EXPOSING DIVERSITY BIAS IN DEEP GENERATIVE MODELS: STATISTICAL ORIGINS AND CORRECTION OF DIVERSITY ERROR

Farzan Farnia, Mohammad Jalali, Azim Ospanov*
Department of Computer Science and Engineering
The Chinese University of Hong Kong
{farnia,mjalali24,aospanov9}@cse.cuhk.edu.hk

ABSTRACT

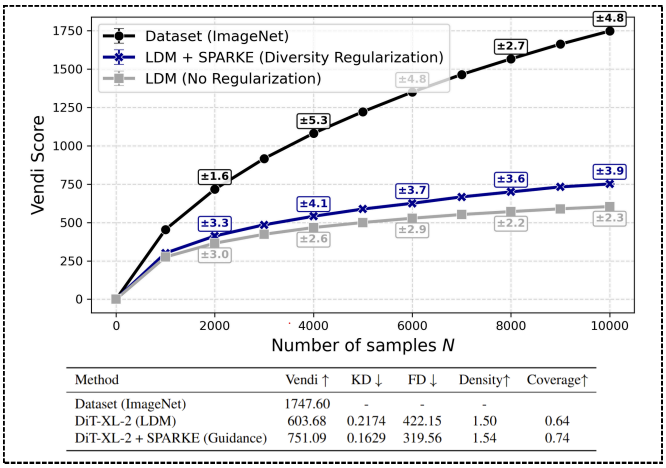
Deep generative models have achieved great success in producing high-quality samples, making them a central tool across machine learning applications. Beyond sample quality, an important yet less systematically studied question is whether trained generative models faithfully capture the diversity of the underlying data distribution. In this work, we address this question by directly comparing the diversity of samples generated by state-of-the-art models with that of test samples drawn from the target data distribution, using recently proposed reference-free entropy-based diversity scores, Vendi and RKE. Across multiple benchmark datasets, we find that test data consistently attains substantially higher Vendi and RKE diversity scores than the generated samples, suggesting a systematic downward diversity bias in modern generative models. To understand the origin of this bias, we analyze the finite-sample behavior of entropy-based diversity scores and show that their expected values increase with sample size, implying that diversity estimated from finite training sets could inherently underestimate the diversity of the true distribution. As a result, optimizing the generators to minimize divergence to empirical data distributions would induce a loss of diversity. Finally, we discuss potential diversity-aware regularization and guidance strategies based on Vendi and RKE as principled directions for mitigating this bias, and provide empirical evidence suggesting their potential to improve the results.

1 INTRODUCTION

Deep generative models have achieved substantial progress in synthesizing complex, high-dimensional data across vision, language, and audio modalities. Early generative modeling approaches including variational autoencoders (VAEs) (Kingma & Welling, 2013) and generative adversarial networks (GANs) (Goodfellow et al., 2014) established scalable paradigms for likelihood-based and divergence minimization-based data generation, while recently diffusion and score-based models (Song & Ermon, 2019; Ho et al., 2020; Song et al., 2021) further improved training stability by formulating generation as iterative denoising. From a statistical perspective, these models aim to learn a neural net-based sample generator that produces fresh draws approximating the data-generating distribution, without explicitly restricting the distribution to a fixed parametric family as in classical statistics.

Evaluating how well a generative model approximates the target distribution is typically addressed by standard quantitative metrics in the literature. In image generation, the Fréchet Distance (FD) (Heusel et al., 2017; Kynkäänniemi et al., 2023; Stein et al., 2023) compares mean and covariance statistics with a reference data distribution, which occurs after embedding samples using pretrained feature extractors such as Inception (Szegedy et al., 2016), CLIP (Radford et al., 2021), and DINOv2 (Oquab et al., 2023). The Kernel Inception Distance (KID) (Bińkowski et al., 2018) relies instead on the maximum mean discrepancy (MMD) (Gretton et al., 2012) measured in the same embedding spaces. These metrics are widely used for benchmarking and model comparison,

*Authors are listed in alphabetical order.



Method	Vendi ↑	KD ↓	FD ↓	Density↑	Coverage↑
Dataset (ImageNet)	1747.60	-	-	-	
DiT-XL-2 (LDM)	603.68	0.2174	422.15	1.50	0.64
DiT-XL-2 + SPARKE (Guidance)	751.09	0.1629	319.56	1.54	0.74

Figure 1: Numerical evaluation of Vendi scores at different sample sizes (averaged over 100 independent random trials with confidence interval) for the validation set of ImageNet, the pre-trained latent diffusion model (LDM) of DiT-XL-2, and the SPRAKE-guided LDM with entropy diversity regularization proposed in (Jalali et al., 2025). The entropy regularization in the SPARKE Guidance not only improves the Vendi and Coverage diversity scores, but further improves the FD and KD overall evaluation scores.

but they are not designed to reveal systematic effects that arise from finite-sample training of the generative models or to isolate structural biases in learned generative distributions.

In this work, we focus on *diversity* as one such structural aspect that could be systematically biased in the trained models. Specifically, we study whether standard deep generative models can capture the diversity of the underlying data distribution, or whether they exhibit a systematic shortfall. To do so, we adopt reference-free entropy-based diversity measures, including the Vendi score (Friedman & Dieng, 2023) and the Rényi Kernel Entropy (RKE) score (Jalali et al., 2023), which quantify diversity through spectral entropy of kernel similarity matrices computed from samples. Using these measures as well as the reference-based diversity scores Recall (Sajjadi et al., 2018; Kynkäänniemi et al., 2019) and Coverage (Naeem et al., 2020), we evaluate state-of-the-art generative models on standard benchmarks such as ImageNet (Deng et al., 2009), MS-COCO (Lin et al., 2014), and FFHQ (Karras et al., 2019). Across all settings, we observe that generated samples consistently exhibit lower entropy-based diversity than the real datasets, suggesting a universal downward diversity bias.

To understand potential reasons for this bias, we specifically examine the finite size of the training dataset for generative models and whether their empirical distributions capture the full diversity of the underlying population distribution. For basic discrete random variables, the underestimation of entropy from finite empirical samples is well-known in classical statistics (Miller, 1955; Basharin, 1959). It is plausible that similar effects occur for kernel-entropy Vendi diversity measures of finite training datasets. To investigate this, we measure the Vendi score on progressively larger subsamples of real datasets. Within the range where exact spectral computation of Vendi score is feasible, we observe a pronounced increase in measured diversity with sample size, suggesting that common dataset sizes may still lie in a regime where diversity estimates have not stabilized.

This empirical behavior connects to the classical theory of entropy estimation under finite sampling. Bias analyses for plug-in entropy estimators (Miller, 1955) and general nonparametric treatments of entropy estimation (Paninski, 2003) show that empirical procedures can systematically underestimate population-level entropy. While the Vendi score in (Friedman & Dieng, 2023) is the matrix-based spectral entropy rather than standard Shannon entropy, it still depends on a finite-sample empirical kernel operator. Building on this perspective, we prove that the expected logarithm of the Vendi score of an i.i.d. sample set is monotonically increasing with the sample size. This formalizes the intuition that finite datasets could under-represent population diversity, and it suggests one possible route by which a diversity shortfall can arise before a generative model is trained.

Beyond diagnosing the bias, understanding downward diversity bias has practical implications for model training and inference. Our theoretical results suggest that projecting a learned distribution onto suitable Vendi or RKE-based super-level sets can reduce kernel discrepancies measured by MMD when the learned model exhibits a diversity shortfall relative to the population distribution. This motivates entropy-based diversity regularization during training or fine-tuning of the generative models. This also provides a theoretical justification for diversity-aware guidance mechanisms, including recently proposed Vendi Guidance (Askari Hemmat et al., 2024) and RKE Guidance (Jalali et al., 2025), at inference time for diffusion models.

2 Related Works

We defer the discussion of the related works to the Appendix, Section A.

3 Preliminaries

Consider a data random variable $X \in \mathcal{X}$ distributed according to P_X over the sample space $\mathcal{X}$. Then, a trained generative model $\mathcal{G}$ induces a probability distribution $P_{\mathcal{G}}$ on variable X. In a reference-free evaluation process, the evaluator has only access to n independently generated samples from $P_{\mathcal{G}}$, denoted by $x_1, \ldots, x_n \in \mathcal{X}$. In our analysis, we utilize entropy-based metrics that quantify the diversity of the drawn sample set $x_1, \ldots, x_n$ in a reference-free manner without requiring access to a reference distribution. In our evaluation of image generative models, we use DINOv2-embedded data following the reference (Stein et al., 2023).

3.1 Kernel Functions and Kernel Matrices

Following the standard definition, a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a kernel if for every $n \in \mathbb{N}$ and inputs $x_1, \ldots, x_n \in \mathcal{X}$, the corresponding Gram matrix $K \in \mathbb{R}^{n \times n}$ with entries $K_{ij} = k(x_i, x_j)$ is positive semidefinite (PSD). By the classical characterization due to Aronszajn (1950), there exists a (possibly infinite-dimensional) reproducing kernel Hilbert space $\mathcal{H}$ and a feature map $\phi : \mathcal{X} \to \mathcal{H}$ such that

$$k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}. \tag{1}$$

In this work, we focus on *normalized kernels* satisfying $k(x, x) = 1$ for all $x \in \mathcal{X}$. Common examples include cosine similarity applied to unit-normalized embeddings and the Gaussian (RBF) kernel. For a normalized kernel, the Gram matrix satisfies $\mathrm{Tr}(K) = \sum_{i=1}^{n} k(x_i, x_i) = n$, and therefore the normalized kernel matrix $\frac{1}{n}K$ has unit trace. Since $\frac{1}{n}K$ is PSD, its eigenvalues are nonnegative and sum to one, i.e. they form a probability model.

