - (1-c) \log \frac{1}{1-c}$, then we have $h'(c) = \log \frac{1}{c(1-c)} - 2$, which is positive over $c \in [0, c_0]$ ($e^{-2} < c_0 < e^{-1}$ is where $c_0(1-c_0) = e^{-2}$), and then negative over $[c_0, \frac{1}{e}]$, and hence $h(c) \geq \min\{h(0), h(1/e)\} = 0$ for every $c \in [0, 1/e]$.

- **Case 3:** $\alpha \in [0, \frac{1}{e})$ and $\beta \in (\frac{1}{e}, 1]$. When α and β lie on the opposite ends from the maximum point, the inequality becomes:

$$\big|\alpha \log \frac{1}{\alpha} - \beta \log \frac{1}{\beta}\big| \leq \max\Big\{\big|(1/e)\log \frac{1}{1/e} - \beta \log \frac{1}{\beta}\big|, \big|\alpha \log \frac{1}{\alpha} - (1/e)\log \frac{1}{1/e}\big|\Big\} \leq c \log \frac{1}{c}$$

since we pick the side with the largest difference, this difference is upper bounded by either Case 1 or Case 2 because $\max\{|\frac{1}{e} - \beta|, |\alpha - \frac{1}{e}|\} < c$. Therefore, this case is upper-bounded by $c \log \frac{1}{c}$.

All the three cases of placement of α and β are upper-bounded by $c \log \frac{1}{c}$; Therefore, the claim holds.

Lemma 3. *If $\|\mathbf{u}\|_2 \leq \epsilon$ for d-dimensional vector $\mathbf{u} \geq \mathbf{0}$ where $\epsilon \leq \frac{1}{e}$, then we have*

$$\sum_{i=1}^{d} u_i \log \frac{1}{u_i} \leq \epsilon\sqrt{d} \log \frac{\sqrt{d}}{\epsilon}$$

Proof. We prove the above inequality using the KKT conditions for the following maximization problem, representing a convex optimization problem,

$$\begin{aligned} \max_{\mathbf{u}\in\mathbb{R}^d} \quad & \sum_{i=1}^{d} u_i \log(\frac{1}{u_i}) \\ \text{subject to} \quad & u_i \geq 0, \quad \text{for all } i \\ & \sum_{i=1}^{d} u_i^2 \leq \epsilon^2 \quad (\text{equivalent to } \|\mathbf{u}\|_2 \leq \epsilon) \end{aligned}$$

In a concave maximization problem subject to convex constraints, any point that satisfies the KKT conditions is guaranteed to be a global optimum. Let us pick the following solution $\mathbf{u}^* = \frac{\epsilon}{\sqrt{d}}\mathbf{1}$ and slack variables $\lambda^* = \frac{\sqrt{d}}{2\epsilon}\big(\log(\frac{\sqrt{d}}{\epsilon}) - 1\big), \forall_i\ \mu_i^* = 0$. The Lagrangian of the above problem:

$$L(\mathbf{u}, \lambda, \mu_1, \ldots, \mu_d) = \sum_{i=1}^{d} u_i \log(\frac{1}{u_i}) + \lambda(\epsilon^2 - \sum_{i=1}^{d} u_i^2) - \sum_{i=1}^{d} \mu_i u_i$$

- **Primal Feasibility.** The solution $\mathbf{u}^*$ satisfies the primal feasibility, since $\epsilon^2 - \sum_{i=1}^{d}(\frac{\epsilon}{\sqrt{d}})^2 = \epsilon^2 - d\frac{\epsilon^2}{d} = 0$ and $\frac{\epsilon}{\sqrt{d}} \geq 0$.
- **Dual Feasibility.** $\lambda^* \geq 0$ is feasible because of the assumption $\epsilon \leq \frac{1}{e}$ implying that $\frac{\sqrt{d}}{\epsilon} \geq e$ for every integer dimension $d \geq 1$. Note that this implies $\lambda^* = \frac{\sqrt{d}}{2\epsilon}\big(\log(\frac{\sqrt{d}}{\epsilon}) - 1\big) \geq 0$.
- **Complementary Slackness.** Since $\lambda^*\big(\epsilon^2 - \sum_{i=1}^{d}(\frac{\epsilon}{\sqrt{d}})^2\big) = \lambda^* \cdot 0 = 0$, the condition is satisfied.
- **Stationarity.** The condition is satisfied as follows:

$$\frac{\partial}{\partial u_i} L(\mathbf{u}^*) = -\log(u_i^*) - 1 - 2\lambda^* u_i^* + \mu_i^* = -\log(\frac{\epsilon}{\sqrt{d}}) - 1 - 2 \cdot \frac{\sqrt{d}}{2\epsilon}\big(-\log(\frac{\epsilon}{\sqrt{d}}) - 1\big) \cdot \frac{\epsilon}{\sqrt{d}} = 0$$

Since all KKT conditions are satisfied and sufficient for global optimality, $\mathbf{u}^* = \frac{\epsilon}{\sqrt{d}}\mathbf{1}$ is a global optimum of the specified concave maximization problem. We note that this result is also implied by the Schur-concavity property of entropy. Following this result, the specified objective is upper-bounded as follows:

$$\sum_{i=1}^{d} u_i \log \frac{1}{u_i} \leq \epsilon\sqrt{d} \log \frac{\sqrt{d}}{\epsilon}$$

Therefore, the lemma's proof is complete.

Following the above lemmas, knowing that $\|\widehat{\lambda}_n - \widetilde{\lambda}\|_2 \leq \sqrt{\frac{32\log(2/\delta)}{n}}$ from Theorem 1 and using the assumption $n \geq 32e^2\log(2/\delta) \approx 236.5\log(2/\delta)$ that ensures the upper-bound satisfies $\sqrt{\frac{32\log(2/\delta)}{n}} \leq \frac{1}{e}$, we can apply the above two lemmas to show that with probability $1-\delta$:

$$\Big|H_1(\widehat{C}_n) - H_1(C_X)\Big| = \Big|H_1(\widehat{\lambda}_n) - H_1(\widetilde{\lambda})\Big| \leq \sqrt{\frac{8d\log(2/\delta)}{n}}\log\Big(\frac{nd}{32\log(2/\delta)}\Big)$$

Note that under a kernel function with finite dimension d, the above bound will be $\mathcal{O}\Big(\sqrt{\frac{d}{n}}\log\big(nd\big)\Big)$.

The case of $1 < \alpha < 2$**.** Note that the inequality $\|v\|_\alpha \leq d^{\frac{2-\alpha}{2}}\|v\|_2$ holds for every d-dimensional vector $v \in \mathbb{R}^d$. Therefore, we can repeat the proof of Corollary 1 to show the following for every $1 < \alpha < 2$

$$\begin{aligned}\Big|\mathrm{Vendi}_\alpha(x_1,\ldots,x_n)^{\frac{1-\alpha}{\alpha}} - \mathrm{Vendi}_\alpha(P_x)^{\frac{1-\alpha}{\alpha}}\Big| &= \Big|\big\|\widehat{\boldsymbol{\lambda}}_n\big\|_\alpha - \big\|\widetilde{\boldsymbol{\lambda}}\big\|_\alpha\Big| \\ &\leq \big\|\widehat{\boldsymbol{\lambda}}_n - \widetilde{\boldsymbol{\lambda}}\big\|_\alpha \\ &\leq d^{\frac{2-\alpha}{2}}\big\|\widehat{\boldsymbol{\lambda}}_n - \widetilde{\boldsymbol{\lambda}}\big\|_2.\end{aligned}$$

Consequently, Theorem 1 impies that for every $1 \leq \alpha < 2$ and $\delta \geq \exp((2-n)/8)$, the following holds with probability at least $1-\delta$

$$\Big|\mathrm{Vendi}_\alpha(x_1,\ldots,x_n)^{\frac{1-\alpha}{\alpha}} - \mathrm{Vendi}_\alpha(P_x)^{\frac{1-\alpha}{\alpha}}\Big| \leq d^{\frac{2-\alpha}{2}}\sqrt{\frac{32\log(2/\delta)}{n}} = \sqrt{\frac{32d^{2-\alpha}\log(2/\delta)}{n}}$$

A.3 PROOF OF COROLLARY 1

Considering the α-norm definition $\|\mathbf{v}\|_\alpha = \big(\sum_{i=1}^d |v_i|^\alpha\big)^{1/\alpha}$, we can rewrite the order-α Vendi definition as

$$\mathrm{Vendi}_\alpha(x_1,\ldots,x_n) = \big\|\widehat{\boldsymbol{\lambda}}_n\big\|_\alpha^{\frac{\alpha}{1-\alpha}} \quad\Longleftrightarrow\quad \mathrm{Vendi}_\alpha(x_1,\ldots,x_n)^{\frac{1-\alpha}{\alpha}} = \big\|\widehat{\boldsymbol{\lambda}}_n\big\|_\alpha$$

where $\widehat{\boldsymbol{\lambda}}_n$ is defined in Theorem 1. Similarly, given the definition of $\widetilde{\boldsymbol{\lambda}}$ we can write

$$\mathrm{Vendi}_\alpha(P_x)^{\frac{1-\alpha}{\alpha}} = \big\|\widetilde{\boldsymbol{\lambda}}\big\|_\alpha$$