3.2 Kernel Covariance and Population Operator

Let $\Phi \in \mathbb{R}^{n \times d}$ denote the feature matrix whose i-th row is $\phi(x_i)^\top$ in the finite-dimensional case (and the same identities hold at the operator level in general). Then the normalized kernel matrix can be written as $\frac{1}{n}K = \frac{1}{n}\Phi\Phi^\top$. Define the empirical kernel covariance matrix $\widehat{C}_X \in \mathbb{R}^{d \times d}$:

$$\widehat{C}_X := \frac{1}{n} \sum_{i=1}^{n} \phi(x_i)\phi(x_i)^\top = \frac{1}{n}\Phi^\top\Phi. \tag{2}$$

The matrices $\frac{1}{n}\Phi\Phi^\top$ and $\widehat{C}_X$ share the same non-zero eigenvalues (singular values of $\frac{1}{\sqrt{n}}\Phi$), and thus any spectral functional applied to $\frac{1}{n}K$ can equivalently be applied to $\widehat{C}_X$.

At the population level, we define the kernel covariance operator

$$\widetilde{C}_X := \mathbb{E}_{X \sim P_X}\left[\phi(X)\phi(X)^\top\right]. \tag{3}$$

Under the normalized kernel assumption, $\widetilde{C}_X$ is positive semidefinite and satisfies $\mathrm{Tr}(\widetilde{C}_X) = \mathbb{E}[k(X, X)] = 1$, thus it is a density matrix whose eigenvalues form a valid probability model.

3.3 MATRIX-BASED ENTROPY: VENDI AND RKE DIVERSITY SCORES

Let $A \in \mathbb{R}^{d\times d}$ be a PSD matrix with unit trace $\mathrm{Tr}(A) = 1$, and let $\{\lambda_i\}_{i=1}^d$ denote its eigenvalues. The *von Neumann entropy* of A is defined as

$$H(A) := \sum_{i=1}^{d} \lambda_i \log \frac{1}{\lambda_i} \tag{4}$$

Given samples $x_1, \ldots, x_n$ and their normalized kernel matrix $\frac{1}{n}K$, Friedman & Dieng (2023) define the *Vendi score* as the exponential of the von Neumann entropy,

$$\mathrm{Vendi}(x_1, .., x_n) := \exp\big(H\big(\frac{1}{n}K\big)\big) = \exp\big(H(\widehat{C}_X)\big) \tag{5}$$

Following the analysis in (Bach, 2022; Ospanov & Farnia, 2025), the population Vendi score of a distribution P_X is defined analogously by applying the same construction to the population kernel covariance operator $\widetilde{C}_X$,

$$\mathrm{Vendi}(P_X) := \exp\big(H(\widetilde{C}_X)\big). \tag{6}$$

Remark 1. *As discussed in (Friedman & Dieng, 2023; Ospanov et al., 2024), the* exact *computation of the Vendi score with a Gaussian kernel (and more generally infinite-dimensional kernels) requires an eigendecomposition of the $n \times n$ kernel matrix, which becomes computationally prohibitively expensive for dataset sizes larger than ≈ 20000. Therefore, our Vendi score evaluations remain constrained to 20000 samples, and we report the Vendi score plots at sample sizes below this threshold. For image data, we use DINOv2 backbone embedding following (Stein et al., 2023).*

Also, as we prove in the Appendix (Proposition 2), the Vendi score $\mathrm{Vendi}(x_1, \ldots, x_n)$ of an instance of n i.i.d. samples $x_1, \ldots, x_n \overset{iid}{\sim} P_X$ concentrates around the expected value $\mathbb{E}[\mathrm{Vendi}(x_1, \ldots, x_n)]$ where the expectation is over the randomness of n i.i.d. samples from P_X (this is different from the Vendi score of the population distribution $\mathrm{Vendi}(P_X)$). To further ensure our reported scores are properties of the underlying distribution, we perform the Vendi score evaluation for $M = 10$ independent trials and report the averaged score and its confidence interval.

Also, note that the *RKE score* (Jalali et al., 2023) corresponds to order-2 Rényi entropy of the kernel matrix, which for samples $x_1, \ldots, x_n$ and their kernel matrix K will be ($\|\cdot\|_F$ denotes the Frobenius norm):

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

As discussed in (Jalali et al., 2023), the population RKE admits a simple characterization: if X, X' are i.i.d. samples from P_X, then $\mathrm{RKE}(P_X) = 1/\mathbb{E}\big[k^2(X, X')\big]$. This identity follows from interpreting the order-2 kernel entropy through the Hilbert–Schmidt norm of the population kernel covariance operator.

4 FINITE-SAMPLE ENTROPY BIAS AND ITS EFFECTS ON DIVERSITY IN GENERATIVE MODELS

A central question in this work is how reliably entropy-based diversity scores of finite samples $x_1, \ldots, x_n$ reflect the true entropy score of the underlying data distribution P_X. This question is particularly relevant for generative modeling, where the generative model is trained on finite datasets. If the expected entropy score of the empirical distribution $\widehat{P}_n$ of n i.i.d. samples from P exhibits a noticeable downward bias relative to the entropy of population distribution P (entropy of infinite samples from P), then the finite training set would expose the model to less diversity than the true diversity of P. Therefore, such an entropy-based diversity gap between the empirical distribution of training data $\widehat{P}_n$ and the underlying P can plausibly propagate to trained generative models on the empirical samples.

In this section, we aim to analyze this phenomenon from a statistical perspective. We begin by reviewing the classical downward bias of the Shannon entropy of finite sample sets for basic discrete random variables. We then extend the same line of reasoning to the von Neumann entropy of the

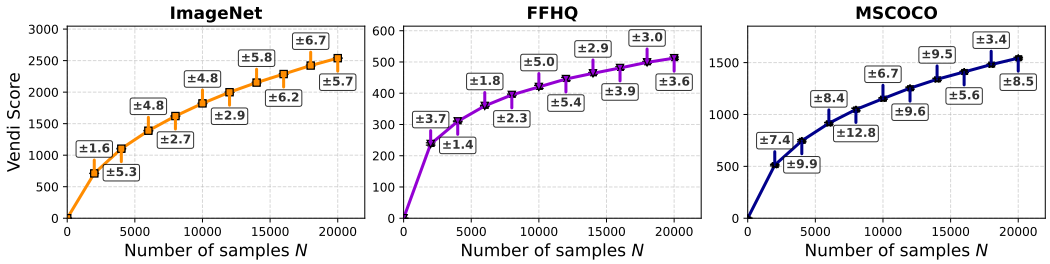


Figure 2: Vendi score curves (mean and confidence intervals over 10 independent sample sets) for ImageNet, FFHQ, and MSCOCO (sample size values of $n \leq 20$K), computed using DINOv2 embeddings and a Gaussian (RBF) kernel with bandwidth $\sigma = 35$. The Vendi scores continue to increase at a significant rate across all sample sizes below the 20000-computational-feasible size limit for exact Vendi computation.

empirical kernel covariance operator, which underlies the Vendi score. Two controlled experiments, one discrete and one continuous, are used to illustrate how entropy-based scores increase with the evaluation sample size n and exhibit systematic downward bias at moderate n. These observations provide the conceptual foundation for the next section, where we empirically study diversity bias in pretrained generative models.

4.1 Classical background: downward bias of finite-sample Shannon entropy

Let P be a discrete distribution on an alphabet of size r, and let $\widehat{P}_n$ denote the empirical distribution formed by n i.i.d. samples from P. The plug-in (maximum-likelihood) estimator of Shannon entropy is $H(\widehat{P}_n)$, i.e., the Shannon entropy of the n empirical samples from P. It is well-known in statistics and information theory that this estimator is *downward biased*, as

$$\mathbb{E}\big[H(\widehat{P}_n)\big] \;\leq\; H(P). \tag{8}$$

Note that the above inequality follows directly from the well-known concavity of Shannon entropy and Jensen's inequality as $\mathbb{E}[\widehat{P}_n] = P$. Beyond the negative sign of this bias, sharp asymptotic expansions are already known. In particular, for fixed r and large n, the leading bias term scales as $\frac{r-1}{2n}$ (in nats), i.e.

$$H(P) - H(\widehat{P}_n) \;\gtrsim\; \frac{r-1}{2n}, \tag{9}$$

as shown by Miller (1955) and Basharin (1959) and refined in (Paninski, 2003; Schürmann, 2004).

4.2 From Shannon entropy to log-Vendi: a kernel-based analogue

We now extend the above intuition to the matrix-based entropy underlying the Vendi score. Recall the normalized kernel feature map $\phi : \mathcal{X} \to \mathcal{H}$ satisfying $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$ and $\|\phi(x)\|^2_{\mathcal{H}} = k(x, x) = 1$. Given i.i.d. samples $X_1, \ldots, X_n \sim P$, define the empirical kernel covariance operator

$$\widehat{C}_n := \frac{1}{n} \sum_{i=1}^{n} \phi(X_i)\phi(X_i)^\top, \tag{10}$$

which is positive semidefinite and has unit trace. The von Neumann entropy $H(\widehat{C}_n) = -\mathrm{Tr}\big(\widehat{C}_n \log \widehat{C}_n\big)$ equals the logarithm of the Vendi score, $\mathrm{Vendi}(X_{1:n}) = \exp(H(\widehat{C}_n))$. We therefore refer to $H(\widehat{C}_n)$ as *log-Vendi*.

Proposition 1 (Monotone increase of expected log-Vendi)**.** *For i.i.d. samples $X_1, \ldots, X_n \overset{iid}{\sim} P$, the sequence $\mathbb{E}\big[H(\widehat{C}_n)\big] = \mathbb{E}\big[\log\big(\mathrm{Vendi}(X_1, \ldots, X_n)\big)\big]$ increases monotonically in sample size n, i.e.,*

$$\forall m, n \in \mathbb{N}: \quad m \leq n \implies \mathbb{E}\big[H(\widehat{C}_m)\big] \leq \mathbb{E}\big[H(\widehat{C}_n)\big]$$

Note that the expectation is with respect to the randomness of n i.i.d. drawn samples from P.

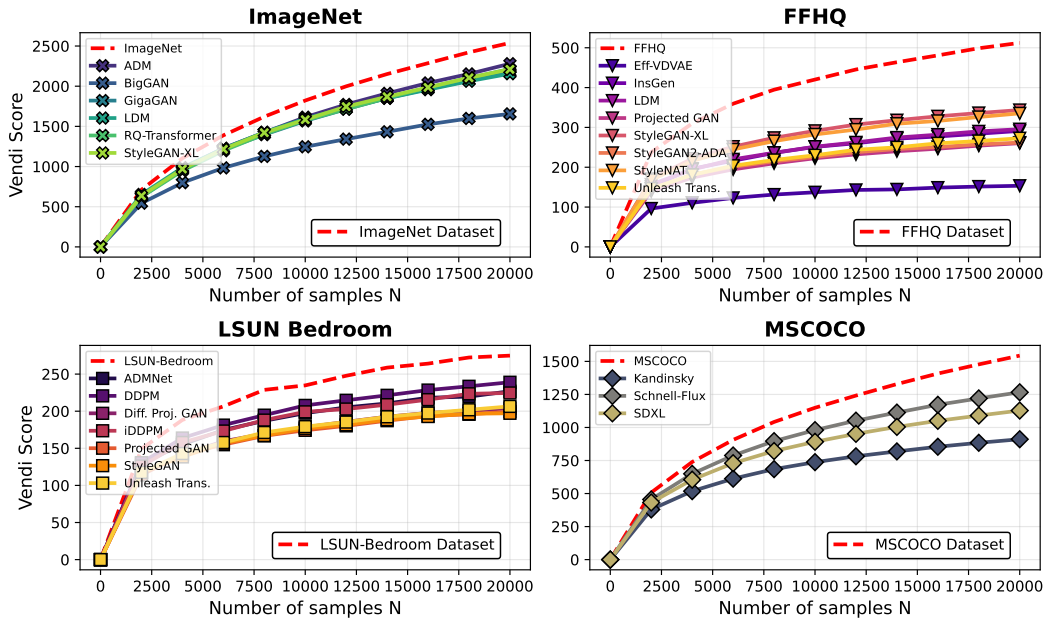


Figure 3: Comparison of Vendi scores of the test sample set (the dashed red curve) and the generated samples by pre-trained generative models across four datasets. The backbone embedding is DINOv2 embeddings using Gaussian (RBF) kernel with bandwidth $\sigma = 35$.

The results in this section suggest that the empirical distribution $\widehat{P}_n$ of finite n samples could exhibit an entropy gap compared to the true distribution P. In the next section, we numerically analyze this entropy gap, which can cause a downward diversity bias in trained generative models over finite datasets relative to the underlying distribution of the test data.

5 Evaluating Downward Entropy Bias in Generative Models

We empirically evaluate whether entropy-based diversity scores reveal a systematic gap between real datasets and samples produced by standard pretrained generative models. Following Section 4, we frame diversity evaluation as a finite-sample statistical problem: for any distribution P, either the empirical distribution induced by a dataset or a learned model distribution, we consider the expected Vendi score of an i.i.d. sample of size n drawn from P. Our goal is to compare these expected scores under a matched evaluation protocol.

We focus on ImageNet (Deng et al., 2009), FFHQ (Karras et al., 2019), and MS-COCO (Lin et al., 2014), and use pretrained models and generated samples from the `dgm-eval` repository (Stein et al., 2023). For each dataset and each model, finite-dataset-size expected Vendi scores are estimated by averaging the Vendi score over 5 independently sampled subsets of each size n, and we report the mean and 95% confidence interval.

Figures 2 and 3 highlight two consistent empirical patterns: 1) For each dataset, the Vendi score of the real dataset samples increases substantially with the evaluation sample size n across the computationally feasible range (up to 20000). 2) Across all evaluated n values, the Vendi curves of samples generated by standard pretrained models remain consistently below the corresponding curve of the dataset itself. This separation persists across ImageNet, FFHQ, and MS-COCO.

To complement Vendi, we also report more scalable diversity metrics, including RKE (order-2 kernel entropy in the same embedding) as well as Recall and Coverage metrics implemented in `dgm-eval`. Across the same datasets and pretrained models, these metrics similarly indicate that generated samples do not match the diversity of true dataset samples, supporting what is suggested by the Vendi curves. Detailed results are provided in Tables 2, 3 and 4 found in the Appendix.

Finally, we examine whether training-set size contributes to the observed diversity gap. We train StyleGAN-XL (Sauer et al., 2022) on ImageNet using subsets of sizes 1,000,000, 500,000, 100,000, and observe that the Vendi diversity of generated samples increases markedly with the size of the training set. A similar trend is observed for an unconditional latent diffusion model (Rombach et al., 2022) trained on FFHQ subsets of sizes 70,000, 35,000, and 17,500. These controlled experiments support the view that limited training-set size can be a substantive contributor to the entropy-based diversity gap observed between real datasets and standard pretrained generative models.

6 Entropy-Based Diversity Regularization for Generative Models

The preceding sections reveal a consistent entropy-based diversity gap: under the same reference-free evaluation protocol, the Vendi/RKE scores of samples from a learned model distribution Q_θ can fall below those of samples from the data distribution P. Motivated by the monotone growth of the expected log-Vendi score with sample size, we view this gap through a regularization lens: if P lies in a high-entropy region of the distribution space, then encouraging Q_θ to move toward that region is a principled way to reduce an entropy-defined notion of diversity error.

High-entropy regions. Let $\mathcal{P}$ be a convex set of probability distributions on $\mathcal{X}$, and let $H : \mathcal{P} \to \mathbb{R}$ be a concave entropy functional. For a target level $\rho \in \mathbb{R}$, define the entropy superlevel set

$$\mathcal{C}_\rho := \big\{Q \in \mathcal{P} : \; H(Q) \geq \rho\big\}. \tag{11}$$

By concavity of H, $\mathcal{C}_\rho$ is a convex set, i.e. for every $Q_1, Q_2 \in \mathcal{C}_\rho$ and $\alpha \in [0, 1]$, then $\alpha Q_1 + (1 - \alpha)Q_2 \in \mathcal{C}_\rho$. In discrete settings, H corresponds to Shannon entropy. In our kernelized setting, H corresponds to the population von Neumann entropy of the kernel covariance operator (population log-Vendi), which is concave in the distribution.

Projection principle. The key implication of the convexity of $\mathcal{C}_\rho$ distribution set is that projecting a learned distribution onto this high-entropy region will improve its proximity to any target distribution already in the region, under common discrepancy families used in generative modeling.

Theorem 1 (Entropy projection principle)**.** *Assume $\mathcal{C}_\rho$ is a nonempty and convex set. Let $P \in \mathcal{C}_\rho$ and let $Q_\theta \in \mathcal{P}$.*

(i) Hilbertian distances. *If* dist *is a Hilbertian distance (including MMD and KD), and $Q^\star$ is the metric projection of Q_θ onto $\mathcal{C}_\rho$, i.e. $Q^\star \in \arg\min_{Q \in \mathcal{C}_\rho} \mathrm{dist}(Q_\theta, Q)$, then*

$$\mathrm{dist}(P, Q^\star) \leq \mathrm{dist}(P, Q_\theta). \tag{12}$$

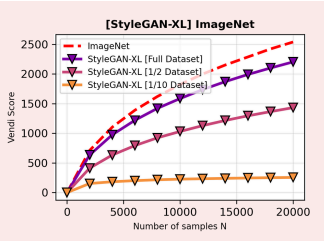
(ii) Bregman divergences. *If D_Φ is a Bregman divergence (in particular, the KL divergence), and $Q^\star$ is the corresponding I-projection $Q^\star \in \arg\min_{Q \in \mathcal{C}_\rho} D_\Phi(Q, Q_\theta)$, then*

$$D_\Phi(P, Q^\star) \leq D_\Phi(P, Q_\theta). \tag{13}$$

Theorem 1 identifies $\mathcal{C}_\rho$ as a meaningful high-entropy region: if the target distribution lies in $\mathcal{C}_\rho$, then moving Q_θ toward $\mathcal{C}_\rho$ cannot increase its discrepancy to P under standard training/evaluation distances. Formal definitions and proofs are provided in Appendix C.3. The projection process is illustrated in Figure 6.

6.1 Practical regularization, guidance, and empirical-support entropy projection

This subsection translates the projection principle of Theorem 1 into concrete, implementable procedures. We first discuss two mechanisms that act *during* model construction, that are training-time regularization for GAN-style objectives and sampling-time guidance for diffusion models. We then present a complementary mechanism that acts *post* sampling: an empirical-support entropy projection that reweights a fixed set of generated samples to satisfy an entropy (Vendi/VNE) or order-2 (RKE) diversity constraint while remaining close to the baseline empirical measure under a kernelized Hilbertian geometry. Detailed derivations are deferred to Appendix, Section B.