Therefore, for every $\alpha \geq 2$, the following hold due to the triangle inequality:

$$\begin{aligned}\Big|\mathrm{Vendi}_\alpha(x_1,\ldots,x_n)^{\frac{1-\alpha}{\alpha}} - \mathrm{Vendi}_\alpha(P_x)^{\frac{1-\alpha}{\alpha}}\Big| &= \Big|\big\|\widehat{\boldsymbol{\lambda}}_n\big\|_\alpha - \big\|\widetilde{\boldsymbol{\lambda}}\big\|_\alpha\Big| \\ &\leq \big\|\widehat{\boldsymbol{\lambda}}_n - \widetilde{\boldsymbol{\lambda}}\big\|_\alpha \\ &\leq \big\|\widehat{\boldsymbol{\lambda}}_n - \widetilde{\boldsymbol{\lambda}}\big\|_2.\end{aligned}$$

As a result, Theorem 1 shows that for every $\alpha \geq 2$ and $\delta \geq \exp((2-n)/8)$, the following holds with probability at least $1-\delta$

$$\Big|\mathrm{Vendi}_\alpha(x_1,\ldots,x_n)^{\frac{1-\alpha}{\alpha}} - \mathrm{Vendi}_\alpha(P_x)^{\frac{1-\alpha}{\alpha}}\Big| \leq \sqrt{\frac{32\log(2/\delta)}{n}}$$

A.4 PROOF OF THEOREM 2

We begin by proving the following lemma showing that the eigenvalues used in the definition of the t-truncated Vendi score are the projection of the original eigenvalues onto a t-dimensional probability simplex.

Lemma 4. *Consider* $\mathbf{v} \in [0,1]^d$ *that satisfies* $\mathbf{1}^\top\mathbf{v} = 1$*. i.e., the sum of* $\mathbf{v}$*'s entries equals* 1*. Given integer* $1 \leq t \leq d$*, define vector* $\mathbf{v}^{(t)} \in [0,1]^d$ *whose last* $d-t$ *entries are* 0*, i.e.,* $v_i^{(t)} = 0$ *for* $t+1 \leq i \leq d$*, and its first* t *entries are defined as* $v_j^{(t)} = v_j + \frac{1-S_t}{t}$ *where* $S_t = v_1 + \cdots + v_t$*. Then,* $\mathbf{v}^{(t)}$ *is the projection of* $\mathbf{v}$ *onto the following simplex set and has the minimum* ℓ_2*-norm distance to this set*

$$\Delta_t := \Big\{\mathbf{u} \in [0,1]^d : \ v_i = 0 \text{ for all } t+1 \leq i \leq d, \quad \sum_{i=1}^t v_i = 1\Big\}.$$

Proof. To prove the lemma, first note that $\mathbf{v}^{(t)} \in \Delta_t$, i.e. its first t entries are non-negative and add up to 1, and also its last $d-t$ entries are zero. Then, consider the projection problem discussed in the lemma:

$$\begin{aligned} \min_{\mathbf{u}\in\mathbb{R}^t} \quad & \sum_{i=1}^{t} \big(u_i - v_i\big)^2 \\ \text{subject to} \quad & u_i \geq 0, \quad \text{for all } i \\ & \sum_{i=1}^{t} u_i = 1 \end{aligned}$$

Then, since we know from the assumptions that $v_i \geq 0$ and $\sum_{i=1}^{t} v_i \leq 1$, the discussed $\mathbf{u}^* \in \mathbb{R}^t$ where $u_i^* = v_i + (1-S_t)/t$ together with Lagrangian coefficients $\mu_i = 0$ (for inequality constraint $u_i \geq 0$) and $\lambda = (1-S_t)/t$ (for equality constraint) satisfy the KKT conditions. The primal and dual feasibility conditions as well as the complementary slackness clearly hold for these selection of primal and dual variables. Also, the KKT stationarity condition is satisfied as for every i we have $u_i^* - v_i - \lambda - \mu_i = 0$. Since the optimization problem is a convex optimization task with affine constraints, the KKT conditions are sufficient for optimaility which proves the lemma.

Based on the above lemma, the eigenvalues $\widehat{\boldsymbol{\lambda}}_n^{(t)}$ used to calculate the t-truncated Vendi score $\mathrm{Vendi}_\alpha^{(t)}(x_1,\ldots,x_n)$ are the projections of the top-t eigenvalues in $\widehat{\boldsymbol{\lambda}}_n$ for the original score $\mathrm{Vendi}_\alpha(x_1,\ldots,x_n)$ onto the t-simplex subset of $\mathbb{R}^d$ according to the ℓ_2-norm. Similarly, the eigenvalues $\widetilde{\boldsymbol{\lambda}}_n^{(t)}$ used to calculate the t-truncated population Vendi $\mathrm{Vendi}_\alpha^{(t)}(P_X)$ are the projections of the top-t eigenvalues in $\widetilde{\boldsymbol{\lambda}}$ for the original population Vendi $\mathrm{Vendi}_\alpha(P_x)$ onto the t-simplex subset of $\mathbb{R}^d$.

Since ℓ_2-norm is a Hilbert space norm and the t-simplex subset Δ_t is a convex set, we know from the convex analysis that the ℓ_2-distance between the projected points $\widehat{\boldsymbol{\lambda}}_n^{(t)}$ and $\widetilde{\boldsymbol{\lambda}}^{(t)}$ is upper-bounded by the ℓ_2-distance between the original points $\widehat{\boldsymbol{\lambda}}_n$ and $\widetilde{\boldsymbol{\lambda}}$. As a result, Theorem 1 implies that

$$\begin{aligned} & \mathbb{P}\Big(\big\|\widehat{\boldsymbol{\lambda}}_n - \widetilde{\boldsymbol{\lambda}}\big\|_2 \leq \sqrt{\frac{32\log(2/\delta)}{n}}\Big) \;\geq\; 1-\delta \\ \implies & \mathbb{P}\Big(\big\|\widehat{\boldsymbol{\lambda}}_n^{(t)} - \widetilde{\boldsymbol{\lambda}}^{(t)}\big\|_2 \leq \sqrt{\frac{32\log(2/\delta)}{n}}\Big) \;\geq\; 1-\delta \end{aligned}$$

However, note that the eigenvalue vectors $\widehat{\boldsymbol{\lambda}}_n^{(t)}$ and $\widetilde{\boldsymbol{\lambda}}^{(t)}$ can be analyzed in a bounded t-dimensional space as their entries after index $t+1$ are zero. Therefore, we can apply the proof of Corollary 2 to show that for every $1 \leq \alpha < 2$ and $\delta \geq \exp((2-n)/8)$, the following holds with probability at least $1-\delta$

$$\Big|\mathrm{Vendi}_\alpha(x_1,\ldots,x_n)^{\frac{1-\alpha}{\alpha}} - \mathrm{Vendi}_\alpha(P_x)^{\frac{1-\alpha}{\alpha}}\Big| \;\leq\; \sqrt{\frac{32t^{2-\alpha}\log(2/\delta)}{n}}$$

To extend the result to a general $\alpha > 1$, we reach the following inequality covering the above result as well as the result of Corollary 1 in one inequality

$$\Big|\mathrm{Vendi}_\alpha(x_1,\ldots,x_n)^{\frac{1-\alpha}{\alpha}} - \mathrm{Vendi}_\alpha(P_x)^{\frac{1-\alpha}{\alpha}}\Big| \;\leq\; \sqrt{\frac{32\max\{1,t^{2-\alpha}\}\log(2/\delta)}{n}}$$

A.5 PROOF OF THEOREM 3

Proof of Part (a). As defined by Ospanov et al. [2024b], the FKEA method uses the eigenvalues of t random Fourier frequencies $\omega_1,\ldots,\omega_t$ where for each ω_i they consider two features $\cos(\omega_i^\top x)$ and $\sin(\omega_i^\top x)$. Following the definitions, it can be seen that $k(x,x') = \mathbb{E}_{\omega\sim p_\omega}\big[\cos(\omega^\top(x-x'))\big]$ which is approximated by FKEA as $\frac{1}{t}\sum_{i=1}^{t}\cos(\omega_i^\top(x-x'))$. Therefore, if we define kernel matrix K_i as the kernel matrix for $k_i(x,x') = \cos(\omega_i^\top(x-x'))$, then we will have

$$\frac{1}{n}K^{\mathrm{FKEA}(t)} = \frac{1}{t}\sum_{i=1}^{t}\frac{1}{n}K_i$$

where $\mathbb{E}_{\omega_i \sim p_\omega}[\frac{1}{n}K_i] = \frac{1}{n}K$.

On the other hand, we note that $\|\frac{1}{n}K\|_{\mathrm{HS}} \leq 1$ holds as the kernel function is normalized and hence $|k(x, x')| \leq 1$. Since the Frobenius norm is the ℓ_2-norm of the vectorized version of the matrix, we can apply Vector Bernstein inequality in Lemma 1 to show that for every $0 \leq \epsilon \leq 2$:

$$\mathbb{P}\Big(\Big\|\frac{1}{t}\sum_{i=1}^{t}[\frac{1}{n}K_i] - \frac{1}{n}K\Big\|_F \geq \epsilon\Big) \leq \exp\Big(\frac{8 - t\epsilon^2}{32}\Big)$$

$$\implies \mathbb{P}\Big(\Big\|\frac{1}{n}K^{\mathrm{FKEA}(t)} - \frac{1}{n}K\Big\|_F \geq \epsilon\Big) \leq \exp\Big(\frac{8 - t\epsilon^2}{32}\Big)$$

Then, we apply the Hoffman-Wielandt inequality to show that for the sorted eigenvalue vectors of $\frac{1}{n}K$ (denoted by $\widehat{\boldsymbol{\lambda}}_n$) and $\frac{1}{n}K^{\mathrm{FKEA}(t)}$ (denoted by $\boldsymbol{\lambda}^{\mathrm{FKEA}(t)}$) we will have $\|\widehat{\boldsymbol{\lambda}}_n - \boldsymbol{\lambda}^{\mathrm{FKEA}(t)}\|_2 \leq \|\frac{1}{n}K^{\mathrm{FKEA}(t)} - \frac{1}{n}K\|_{\mathrm{HS}}$, which together with the previous inequality leads to

$$\mathbb{P}\Big(\Big\|\widehat{\boldsymbol{\lambda}}_n - \boldsymbol{\lambda}^{\mathrm{FKEA}(t)}\Big\|_2 \geq \epsilon\Big) \leq \exp\Big(\frac{8 - t\epsilon^2}{32}\Big)$$

Furthermore, as we shown in the proof of Theorem 1 for every $0 \leq \gamma \leq 2$

$$\mathbb{P}\Big(\big\|\widehat{\boldsymbol{\lambda}}_n - \widetilde{\boldsymbol{\lambda}}\big\|_2 \geq \gamma\Big) \leq \exp\Big(\frac{8 - n\gamma^2}{32}\Big)$$

which, by applying the union bound for $\gamma = \epsilon$, together with the previous inequality shows that

$$\begin{aligned}\mathbb{P}\Big(\Big\|\widetilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}^{\mathrm{FKEA}(t)}\Big\|_2 \geq 2\epsilon\Big) &\leq \exp\Big(\frac{8 - t\epsilon^2}{32}\Big) + \exp\Big(\frac{8 - n\epsilon^2}{32}\Big) \\ &\leq 2\exp\Big(\frac{8 - \min\{n, t\}\epsilon^2}{32}\Big)\end{aligned}$$

Therefore, Lemma 4 implies that

$$\mathbb{P}\Big(\Big\|\widetilde{\boldsymbol{\lambda}}^{(t)} - \boldsymbol{\lambda}^{\mathrm{FKEA}(t)}\Big\|_2 \geq \epsilon\Big) \leq 2\exp\Big(\frac{32 - \min\{n, t\}\epsilon^2}{128}\Big)$$

If we define $\delta = 2\exp\big(\frac{32 - \min\{n,t\}\epsilon^2}{128}\big)$, implying that $\epsilon \leq \sqrt{\frac{128\log(3/\delta)}{\min\{n,t\}}}$, then the above inequality shows that

$$\mathbb{P}\Big(\Big\|\widetilde{\boldsymbol{\lambda}}^{(t)} - \boldsymbol{\lambda}^{\mathrm{FKEA}(t)}\Big\|_2 \leq \sqrt{\frac{128\log(3/\delta)}{\min\{n, t\}}}\Big) \geq 1 - \delta$$

Therefore, if we follow the same steps of the proof of Theorem 2, we can show

$$\Big|\mathrm{FKEA\text{-}Vendi}_\alpha^{(t)}(x_1, \ldots, x_n)^{\frac{1-\alpha}{\alpha}} - \mathrm{Vendi}_\alpha^{(t)}(P_x)^{\frac{1-\alpha}{\alpha}}\Big| \leq \sqrt{\frac{128\max\{1, t^{2-\alpha}\}\log(3/\delta)}{\min\{n, t\}}}$$

Proof of Part (b). To show this theorem, we use Theorem 3 from [Xu et al., 2015], which shows that if the rth largest eigenvalue of the kernel matrix $\frac{1}{n}K$ satisfies $\lambda_r \leq \frac{\tau}{n}$, then given $t \geq Cr\log(n)$ (C is a universal constant), the following spectral norm bound will hold with probability $1 - \frac{2}{n^3}$:

$$\big\|\frac{1}{n}K - \frac{1}{n}K^{\mathrm{Nystrom(t)}}\big\|_{sp} \leq \mathcal{O}\Big(\frac{\tau\log(n)}{\sqrt{nt}}\Big).$$

Therefore, Weyl's inequality implies the following for the vector of sorted eigenvalues of $\frac{1}{n}K$, i.e. $\widehat{\boldsymbol{\lambda}}_n$, and that of $\frac{1}{n}K^{\mathrm{Nystrom(t)}}$, i.e., $\boldsymbol{\lambda}^{\mathrm{Nystrom(t)}}$,

$$\big\|\widehat{\boldsymbol{\lambda}}_n - \boldsymbol{\lambda}^{\mathrm{Nystrom(t)}}\big\|_\infty \leq \mathcal{O}\Big(\frac{\tau\log(n)}{\sqrt{nt}}\Big).$$

As a result, considering the subvectors $\widehat{\boldsymbol{\lambda}}_n[1:t]$ and $\boldsymbol{\lambda}^{\mathrm{Nystrom(t)}}[1:t]$ with the first t entries of the vectors, we will have:

$$\left\|\widehat{\boldsymbol{\lambda}}_n[1:t] - \boldsymbol{\lambda}^{\mathrm{Nystrom(t)}}[1:t]\right\|_\infty \leq \mathcal{O}\Big(\frac{\tau\log(n)}{\sqrt{nt}}\Big) \quad \Longrightarrow \quad \left\|\widehat{\boldsymbol{\lambda}}_n[1:t] - \boldsymbol{\lambda}^{\mathrm{Nystrom(t)}}[1:t]\right\|_2 \leq \mathcal{O}\Big(\tau\log(n)\sqrt{\frac{t}{n}}\Big)$$

Noting that the non-zero entries of $\boldsymbol{\lambda}^{\mathrm{Nystrom(t)}}$ are all included in the first-t elements, we can apply Lemma 4 which shows that with probability $1-2n^{-3}$ we have

$$\left\|\widehat{\boldsymbol{\lambda}}_n^{(t)} - \boldsymbol{\lambda}^{\mathrm{Nystrom(t)}}\right\|_2 \leq \mathcal{O}\Big(\tau\log(n)\sqrt{\frac{t}{n}}\Big)$$

Also, in the proof of Theorem 2, we showed that

$$\mathbb{P}\Big(\|\widehat{\boldsymbol{\lambda}}_n^{(t)} - \widetilde{\boldsymbol{\lambda}}^{(t)}\|_2 \leq \sqrt{\frac{32\log(2/\delta)}{n}}\Big) \;\geq\; 1-\delta$$

Combining the above inequalities using a union bound, shows that with probability at least $1-\delta-2n^{-3}$ we have

$$\begin{aligned}
\left\|\boldsymbol{\lambda}^{\mathrm{Nystrom(t)}} - \widetilde{\boldsymbol{\lambda}}^{(t)}\right\|_2 &\leq \left\|\boldsymbol{\lambda}^{\mathrm{Nystrom(t)}} - \widehat{\boldsymbol{\lambda}}_n^{(t)}\right\|_2 + \left\|\widehat{\boldsymbol{\lambda}}_n^{(t)} - \widetilde{\boldsymbol{\lambda}}^{(t)}\right\|_2 \\
&\leq \sqrt{\frac{32\log(2/\delta)}{n}} + \mathcal{O}\Big(\tau\log(n)\sqrt{\frac{t}{n}}\Big) \\
&= \mathcal{O}\Big(\sqrt{\frac{\log(2/\delta)+t\log(n)^2\tau^2}{n}}\Big)
\end{aligned}$$