Method	Recall ↑	Coverage ↑	RKE ↑	Vendi ↑
Dataset	-	-	**36.82**	**2533.93**
StyleGAN-XL: full dataset	0.74	0.96	31.27	2203.39
StyleGAN-XL: 1/2 dataset	0.59	0.94	20.80	1432.20
StyleGAN-XL: 1/10 dataset	0.42	0.93	14.95	255.44

Method	Recall ↑	Coverage ↑	RKE ↑	Vendi ↑
Dataset	-	-	**14.46**	**514.02**
LDM: full dataset	0.44	0.74	10.21	294.60
LDM: 1/2 dataset	0.22	0.74	9.17	260.64
LDM: 1/4 dataset	0.03	0.52	7.06	170.44

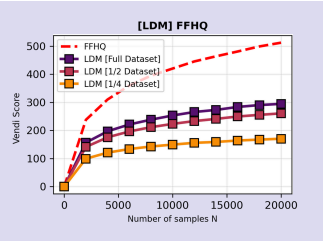


Figure 4: Vendi Score comparison across varying training set sizes. Results are shown for StyleGAN-XL (shaded red) and LDMs (shaded blue), with the training dataset baseline indicated by a dashed red line. Corresponding RKE, Recall, and Coverage metrics are also presented.

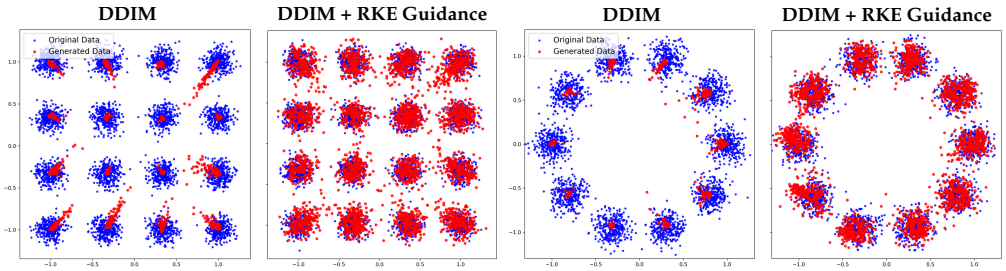


Figure 5: Comparison of the diffusion model (DDIM) with RKE regularized DDIM on an underlying 2D-Gaussian mixture model.

7 Numerical Results

Model comparison. Extending Figure 2, we evaluated diverse architectures on ImageNet, FFHQ, LSUN, and MSCOCO. Figure 3 confirms a systematic downward diversity bias: generative models consistently rank below real data across all sample sizes and embeddings. We provide additional results in the Appendix. To further validate our findings, we trained three versions of StyleGAN-XL using the full training set, 50%, and 10% of the data, respectively. As shown in Figure 4, model diversity drops as the training set size decreases. Notably, even when overfitted to smaller datasets, the model fails to replicate training samples sufficiently to recover diversity. Similarly, Latent Diffusion Models (LDMs) trained on the FFHQ dataset exhibit a comparable decline in diversity. To provide a comprehensive evaluation, we supplement the Vendi Score with RKE, Recall, and Coverage.

Entropy-based guidance for generative models. We evaluated the impact of RKE and Vendi regularization on generative models, demonstrating that guidance can effectively mitigate diversity biases while improving FD and KD scores. As shown in Figure 1, applying inverse-RKE guidance to the DiT-XL-2 diffusion model Peebles & Xie (2022) enhances both diversity and fidelity metrics. We observe similar results when guiding the DDIM model Song et al. (2020) on 2D Gaussian mixtures (Figure 5); increases in Vendi, RKE, and Recall confirm improved diversity, while better FD, KD, and Precision scores indicate a closer alignment with the dataset distribution. Extended quantitative results are provided in the Appendix.

8 Conclusion

We studied entropy-based diversity in generative models using reference-free metrics and showed that finite-sample entropy exhibits a systematic downward bias relative to the underlying data distribution. Empirically, real datasets consistently achieve higher Vendi/RKE diversity than samples from standard pretrained generative models under matched evaluation protocols, and increasing training-set size improves model diversity. These results suggest that finite-sample effects and training-data limitations can contribute to observed diversity gaps. Our numerical studies rely mainly on controlled synthetic settings and finite-sample evaluations, and while entropy-based regularization and guidance are theoretically motivated, we do not implement exact entropy projection operators due to computational and optimization challenges. Extending these ideas to scalable implementations and broader benchmarks remains a direction for future work.

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 14210725, and is partially supported by CUHK Direct Research Grants with CUHK Project No. 4055164 and 4937054. The work is also supported by a grant under 1+1+1 CUHK-CUHK(SZ)-GDSTC Joint Collaboration Fund.

REFERENCES

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. In *Proceedings of the International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 290–306. PMLR, 2022. URL https://proceedings.mlr.press/v162/alaa22a.html.

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

Sanjeev Arora, Andrej Risteski, and Yi Zhang. Do GANs learn the distribution? some theory and empirics. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=BJehNfW0-.

Reyhane Askari Hemmat, Melissa Hall, Alicia Sun, Candace Ross, Michal Drozdzal, and Adriana Romero-Soriano. Improving geo-diversity of generated images with contextualized vendi score guidance. In *European Conference on Computer Vision*, pp. 213–229. Springer, 2024.

Koenraad MR Audenaert. A sharp continuity estimate for the von neumann entropy. *Journal of Physics A: Mathematical and Theoretical*, 40(28):8127, 2007.

Francis Bach. Information theory with kernel methods. *arXiv preprint arXiv:2202.08545*, 2022. URL https://arxiv.org/abs/2202.08545.

G. P. Basharin. On a statistical estimate for the entropy of a sequence of independent random variables. *Theory of Probability and Its Applications*, 4(3):333–336, 1959.

Mikolaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.

Sam Bond-Taylor, Peter Hessey, Hiroshi Sasaki, Toby P Breckon, and Chris G Willcocks. Unleashing transformers: Parallel token prediction with discrete absorbing diffusion for fast high-resolution image generation from vector-quantized codes. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIII*, pp. 170–188. Springer, 2022.

Andrew Brock, Jeff Donahue, and K. Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations (ICLR)*, 2018.

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.

Gabriele Corso, Yilun Xu, Valentin De Bortoli, Regina Barzilay, and Tommi S Jaakkola. Particle guidance: non-iid diverse sampling with diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024.

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)*, pp. 248–255. IEEE, 2009.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021.

Mischa Dombrowski, Wenjia Zhang, Sonia Cechnicka, Hugo Reynaud, and Bernhard Kainz. Image generation diversity issues and how to tame them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3029–3039, 2025.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=g97OHbQyk1.

Ian J. 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*, 2014.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

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. URL https://arxiv.org/abs/2206.08549.

Louay Hazami, Rayhane Mama, and Ragavan Thurairatnam. Efficient-VDVAE: Less is more. *arXiv:2203.13751*, 2022.

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*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Xiaoyan Hu, Ho-fung Leung, and Farzan Farnia. A multi-armed bandit approach to online selection and evaluation of generative models. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025a.

Xiaoyan Hu, Ho-fung Leung, and Farzan Farnia. Pak-ucb contextual bandit: An online learning approach to prompt-aware selection of generative models and llms. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, 2025b.

Donya Jafari and Farzan Farnia. Diversity-aware online prompt assignment to generative models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=nnN2TKlS5C.

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, pp. 9931–9943, 2023.

Mohammad Jalali, LEI Haoyu, Amin Gohari, and Farzan Farnia. Sparke: Scalable prompt-aware diversity and novelty guidance in diffusion models via rke score. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Mohammad Jalali, Azim Ospanov, Amin Gohari, and Farzan Farnia. Conditional Vendi Score: Prompt-aware diversity evaluation for text-guided generative ai models. In *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, 2026.

Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 61(5):2835–2885, 2015. doi: 10.1109/TIT.2015.2412945. URL https://arxiv.org/abs/1406.6956.

Marco Jiralerspong, Avishek Joey Bose, Ian Gemp, Chongli Qin, Yoram Bachrach, and Gauthier Gidel. Feature likelihood divergence: Evaluating the generalization of generative models using samples. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://openreview.net/forum?id=l2VKZkolT7.

Jaehun Jung, Seungju Han, Ximing Lu, Skyler Hallinan, David Acuna, Shrimai Prabhumoye, Mostafa Patwary, Mohammad Shoeybi, Bryan Catanzaro, and Yejin Choi. Prismatic synthesis: Gradient-based data diversification boosts generalization in LLM reasoning. *arXiv preprint arXiv:2505.20161*, 2025. doi: 10.48550/arXiv.2505.20161. URL https://arxiv.org/abs/2505.20161.

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.

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)*, pp. 4401–4410, 2019.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12104–12114, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/8d30aa96e72440759f74bd2306c1fa3d-Paper.pdf.

Jaeill Kim, Suhyun Kang, Duhun Hwang, Jungwook Shin, and Wonjong Rhee. VNE: An effective method for improving deep representation by manipating eigenvalue distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3799–3810, June 2023.

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

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

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.

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*, pp. 11523–11532, 2022.

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 *European Conference on Computer Vision*, pp. 740–755, 2014.

George A. Miller. Note on the bias of information estimates. In Henry Quastler (ed.), *Information Theory in Psychology: Problems and Methods*, pp. 95–100. Free Press, 1955.