Hence, repeating the final steps in the proof of Theorem 2, we can prove

$$\Big|\mathrm{Nystrom\text{-}Vendi}_\alpha^{(t)}(x_1,\ldots,x_n)^{\frac{1-\alpha}{\alpha}} - \mathrm{Vendi}_\alpha^{(t)}(P_x)^{\frac{1-\alpha}{\alpha}}\Big| \;\leq\; \mathcal{O}\Big(\sqrt{\frac{\max\{t^{2-\alpha},1\}\big(\log(2/\delta)+t\log(n)^2\tau^2\big)}{n}}\Big)$$

B ADDITIONAL NUMERICAL RESULTS

In this section, we present supplementary results concerning the evaluation of diversity and the convergence behavior of different variants of the Vendi score. We extend the convergence experiments discussed in the main text to include the truncated StyleGAN3-t FFHQ dataset (Figure 12) and the StyleGAN-XL ImageNet dataset (Figure 13). Furthermore, we demonstrate that the truncated Vendi statistic effectively captures the diversity characteristics across various data modalities. Specifically, we conducted similar experiments as shown in Figures 7 and 6 on text data (Figure 9) and video data (Figure 11), showcasing the applicability of the metric across different domains.

We observe in Figure 12 that the convergence behavior is illustrated across various values of ψ. The results indicate that, for a fixed bandwidth σ, the truncated, Nyström, and FKEA variants of the Vendi score converge to the truncated Vendi statistic. As demonstrated in Figure 6 of the main text, this truncated Vendi statistic effectively captures the diversity characteristics inherent in the underlying dataset.

We note that in presence of incremental changes to the diversity of the dataset, finite-dimensional kernels, such as cosine similarity kernel, remain relatively constant. This effect is illustrated in Figure 13, where increase in truncation factor ψ results in incremental change in diversity. This is one of the cases where infinite-dimensional kernel maps with a sensitivity (bandwidth) parameter σ are useful in controlling how responsive the method should be to the change in diversity.

B.1 BANDWIDTH σ SELECTION

In our experiments, we select the Gaussian kernel bandwidth, σ, to ensure that the Vendi metric effectively distinguishes the inherent modes within the dataset. The kernel bandwidth directly controls the sensitivity of the metric to the underlying data clusters. As illustrated in Figure 10, varying σ significantly impacts the diversity computation on the ImageNet dataset. A smaller bandwidth (e.g., $\sigma = 20, 30$) results in the metric treating redundant samples as distinct modes, artificially inflating the number of clusters, which in turn slows down the convergence of the metric. On the other hand, large bandwidth results in instant convergence of the metric, i.e. in $\sigma = 60$ $n = 100$ and $n = 1000$ have almost the same amount of diversity.

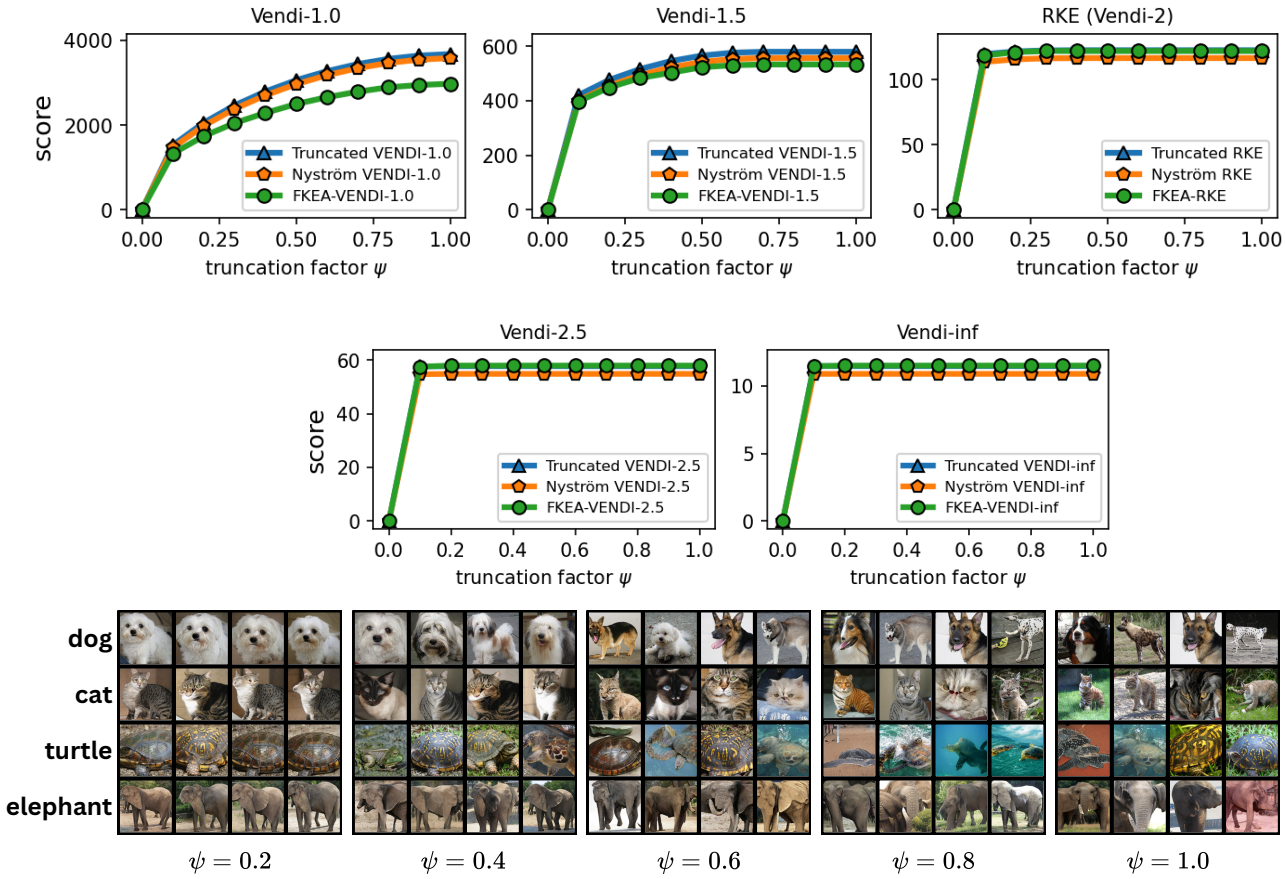

Figure 8: Diversity evaluation of Vendi scores on StyleGAN-XL generated ImageNet dataset with varying truncation parameter ψ. The setting is based on *DinoV2* embedding and bandwidth $\sigma = 30$

Table 1: Statistical convergence of diversity scores for different sample size on DALL-E 3 generated MSCOCO data

n	VENDI-1.0	RKE	Vendi-t	FKEA-Vendi	Nystrom-Vendi	Recall	Coverage
2000	239.91	13.47	239.91	228.69	239.91	0.76	0.86
4000	315.35	13.51	315.35	280.68	315.35	0.81	0.87
6000	357.15	13.56	346.27	310.9	345.49	0.83	0.91
8000	392.36	13.56	354.8	329.56	357.41	0.87	0.91

Table 2: Statistical convergence of diversity scores for different sample size on SDXL generated MSCOCO data

n	VENDI-1.0	RKE	Vendi-t	FKEA-Vendi	Nystrom-Vendi	Recall	Coverage
2000	187.17	10.65	187.17	173.06	187.18	0.78	0.85
4000	236.49	10.7	236.49	222.78	236.08	0.82	0.87
6000	264.82	10.7	258.21	236.37	257.34	0.86	0.87
8000	289.08	10.71	265.84	251.59	266.23	0.86	0.86
10000	304.44	10.72	267.39	256.24	268.34	0.86	0.87