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 *Proceedings of Machine Learning Research*, pp. 7176–7185. PMLR, 2020.

Quan Nguyen and Adji Bousso Dieng. Quality-weighted vendi scores and their application to diverse experimental design. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 37667–37682. PMLR, 2024. URL https://proceedings.mlr.press/v235/nguyen24d.html.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, volume 139, pp. 8162–8171. PMLR, 2021.

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 and Farzan Farnia. Do vendi scores converge with finite samples? truncated vendi score for finite-sample convergence guarantees. In *The 41st Conference on Uncertainty in Artificial Intelligence*, 2025.

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, 2024.

Azim Ospanov, Mohammad Jalali, and Farzan Farnia. Scendi score: Prompt-aware diversity evaluation via schur complement of clip embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16927–16937, October 2025.

Liam Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, 2003.

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

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*, pp. 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: Optimal mixtures of generative models via mixture-UCB bandit algorithms. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=2Chkk5Ye2s.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M Weber. Cads: Unleashing the diversity of diffusion models through condition-annealed sampling. In *The Twelfth International Conference on Learning Representations*, 2024.

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, 2018.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://arxiv.org/abs/1606.03498.

Luis G. Sanchez Giraldo, Murali Rao, and Jose C. Principe. Measures of entropy from data using infinitely divisible kernels. *arXiv preprint arXiv:1211.2459*, 2012. URL https://arxiv.org/abs/1211.2459.

Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected GANs converge faster. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17480–17492, 2021.

Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. *arXiv preprint arXiv:2202.00273*, 2022.

Thomas Schürmann. Bias analysis in entropy estimation. *Journal of Physics A: Mathematical and General*, 37(27):L295–L301, 2004.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. URL https://arxiv.org/abs/2508.10104.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 3732–3784. Curran Associates, Inc., 2023.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

Paul Valiant and Gregory Valiant. Estimating the unseen: Improved estimators for entropy and other properties. In *Advances in Neural Information Processing Systems*, 2013.

Steven Walton, Ali Hassani, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. StyleNAT: Giving each head a new perspective. *arXiv:2211.05770*, 2022.

Zhendong Wang, Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Diffusion-GAN: Training GANs with diffusion. *arXiv:2206.02262*, 2022.

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

Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Jingwei Zhang, Cheuk Ting Li, and Farzan Farnia. An interpretable evaluation of entropy-based novelty of generative models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhang24ac.html.

Jingwei Zhang, Mohammad Jalali, Cheuk Ting Li, and Farzan Farnia. Unveiling differences in generative models: A scalable differential clustering approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

A Related Works

Diversity and evaluation of generative models. Evaluation of deep generative models has been extensively studied through sample-based metrics, including the Inception Score (Salimans et al., 2016) and embedding-based distances such as FID (Heusel et al., 2017) and KID (Bińkowski et al., 2018). To separate fidelity from coverage, precision–recall-style metrics were proposed by Sajjadi et al. (2018) and refined by improved precision/recall estimators based on manifold estimation (Kynkäänniemi et al., 2019), with related density and coverage scores by Naeem et al. (2020). The metrics have been further updated by employing stronger backbone embeddings, including CLIP (Kynkäänniemi et al., 2023) and DINOv2 (Stein et al., 2023). The prompt-aware evaluation of diversity in conditional prompt-guided generative models has also been studied by Ospanov et al. (2025) and Jalali et al. (2026). A complementary perspective on diversity limitations is provided by the birthday-paradox-based test of Arora et al. (2018), which estimates the effective support size of a model's distribution, suggesting that (at the time) GAN samples could output relatively small support sets. Recent work by Dombrowski et al. (2025) also highlights that current image generative models can exhibit substantial diversity limitations that are not well captured by common evaluation metrics, and proposes an image-retrieval-based diversity measure together with diversity-aware diffusion modeling. Furthermore, the recent works (Hu et al., 2025a; Rezaei et al., 2025; Hu et al., 2025b; Jafari & Farnia, 2026) consider the diversity-based scores in an online multi-armed bandit setting and propose diversity-aware online learning methods to address the model selection task.

Reference-free diversity metrics and kernel-entropy scores. Reference-free diversity measures quantify variability within a set of generated samples without requiring an explicit reference dataset. The Vendi score (Friedman & Dieng, 2023) defines diversity as the exponential of the von Neumann entropy of a kernel similarity matrix, and Rényi Kernel Entropy (RKE) (Jalali et al., 2023) provides an order-2 Rényi analogue with tractable structure and mode-count interpretations in certain settings. These kernel-entropy ideas have also been used to steer generation: contextualized Vendi score guidance (VSG) (Askari Hemmat et al., 2024) injects Vendi-based gradients into latent diffusion sampling to encourage diversity (under the chosen kernel/representation) while controlling drift. Related extensions incorporate quality into diversity scoring, e.g., quality-weighted Vendi (Nguyen & Dieng, 2024), and recent work proposes gradient-space entropy measures such as G-Vendi (Jung et al., 2025), linking diversity of data to generalization in downstream reasoning tasks.

Novelty, memorization, rarity, and fine-grained diagnostics. Beyond diversity, several works target novelty and memorization behavior of generative models. Alaa et al. (2022) propose sample-level auditing via the Authenticity score. Jiralerspong et al. (2023) propose Feature Likelihood Divergence, a parametric feature-space likelihood approach to jointly evaluate novelty, fidelity, and diversity from samples. Han et al. (2022) propose the Rarity score to quantify per-sample uncommonness. Also, Zhang et al. (2024; 2025) propose the Kernel-based Entropic Novelty score for identifying sample types generated more frequently.

Diversity in diffusion models. Recent work has examined inference-time mechanisms for improving diversity in diffusion models. Sadat et al. (2024) propose CADS which injects scheduled noise into the conditioning signal during inference to trade off conditioning strength and diversity. Corso et al. (2024) develop Particle Guidance for non-i.i.d. diversity-aware sampling, guiding a set of simultaneously generated samples to increase diversity via particle-style interactions. Also, Vendi and RKE-based guidance by Askari Hemmat et al. (2024); Jalali et al. (2025) use matrix-based entropy as an explicit diversity objective. Our work examines the role of entropy-guided methods in addressing diversity biases in generative models.

Entropy, kernel operators, and bias in entropy estimation. Kernel/operator entropy functionals connect generative-model diversity evaluation to information theory and quantum information, where von Neumann entropy is defined on positive semidefinite operators. In machine learning, matrix-based and kernel-based entropy estimators were developed to avoid explicit density estimation, including kernel constructions motivated by Rényi-style axioms (Sanchez Giraldo et al., 2012) and RKHS information-theoretic frameworks based on covariance operators (Bach, 2022). Separately, classical statistics shows that plug-in entropy estimation from finite samples can exhibit systematic downward bias, starting from early bias analyses (Miller, 1955) and nonparametric treatments (Paninski, 2003), with modern minimax-optimal estimation of distribution functionals including entropy (Jiao et al., 2015) and improved estimators for the unseen regime (Valiant & Valiant,

2013). These results inspire the study of finite-sample effects when diversity is quantified through entropy-like functionals.

B Detailed Methods for Practical regularization, guidance, and empirical-support entropy projection

Training-time regularization. A direct way to promote entropy-based diversity is to augment a divergence-based training objective $\mathcal{L}(\theta)$ with an entropy term,

$$\min_{\theta} \; \mathcal{L}(\theta) - \lambda\, H_{\mathrm{VNE}}(Q_\theta), \qquad \lambda > 0. \tag{14}$$

In our kernelized setting, $H_{\mathrm{VNE}}(Q_\theta)$ corresponds to the population log-Vendi, i.e., the von Neumann entropy (VNE) of a kernel covariance operator. Since VNE involves eigendecomposition, we also consider the order-2 surrogate inverse-RKE (Jalali et al., 2023):

$$\text{Inverse-RKE}(Q_\theta) := \mathbb{E}_{X,X'\sim Q_\theta}\big[k(X,X')^2\big], \tag{15}$$

Related to this proposal, Kim et al. (2023) empirically show that the kernel von Neumann entropy (equivalently the Vendi score) regularization of representation autocorrelation matrices provides an effective mechanism in training GANs, further supporting the relevance of spectral entropy functionals in improving the generative modeling performance.

Sampling-time guidance (diffusion/latent diffusion). For diffusion and latent diffusion models, entropy-aligned corrections can be applied along the reverse-time trajectory. Let $\mathcal{S}_\psi$ produce an intermediate proposal $\tilde{z}_{t-1} = \mathcal{S}_\psi(z_t, t)$. Given a memory bank $\mathcal{M} = \{z^{(i)}\}_{i=1}^m$, define

$$\mathcal{J}_{\text{Inverse-RKE}}(z;\mathcal{M}) := \frac{1}{m}\sum_{i=1}^{m} k(z, z^{(i)})^2, \tag{16}$$

leading to the guided update

$$z_{t-1} = \tilde{z}_{t-1} - \eta_t \nabla_z \mathcal{J}_{\text{Inverse-RKE}}(\tilde{z}_{t-1};\mathcal{M}), \tag{17}$$

with $\eta_t \geq 0$. This is the prompt-free analogue of SPARKE-guidance (Jalali et al., 2025); related Vendi/VNE-based guidance has also been explored in (Askari Hemmat et al., 2024).

Empirical-support post-hoc projection via reweighting. Beyond training-time regularization and sampling-time guidance, we consider a post-hoc mechanism that directly instantiates the geometric "projection-to-high-entropy" perspective on a fixed sample set. Given M samples $\{x_i\}_{i=1}^M$ generated from a trained model Q_θ, we form a reweighted empirical measure

$$Q_q = \sum_{i=1}^{M} q_i\, \delta_{x_i}, \qquad q \in \Delta_M := \Big\{q \geq 0,\; \sum_{i=1}^{M} q_i = 1\Big\}, \qquad q_0 = \frac{1}{M}\mathbf{1}.$$

This reweighting preserves the support and only redistributes probability mass across already-generated samples.

The projection viewpoint suggests choosing q to (i) remain close to the baseline empirical model Q_{q_0} under a Hilbertian discrepancy (e.g., MMD/KD), while (ii) enforcing a target diversity level. Concretely, we consider the empirical-support projection program

$$\begin{aligned} \min_{q\in\Delta_M} \quad & \mathrm{KD}\,(Q_q, Q_{q_0}) := (q-q_0)^\top K (q-q_0) \\ \text{s.t.} \quad & H_{\mathrm{VNE}}(Q_q) := H_{\mathrm{VNE}}\big(\mathrm{diag}(q)^{1/2} K\, \mathrm{diag}(q)^{1/2}\big) \;\geq\; \rho, \end{aligned} \tag{18}$$

where KD denotes the kernel distance induced by k (equivalently, squared MMD on the fixed support), which for the kernel matrix K of generated samples $x_1, \ldots, x_n$ will be $(q-q_0)^\top K(q-q_0)$. The above formulation also motivates the following Lagrangian log-Vendi H_{VNE} penalty variant for a regularization coefficient $\lambda > 0$:

$$\min_{q\in\Delta_M} \quad \mathrm{KD}\,(Q_q, Q_{q_0}) \;-\; \lambda\, H_{\mathrm{VNE}}(Q_q) \tag{19}$$

As an efficient order-2 alternative aligned with the RKE score, one can also use a penalized variant based on the inverse-RKE score,

$$\min_{q \in \Delta_M} \quad \mathrm{KD}\left(Q_q, Q_{q_0}\right) \; + \; \lambda \, \text{Inverse-RKE}(Q_q) \tag{20}$$

Both formulations implement the same principle: adjust the probability mass over the generated samples to reach a higher-diversity region while staying close to the original empirical model. Full technical definitions, convexity properties, and optimization algorithms are presented in Appendix D.

C PROOFS

C.1 PROOF OF PROPOSITION 1

The von Neumann entropy is concave on the set of positive semidefinite, unit-trace operators (set of density matrices). Fix $n \geq 1$ and draw $X_1, \ldots, X_{n+1} \sim P$ i.i.d. Let $Z_i := \phi(X_i)\phi(X_i)^\top$. Consider a uniformly random subset $S \subset \{1, \ldots, n+1\}$ of size n, independent of the data, and define $\widehat{C}_S = \frac{1}{n}\sum_{i \in S} Z_i$. Conditioned on $(Z_1, \ldots, Z_{n+1})$, we have $\mathbb{E}[\widehat{C}_S] = \widehat{C}_{n+1}$. Therefore, by the concavity of the entropy function, we can apply Jensen's inequality to obtain

$$H(\widehat{C}_{n+1}) \geq \mathbb{E}\left[H(\widehat{C}_S) \,\middle|\, Z_1, \ldots, Z_{n+1}\right].$$

Taking expectation and noting that $\widehat{C}_S$ has the same marginal distribution as $\widehat{C}_n$ completes the proof.

C.2 PROOF OF CONCENTRATION OF log-Vendi SCORE AROUND ITS FINITE-SAMPLE EXPECTATION

Here, we show that a single-sample estimate of log-Vendi (the von Neumann entropy of the empirical kernel covariance) concentrates around its expectation under i.i.d. sampling. This result provides justification on reporting averaged Vendi/RKE scores over multiple random subsets at a fixed evaluation size.

To review the notation in the result, let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a normalized bounded kernel with $k(x, x) = 1$ and $|k(x, x')| \leq 1$ for all x, x'. Let $\phi : \mathcal{X} \to \mathcal{H}$ be a feature map into a real Hilbert space such that $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$. For $m \geq 2$ and i.i.d. samples $X_{1:m} \sim P$, define the empirical kernel covariance operator

$$\widehat{C}_m := \frac{1}{m}\sum_{i=1}^{m} \phi(X_i)\phi(X_i)^\top,$$

and the corresponding log-Vendi and Vendi scores

$$\mathsf{Ent}_m := H(\widehat{C}_m), \qquad \mathsf{Vendi}_m := \exp(\mathsf{Ent}_m), \qquad H(\rho) := -\mathrm{Tr}(\rho \log \rho).$$

Note that we define the stability constant c_m as follows

$$c_m := \frac{1}{m}\log(2m-1) + h\left(\frac{1}{m}\right), \qquad h(t) := -t\log t - (1-t)\log(1-t). \tag{21}$$

As we show in the theorem's proof, it can be seen that $c_m \leq \frac{\log(em)}{m}$ holds for every integer m.

Proposition 2. *Fix $m \geq 2$ and let $X_{1:m} \overset{\text{i.i.d.}}{\sim} P$. Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\left|\mathsf{Ent}_m - \mathbb{E}[\mathsf{Ent}_m]\right| \leq c_m \sqrt{\frac{m}{2}\log\left(\frac{2}{\delta}\right)} \leq \sqrt{\frac{\log^2(em)\log(2/\delta)}{2m}}. \tag{22}$$

Equivalently, letting $t_\delta := \sqrt{\frac{\log^2(em)\log(2/\delta)}{2m}}$, with probability at least $1 - \delta$, the following holds

$$e^{\mathbb{E}[\mathsf{Ent}_m] - t_\delta} \leq \mathsf{Vendi}_m \leq e^{\mathbb{E}[\mathsf{Ent}_m] + t_\delta}. \tag{23}$$

Proof. To apply McDiarmid's inequality, we define the function

$$f(x_1, \ldots, x_m) := H\left(\frac{1}{m}\sum_{i=1}^{m}\phi(x_i)\phi(x_i)^\top\right).$$

We will show that replacing any single input changes f by at most c_m.

Fix an index $j \in [m]$ and consider two input sequences that differ only at the j-th coordinate: $(x_1, \ldots, x_j, \ldots, x_m)$ and $(x_1, \ldots, x'_j, \ldots, x_m)$. Let $\widehat{C}$ and $\widehat{C}'$ be the corresponding empirical covariance operators. Write $u = \phi(x_j)$ and $v = \phi(x'_j)$. Since k is normalized, $\|u\|_{\mathcal{H}}^2 = k(x_j, x_j) = 1$ and similarly $\|v\|_{\mathcal{H}} = 1$. Then

$$\widehat{C} - \widehat{C}' = \frac{1}{m}\big(uu^\top - vv^\top\big).$$

Each rank-one operator has trace norm $\|uu^\top\|_1 = \mathrm{Tr}(uu^\top) = \|u\|_{\mathcal{H}}^2 = 1$ (and similarly for $vv^\top$), hence by the triangle inequality

$$\|uu^\top - vv^\top\|_1 \le \|uu^\top\|_1 + \|vv^\top\|_1 = 2 \quad \Longrightarrow \quad \|\widehat{C} - \widehat{C}'\|_1 \le \frac{2}{m}.$$

Let

$$\Delta := \frac{1}{2}\|\widehat{C} - \widehat{C}'\|_1 \le \frac{1}{m}.$$

Both $\widehat{C}$ and $\widehat{C}'$ are PSD with unit trace and rank at most m, since each is an average of m rank-one PSD operators. Let d be the dimension of the subspace spanned by $\mathrm{supp}(\widehat{C}) \cup \mathrm{supp}(\widehat{C}')$. Then

$$d \le \mathrm{rank}(\widehat{C}) + \mathrm{rank}(\widehat{C}') \le 2m.$$

Von Neumann entropy depends only on the nonzero spectrum, thus we may restrict both operators to this d-dimensional subspace. Applying Audenaert's continuity bound for von Neumann entropy Audenaert (2007) results in the following inequality:

$$\big|H(\widehat{C}) - H(\widehat{C}')\big| \le \Delta \log(d-1) + h(\Delta). \tag{24}$$

Using $d \le 2m$ and $\Delta \le 1/m$, and the facts that $\log(d-1)$ is increasing in d and $h(\Delta)$ is increasing for $\Delta \in [0, 1/2]$ (here $\Delta \le 1/m \le 1/2$ since $m \ge 2$), we obtain

$$|H(\widehat{C}) - H(\widehat{C}')| \le \frac{1}{m}\log(2m-1) + h\left(\frac{1}{m}\right) = c_m.$$

Next, we apply McDiarmid's inequality to the defined function. Let $\mathsf{Ent}_m = f(X_1, \ldots, X_m)$. By McDiarmid's inequality, for any $t > 0$,

$$\mathbb{P}\left(\big|\mathsf{Ent}_m - \mathbb{E}[\mathsf{Ent}_m]\big| \ge t\right) \le 2\exp\left(-\frac{2t^2}{\sum_{j=1}^{m} c_j^2}\right) = 2\exp\left(-\frac{2t^2}{mc_m^2}\right).$$

Setting the right-hand side equal to δ yields $t = c_m\sqrt{\frac{m}{2}\log\left(\frac{2}{\delta}\right)}$, which proves equation 22.

To convert log-Vendi concentration to Vendi concentration, we note that the bound equation 22 implies $\mathbb{E}[\mathsf{Ent}_m] - t_\delta \le \mathsf{Ent}_m \le \mathbb{E}[\mathsf{Ent}_m] + t_\delta$ with probability at least $1 - \delta$. Exponentiating all terms gives equation 23.