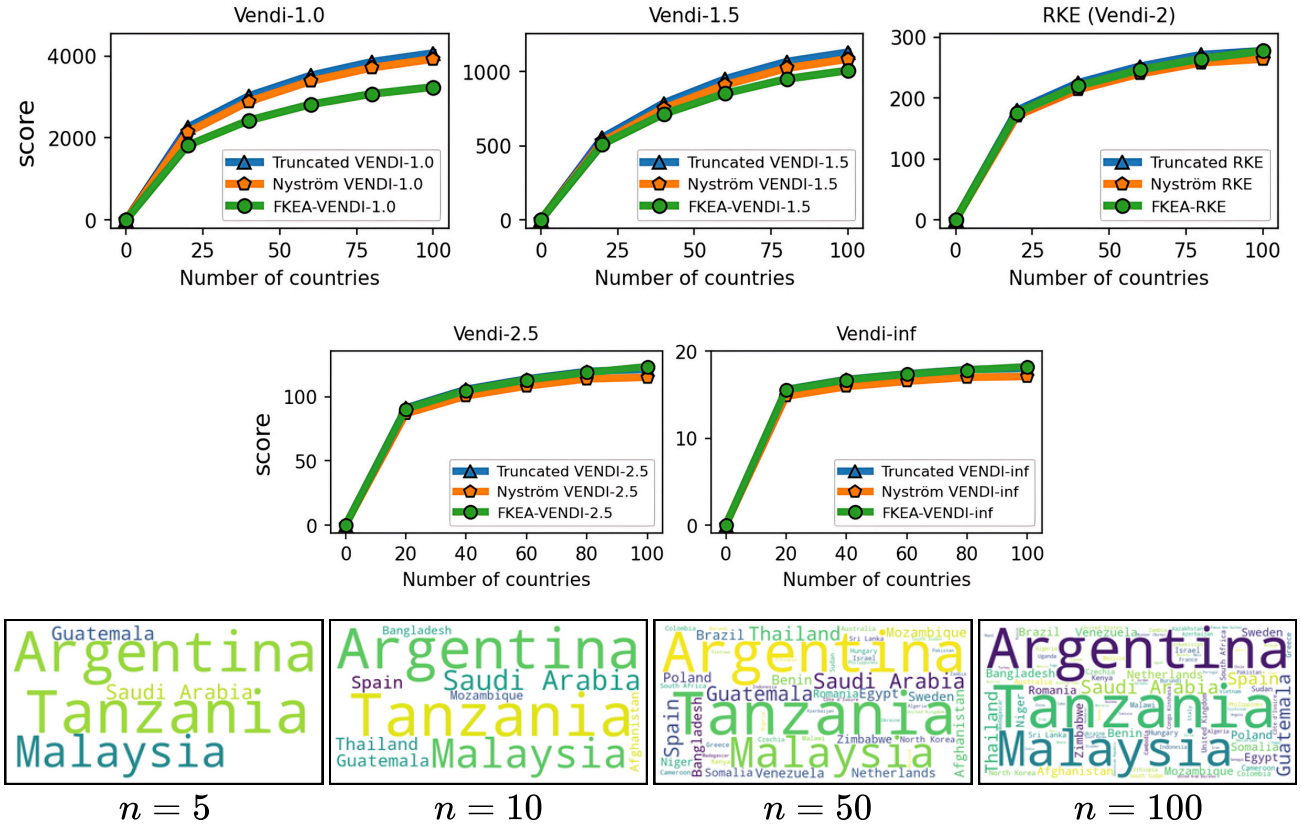

Figure 9: Diversity evaluation of Vendi scores on synthetic text dataset about 100 countries generated by GPT-4 with varying number of countries. The setting is based on *text-embedding-3-large* embedding and bandwidth $\sigma = 0.5$

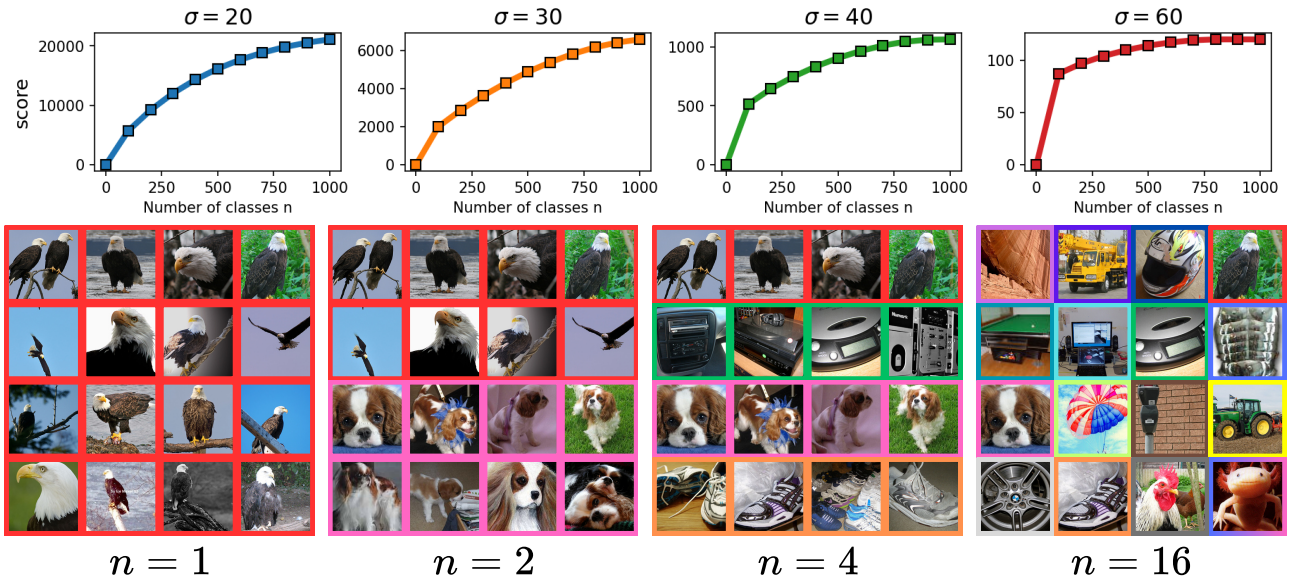

Figure 10: The diagram outlining an intuition behind a kernel bandwidth σ selection in diversity evaluation.

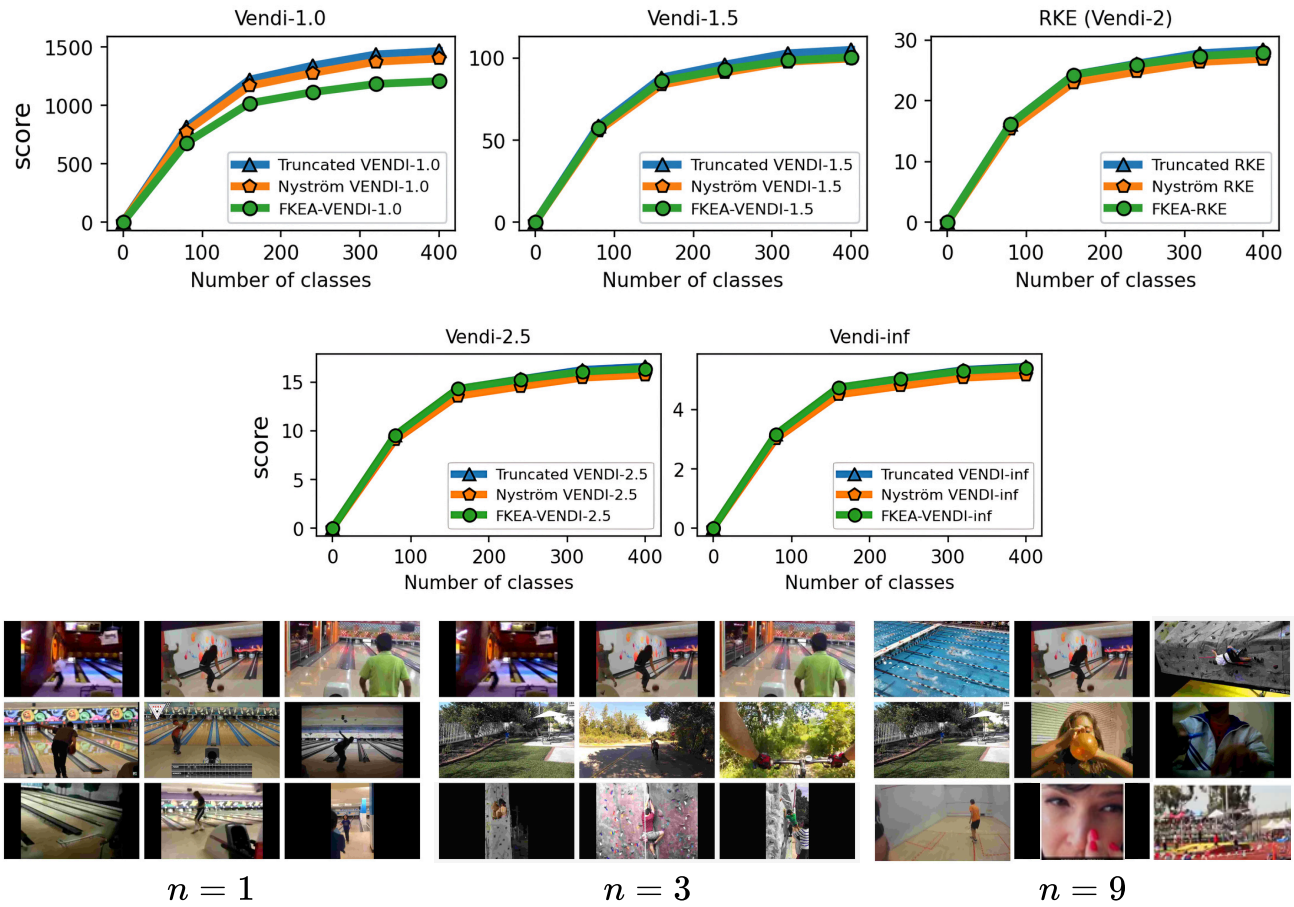

Figure 11: Diversity evaluation of Vendi scores on Kinetics400 dataset with varying number of classes. The setting is based on *I3D* embedding and bandwidth $\sigma = 4.0$