Also, for a simple upper bound on $h(1/m)$, we note that for $0 < p \le 1/2$, a standard inequality is $h(p) \le p\log(e/p)$. Taking $p = 1/m$ (valid for $m \ge 2$) gives

$$h\big(\frac{1}{m}\big) \le \frac{1}{m}\log(em).$$

The proof is therefore complete. □

C.3 PROOF OF THEOREM 1

We first prove the following lemma on the convexity of the entropy's super-level sets and then prove Parts (i) and (ii) separately.

Lemma 1. *Let $\mathcal{P}$ be a convex set and let $H : \mathcal{P} \to \mathbb{R}$ be concave. For any $\rho \in \mathbb{R}$, the set*

$$\mathcal{C}_\rho := \{Q \in \mathcal{P} : H(Q) \geq \rho\}$$

is a convex set.

Proof. Take any $Q_1, Q_2 \in \mathcal{C}_\rho$ and any $\alpha \in [0, 1]$. Since $\mathcal{P}$ is convex, $Q_\alpha := \alpha Q_1 + (1-\alpha)Q_2 \in \mathcal{P}$. By concavity of H,

$$H(Q_\alpha) \;\geq\; \alpha H(Q_1) + (1-\alpha)H(Q_2) \;\geq\; \alpha\rho + (1-\alpha)\rho \;=\; \rho$$

Therefore, $Q_\alpha \in \mathcal{C}_\rho$. □

(I) HILBERTIAN DISTANCES

We first review the definitions of Hilbertian distance and Hilbertian distance projection.

We call dist *Hilbertian* on $\mathcal{P}$ if there exist a Hilbert space $\mathcal{H}$ and an affine map $\mu : \mathcal{P} \to \mathcal{H}$ such that for all $P, Q \in \mathcal{P}$,

$$\mathrm{dist}(P, Q) = \|\mu(P) - \mu(Q)\|_{\mathcal{H}}.$$

Also, for the Hilbertian distance projection, we assume $\mathcal{C}_\rho$ is nonempty and convex, and that $\mu(\mathcal{C}_\rho)$ is closed in $\mathcal{H}$ (equivalently, $\mathcal{C}_\rho$ is closed under dist). Then for every $Q_\theta \in \mathcal{P}$, Hilbertian distance projection is

$$Q^\star \in \arg\min_{Q \in \mathcal{C}_\rho} \mathrm{dist}(Q_\theta, Q).$$

Lemma 2 (Pythagorean inequality for Hilbert projections)**.** *Let $C \subseteq \mathcal{H}$ be nonempty, closed, and convex. Let $x^\star = \Pi_C(x)$ denote the metric projection of x onto C. Then for every $y \in C$,*

$$\|y - x\|_{\mathcal{H}}^2 \;\geq\; \|y - x^\star\|_{\mathcal{H}}^2 + \|x - x^\star\|_{\mathcal{H}}^2.$$

Proof. It can be seen a metric projection in Hilbert spaces satisfies the following: $x^\star = \Pi_C(x)$ if and only if

$$\langle x - x^\star,\ y - x^\star\rangle_{\mathcal{H}} \leq 0 \quad \text{for all } y \in C.$$

We then expand the difference norm squared as follows

$$\|y - x\|^2 = \|y - x^\star\|^2 + \|x - x^\star\|^2 + 2\langle y - x^\star,\ x^\star - x\rangle.$$

The last inner product is ≤ 0 by the projection characterization, and thus the inequality follows. □

To derive equation 12, we let $C = \mu(\mathcal{C}_\rho) \subseteq \mathcal{H}$, $x = \mu(Q_\theta)$, $x^\star = \mu(Q^\star)$, and $y = \mu(P)$ (note $P \in \mathcal{C}_\rho$ implies $y \in C$). Applying Lemma 2 results in the following

$$\|y - x\|^2 \geq \|y - x^\star\|^2 + \|x - x^\star\|^2,$$

and hence in particular $\|y - x^\star\| \leq \|y - x\|$. Translating back using $\mathrm{dist}(P, Q) = \|\mu(P) - \mu(Q)\|_{\mathcal{H}}$ yields

$$\mathrm{dist}(P, Q^\star) \leq \mathrm{dist}(P, Q_\theta).$$

(II) BREGMAN DIVERGENCES

We first review the definition of Bregman divergence and I-Projection.

Let Ω be a convex set in a real vector space and let $\Phi : \Omega \to \mathbb{R}$ be differentiable and strictly convex. The associated Bregman divergence is

$$D_\Phi(p, q) := \Phi(p) - \Phi(q) - \langle \nabla\Phi(q),\, p - q\rangle.$$

Regarding the I-projection, we assume $\mathcal{C}_\rho$ is nonempty and convex, $\mathcal{C}_\rho \subseteq \Omega$, and that $\mathcal{C}_\rho$ is closed and the function $p \mapsto D_\Phi(p, Q_\theta)$ is coercive on $\mathcal{C}_\rho$ (for example, if $\mathcal{C}_\rho$ is compact). Then, the I-Projection follows from

$$Q^\star \in \arg\min_{Q \in \mathcal{C}_\rho} D_\Phi(Q, Q_\theta).$$

Lemma 3. *For any $p, u, q \in \Omega$, the following hold*

$$D_\Phi(p, q) = D_\Phi(p, u) + D_\Phi(u, q) + \langle \nabla\Phi(u) - \nabla\Phi(q),\, p - u\rangle.$$

Proof. We expand each D_Φ term using the definition as follows:

$$\begin{aligned} D_\Phi(p, u) + D_\Phi(u, q) &= \Phi(p) - \Phi(u) - \langle \nabla\Phi(u), p - u\rangle + \Phi(u) - \Phi(q) - \langle \nabla\Phi(q), u - q\rangle \\ &= \Phi(p) - \Phi(q) - \langle \nabla\Phi(q), p - q\rangle + \langle \nabla\Phi(q) - \nabla\Phi(u), p - u\rangle \\ &= D_\Phi(p, q) + \langle \nabla\Phi(q) - \nabla\Phi(u), p - u\rangle, \end{aligned}$$

which is equivalent to the stated identity. □

Lemma 4 (Bregman Pythagorean inequality for I-projections)**.** *Let $C \subseteq \Omega$ be nonempty and convex, and let $u \in \arg\min_{v \in C} D_\Phi(v, q)$ be an I-projection of q onto C. Then for every $p \in C$,*

$$D_\Phi(p, q) \;\geq\; D_\Phi(p, u) + D_\Phi(u, q).$$

Proof. We first establish the first-order optimality condition: since $v \mapsto D_\Phi(v, q)$ is convex and differentiable with gradient $\nabla_v D_\Phi(v, q) = \nabla\Phi(v) - \nabla\Phi(q)$, optimality of u over the convex set C implies

$$\langle \nabla\Phi(u) - \nabla\Phi(q),\, p - u\rangle \geq 0 \qquad \text{for all } p \in C. \tag{25}$$

Then, we apply Lemma 3 with (p, u, q) to obtain

$$D_\Phi(p, q) = D_\Phi(p, u) + D_\Phi(u, q) + \langle \nabla\Phi(u) - \nabla\Phi(q),\, p - u\rangle.$$

Using equation 25, the inner product term is nonnegative, hence $D_\Phi(p, q) \geq D_\Phi(p, u) + D_\Phi(u, q)$. □

To derive equation 13, we apply Lemma 4 with $C = \mathcal{C}_\rho$, $q = Q_\theta$, and $u = Q^\star$. For any $P \in \mathcal{C}_\rho$,

$$D_\Phi(P, Q_\theta) \;\geq\; D_\Phi(P, Q^\star) + D_\Phi(Q^\star, Q_\theta) \;\geq\; D_\Phi(P, Q^\star),$$

which gives $D_\Phi(P, Q^\star) \leq D_\Phi(P, Q_\theta)$, i.e. equation 13.

Finally, note that on the probability simplex, forward KL divergence $\mathrm{KL}(p\|q)$ is the Bregman divergence generated by $\Phi(p) = \sum_i p_i \log p_i$ (negative Shannon entropy). Therefore the above I-projection guarantee applies to the KL-divergence. □

D Empirical-support entropy projection: VNE and RKE

This appendix provides the technical details for the empirical-support post-hoc projection described in Section 6.1. The goal is to reweight a fixed set of M generated samples to satisfy a target diversity level while staying close (in a Hilbertian geometry) to the baseline uniform empirical distribution.

D.1 Setup: reweighted empirical measures and kernel geometry

Let $\{x_i\}_{i=1}^M \subset \mathcal{X}$ be samples generated from a trained model distribution Q_θ. For any weight vector $q \in \Delta_M := \{q \geq 0,\ \sum_{i=1}^M q_i = 1\}$ define the reweighted empirical measure

$$Q_q := \sum_{i=1}^M q_i\, \delta_{x_i}, \qquad q_0 := \tfrac{1}{M}\mathbf{1}.$$

Let k be a positive semidefinite kernel with Gram matrix $K \in \mathbb{R}^{M \times M}$, $K_{ij} = k(x_i, x_j)$.

Squared MMD/KD on a fixed support. For two weighted empirical measures Q_q and $Q_{q'}$ supported on the same points $\{x_i\}$, the squared MMD (equivalently KD in our notation) with kernel k satisfies

$$\mathrm{KD}^2(Q_q, Q_{q'}) = \sum_{i,j} (q_i - q'_i)(q_j - q'_j) k(x_i, x_j) = (q - q')^\top K (q - q'). \tag{26}$$

In particular, $\mathrm{KD}^2(Q_q, Q_{q_0}) = (q - q_0)^\top K (q - q_0)$ is a convex quadratic function of q.