Table 3: Compilation time (in seconds) of different Vendi scores with increasing sample size

Metric	samples n						
	10000	20000	30000	40000	50000	60000	70000
Vendi	97s	631s	1868s	-	-	-	-
FKEA-Vendi	19s	36s	53s	71s	88s	105s	124s
Nystrom-Vendi	31s	44s	78s	91s	112s	136s	164s

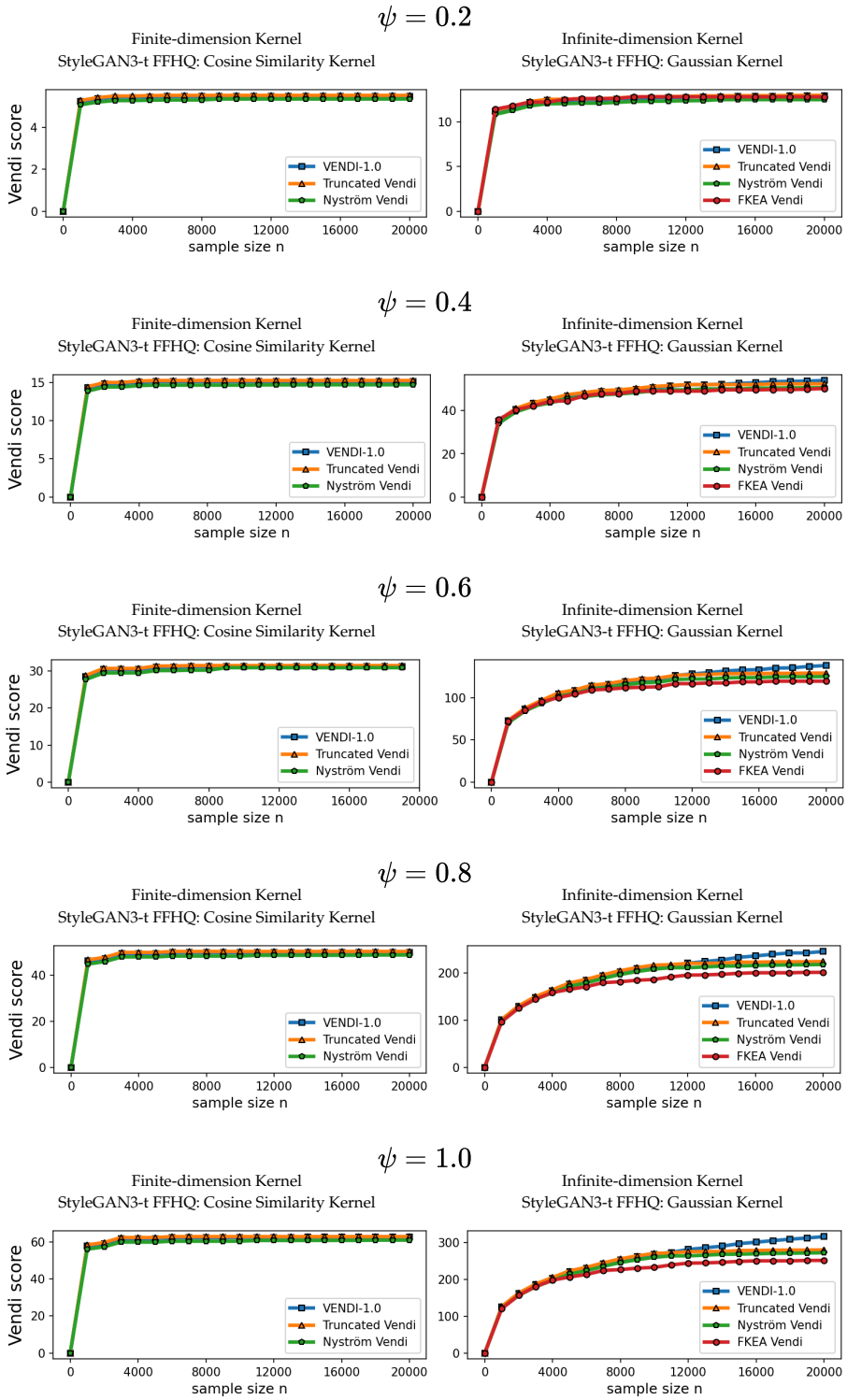

Figure 12: Statistical convergence of Vendi score for different sample sizes on StyleGAN3 generated FFHQ data at various truncation factors ψ: (Left plot) finite-dimension cosine similarity kernel (Right plot) infinite dimension Gaussian kernel with bandwidth $\sigma = 35$. *DinoV2* embedding (dimension 768) is used in computing the scores.

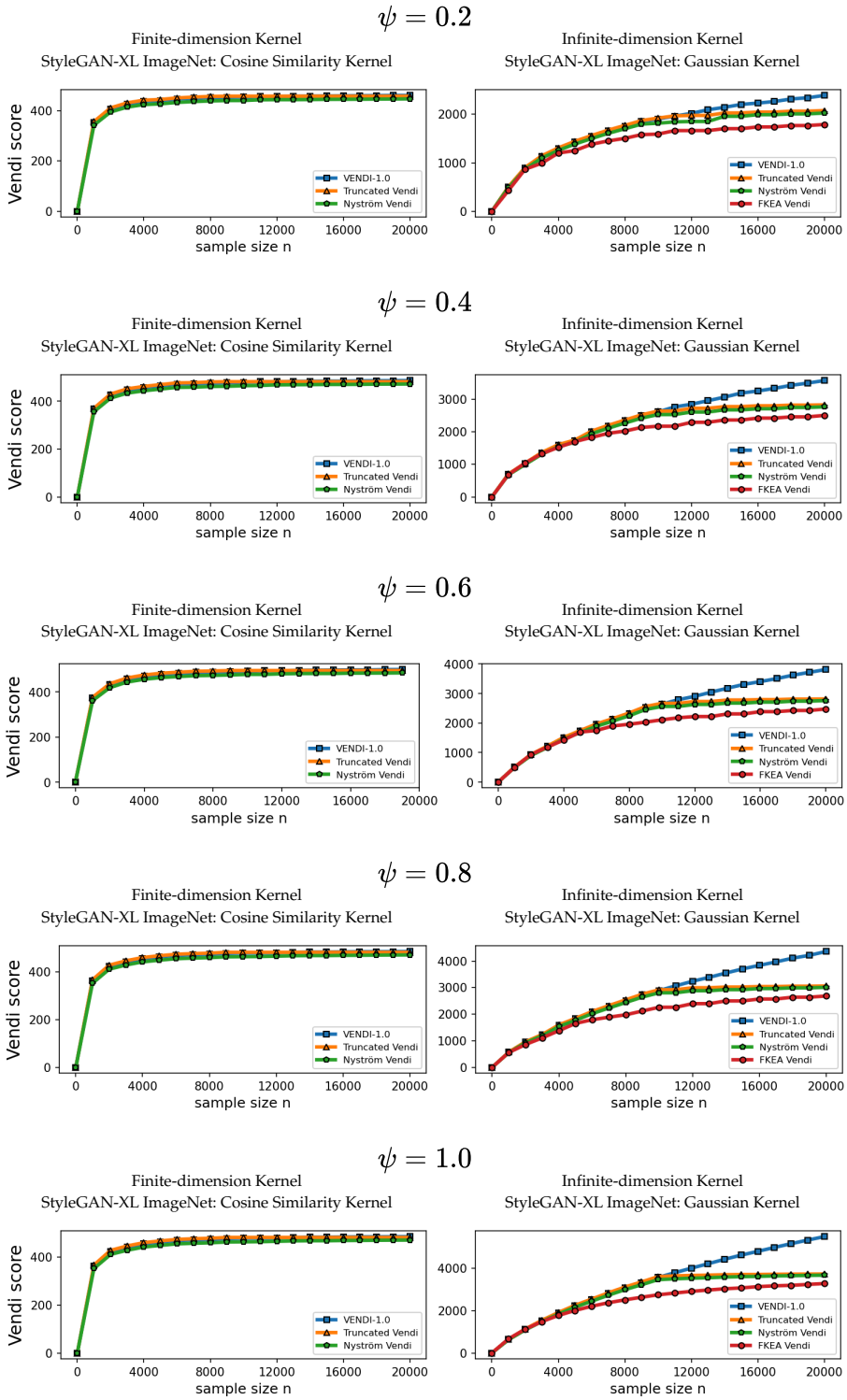

Figure 13: Statistical convergence of Vendi score for different sample sizes on StyleGAN-XL generated ImageNet data: (Left plot) finite-dimension cosine similarity kernel (Right plot) infinite dimension Gaussian kernel with bandwidth $\sigma = 40$. *DinoV2* embedding (dimension 768) is used in computing the scores.

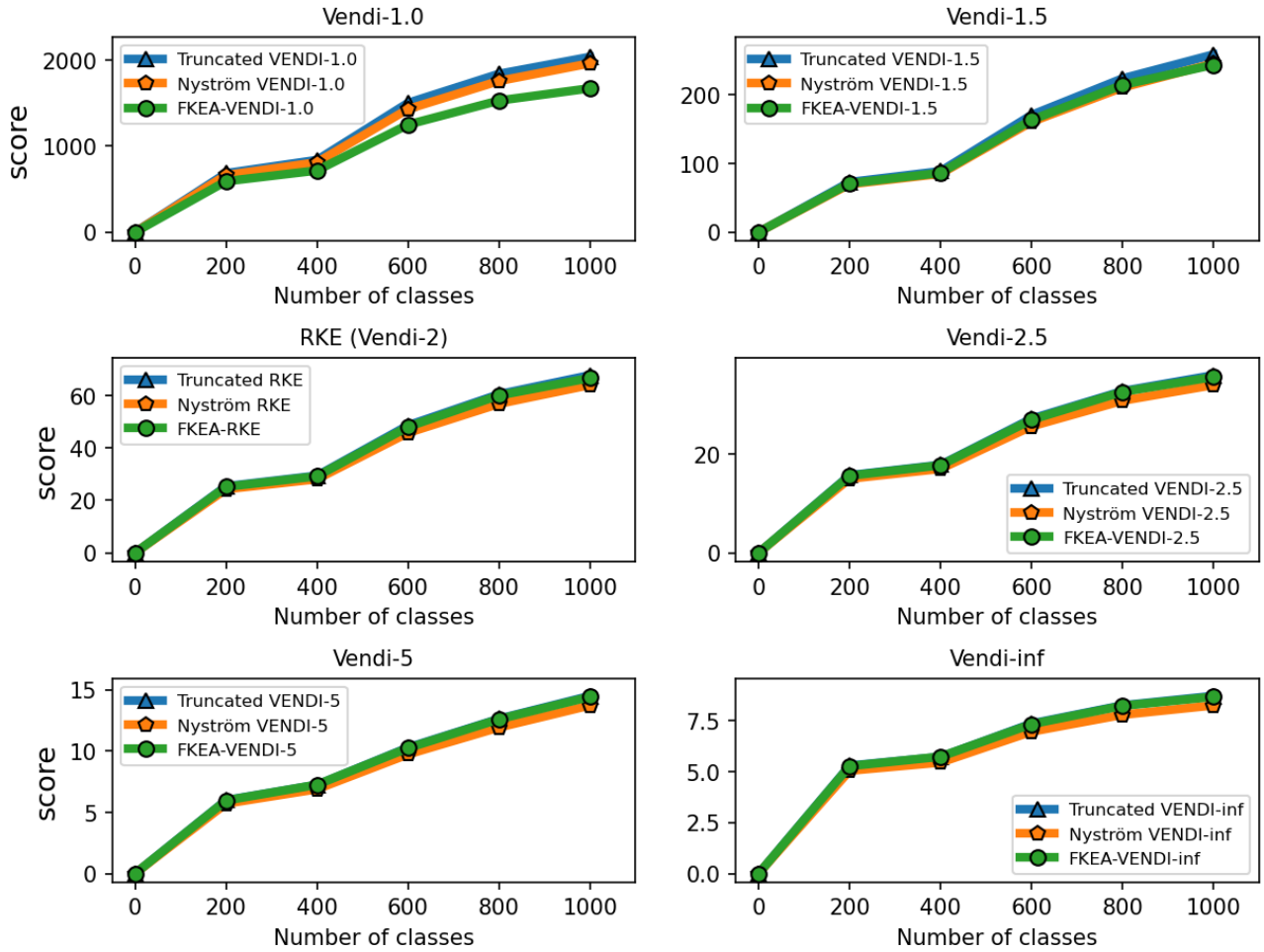

Figure 14: Diversity evaluation of Vendi scores on ImageNet dataset with varying number of classes based on *CLIP* embedding and bandwidth $\sigma = 5.0$

C SELECTION OF EMBEDDING SPACE

To show that proposed truncated Vendi score remains feasible under arbitrary embedding selection, we conducted experiments from Figures 6 and 7. Figures 14, 15, 16 and 17 extend the results to CLIP Radford et al. [2021] and SWaV Caron et al. [2020] embeddings. These experiments demonstrate that FKEA, Nyström and t-truncated Vendi correlate with increasing diversity of the evaluated dataset. We emphasize that proposed statistic remains feasible under arbitrary embedding space that is capable of mapping image samples into a latent space.

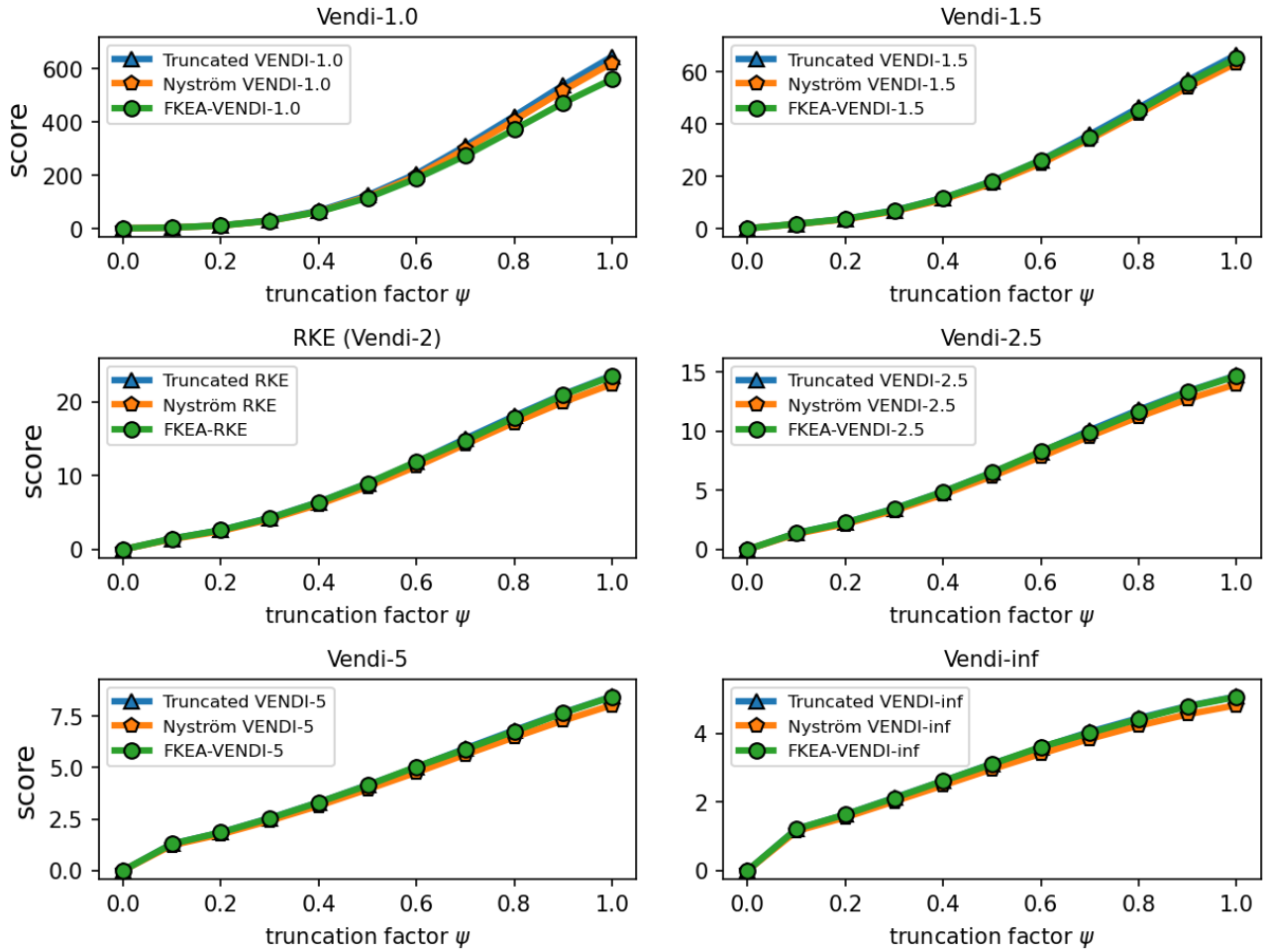

Figure 15: Diversity evaluation of Vendi scores on truncated StyleGAN3 generated FFHQ with varying truncation coefficient ψ based on *CLIP* embedding and bandwidth $\sigma = 5.0$

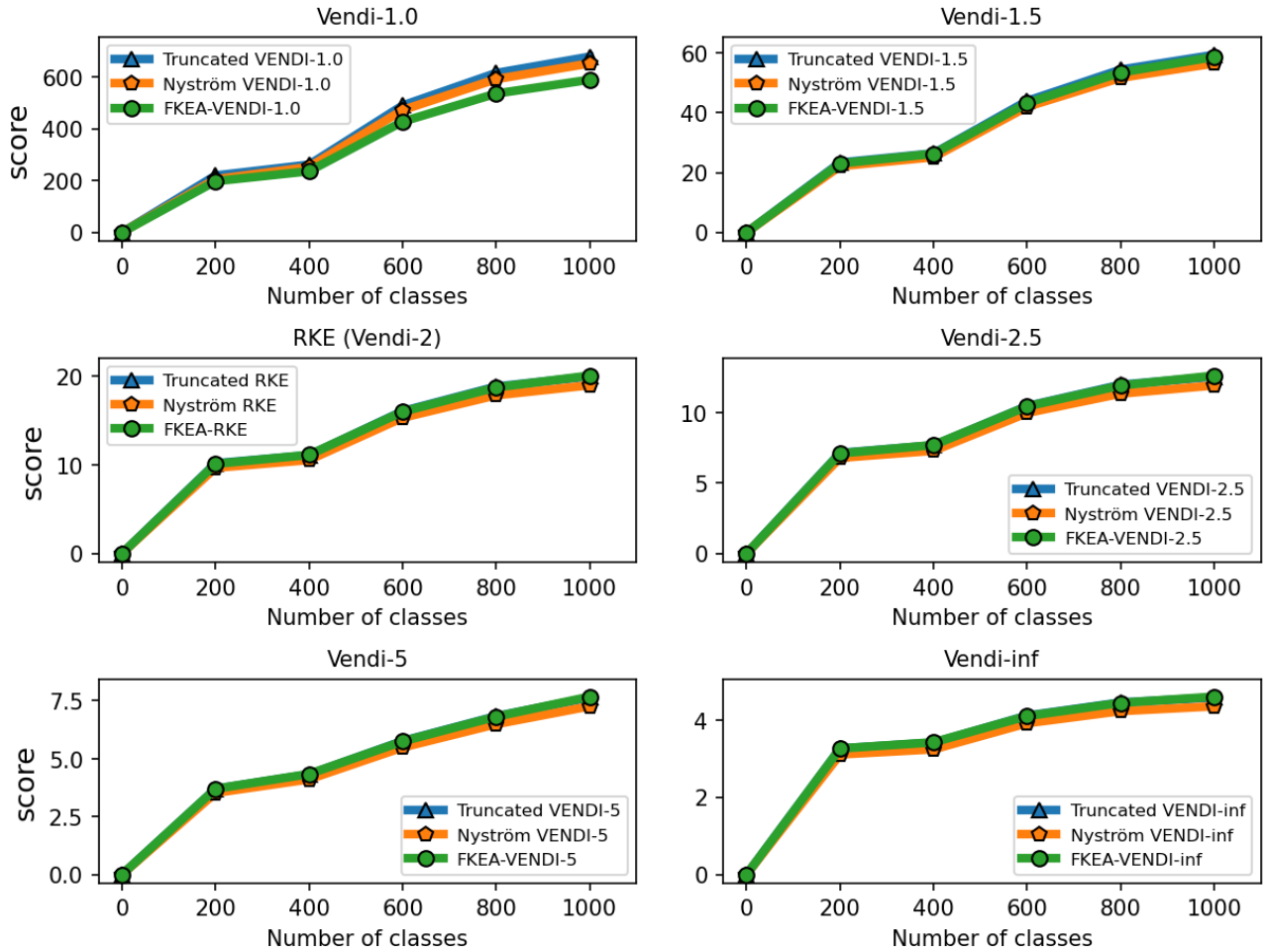

Figure 16: Diversity evaluation of Vendi scores on ImageNet dataset with varying number of classes based on *SWaV* embedding and bandwidth $\sigma = 1.0$

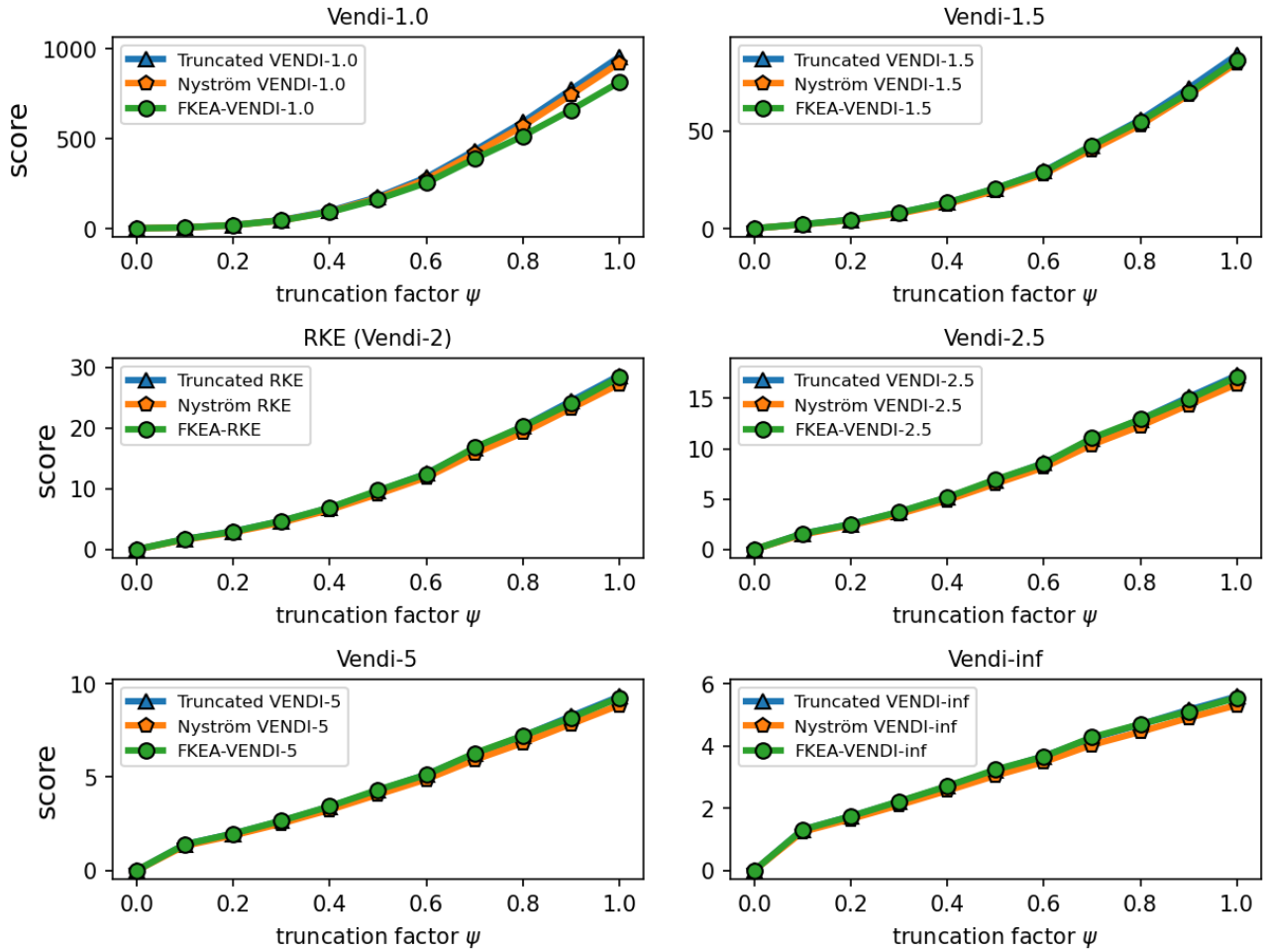

Figure 17: Diversity evaluation of Vendi scores on truncated StyleGAN3 generated FFHQ with varying truncation coefficient ψ based on *SwAV* embedding and bandwidth $\sigma = 1.0$