D.2 VNE (LOG-VENDI) ON EMPIRICAL SUPPORT

Let $k(x, y) = \langle \phi(x), \phi(y) \rangle$ be represented by a feature map ϕ into a (possibly infinite-dimensional) Hilbert space. Define the empirical-support covariance operator

$$C(q) := \sum_{i=1}^{M} q_i \, \phi(x_i)\phi(x_i)^\top. \tag{27}$$

Assume the kernel is normalized on the support: $k(x_i, x_i) = \|\phi(x_i)\|^2 = 1$ for all i. Then $\mathrm{tr}(C(q)) = \sum_i q_i = 1$.

Empirical-support VNE. Define the von Neumann entropy

$$H_{\mathrm{VNE}}(q) := -\mathrm{tr}\big(C(q) \log C(q)\big), \tag{28}$$

with the convention $0 \log 0 := 0$.

Proposition 3 (Concavity of empirical-support VNE in q)**.** *Under the above normalization, $q \mapsto H_{\mathrm{VNE}}(q)$ is concave on Δ_M. Consequently, for any $\rho \in \mathbb{R}$, the following superlevel set $\mathcal{C}_\rho^{\mathrm{VNE}}$ is a convex set:*

$$\mathcal{C}_\rho^{\mathrm{VNE}} := \{q \in \Delta_M : \; H_{\mathrm{VNE}}(q) \geq \rho\}$$

Proof. To prove the statement, let $\mathcal{D} := \{X \succeq 0 : \mathrm{tr}(X) = 1\}$ and define $S(X) := -\mathrm{tr}(X \log X)$ on $\mathcal{D}$. We show S is concave on $\mathcal{D}$ via a variational identity.

For any Hermitian H, define $\sigma_H := e^H / \mathrm{tr}(e^H)$. For $X \in \mathcal{D}$, the quantum relative entropy $D(X \| \sigma_H) := \mathrm{tr}(X(\log X - \log \sigma_H))$ satisfies $D(X \| \sigma_H) \geq 0$. Using $\log \sigma_H = H - \log \mathrm{tr}(e^H)$, we obtain

$$D(X \| \sigma_H) = \mathrm{tr}(X \log X) - \mathrm{tr}(XH) + \log \mathrm{tr}(e^H) \geq 0,$$

hence $S(X) \leq \log \mathrm{tr}(e^H) - \mathrm{tr}(XH)$ for all H. Equality is achieved by taking $\sigma_H = X$ (equivalently $H = \log X$ on the support of X), yielding

$$S(X) = \inf_{H = H^\top} \big\{ \log \mathrm{tr}(e^H) - \mathrm{tr}(XH) \big\}.$$

For each fixed H, the map $X \mapsto \log \mathrm{tr}(e^H) - \mathrm{tr}(XH)$ is affine; an infimum of affine functions is concave, hence S is concave on $\mathcal{D}$.

Since $q \mapsto C(q)$ is affine by equation 27 and $C(q) \in \mathcal{D}$ for all $q \in \Delta_M$ by normalization, the composition $q \mapsto S(C(q)) = H_{\mathrm{VNE}}(q)$ is concave. Convexity of the superlevel set follows. □

Gradient with respect to q. Differentiating through $C(q)$ gives the Fréchet differential identity

$$\mathrm{d}H_{\mathrm{VNE}}(q) = -\mathrm{tr}\big((\log C(q) + I) \, \mathrm{d}C(q)\big), \tag{29}$$

and since $\partial C(q) / \partial q_i = \phi(x_i)\phi(x_i)^\top$,

$$\frac{\partial}{\partial q_i} H_{\mathrm{VNE}}(q) = -\big\langle \phi(x_i), (\log C(q) + I)\phi(x_i) \big\rangle, \qquad i \in [M]. \tag{30}$$

In practice, $\log C(q)$ is computed on the support of $C(q)$ (equivalently using a small spectral floor), which is standard for stable entropy computations.

Finite-dimensional evaluation via weighted Gram matrices. Let $K = \Phi\Phi^\top$ for a (possibly implicit) feature matrix Φ with rows $\phi(x_i)^\top$. Define

$$A(q) := \mathrm{diag}(\sqrt{q}) \, K \, \mathrm{diag}(\sqrt{q}). \tag{31}$$

Then $A(q)$ and $C(q)$ share the same nonzero eigenvalues, so $H_{\mathrm{VNE}}(q)$ can be computed from an eigendecomposition of $A(q)$, and equation 30 can be evaluated using the corresponding eigenbasis.

Algorithm 1: Empirical-support VNE projection (primal–dual exponentiated gradient)

Input: K; target ρ; stepsizes $\eta, \gamma > 0$; iterations T
Output: $q_T \in \Delta_M$

Initialize $q_0 \leftarrow \frac{1}{M}\mathbf{1}$ and $\lambda_0 \leftarrow 0$;
for $t = 0, 1, \ldots, T-1$ **do**
  Compute $H_{\mathrm{VNE}}(q_t)$ and $\nabla_q H_{\mathrm{VNE}}(q_t)$ using an eigendecomposition of $A(q_t) = \mathrm{diag}(\sqrt{q_t})K\mathrm{diag}(\sqrt{q_t})$;
  $g_t \leftarrow 2K(q_t - q_0) - \lambda_t \nabla_q H_{\mathrm{VNE}}(q_t)$;
  $\tilde{q}_{t+1,i} \leftarrow q_{t,i}\exp(-\eta\, g_{t,i})$ for all i; $\quad q_{t+1} \leftarrow \tilde{q}_{t+1}/\sum_j \tilde{q}_{t+1,j}$;
  $\lambda_{t+1} \leftarrow \left[\lambda_t + \gamma(\rho - H_{\mathrm{VNE}}(q_{t+1}))\right]_+$;
return q_T;

D.3 Empirical-support VNE projection problem and solver

We define the empirical-support VNE projection as

$$\min_{q \in \Delta_M} (q - q_0)^\top K (q - q_0) \qquad \text{s.t.} \qquad H_{\mathrm{VNE}}(q) \geq \rho. \tag{32}$$

By Proposition 3 and PSD-ness of K, this is a convex optimization problem over the simplex.

A practical solver is primal–dual KL-mirror descent (exponentiated gradient) on the Lagrangian

$$\mathcal{F}(q, \lambda) = (q - q_0)^\top K (q - q_0) + \lambda(\rho - H_{\mathrm{VNE}}(q)), \qquad \lambda \geq 0.$$

The primal gradient is

$$\nabla_q \mathcal{F}(q, \lambda) = 2K(q - q_0) - \lambda \nabla_q H_{\mathrm{VNE}}(q),$$

with $\nabla_q H_{\mathrm{VNE}}(q)$ from equation 30.

D.4 Order-2 (RKE) empirical-support projection

Define the squared-kernel Gram matrix

$$\widetilde{K}_{ij} := k(x_i, x_j)^2, \qquad \widetilde{K} := K \circ K,$$

where $\circ$ denotes the Hadamard product. For $Q_q = \sum_i q_i \delta_{x_i}$,

$$\text{Inverse-RKE}(Q_q) = \mathbb{E}_{X,X' \sim Q_q}[k(X, X')^2] = \sum_{i,j} q_i q_j k(x_i, x_j)^2 = q^\top \widetilde{K} q. \tag{33}$$

Since $K \succeq 0$ implies $\widetilde{K} = K \circ K \succeq 0$ (Schur product theorem), $q \mapsto q^\top \widetilde{K} q$ is convex on Δ_M.

A convenient penalized formulation is the convex quadratic program

$$\min_{q \in \Delta_M} (q - q_0)^\top K (q - q_0) + \lambda\, q^\top \widetilde{K} q, \qquad \lambda > 0. \tag{34}$$

The gradient is explicit:

$$\nabla_q \Big((q - q_0)^\top K (q - q_0) + \lambda\, q^\top \widetilde{K} q \Big) = 2K(q - q_0) + 2\lambda \widetilde{K} q,$$

so exponentiated-gradient updates yield a simple solver.

E Additional Numerical Results

Post-hoc Projected-Vendi reweighting. To evaluate the effectiveness of the empirical-support projection mechanism, we apply the reweighting procedure in Equation (19) (penalized variant with $\lambda = 0.01$) to fixed sample sets of sizes ranging from 1000 to 10000 drawn from several state-of-the-art generative models on FFHQ and ImageNet. For each model and subsample size, we generate one large pool of samples and then create two empirical measures: (i) the uniform baseline (Uniform

Algorithm 2: Empirical-support order-2 projection (exponentiated gradient)

Input: K; penalty $\lambda > 0$; stepsize $\eta > 0$; iterations T
Output: $q_T \in \Delta_M$

Compute $\widetilde{K} \leftarrow K \circ K$ and initialize $q_0 \leftarrow \frac{1}{M}\mathbf{1}$;
for $t = 0, 1, \ldots, T-1$ **do**
  $g_t \leftarrow 2K(q_t - q_0) + 2\lambda\,\widetilde{K} q_t$;
  $\tilde{q}_{t+1,i} \leftarrow q_{t,i} \exp(-\eta\, g_{t,i})$ for all i; $\quad q_{t+1} \leftarrow \tilde{q}_{t+1} / \sum_j \tilde{q}_{t+1,j}$;
return q_T;

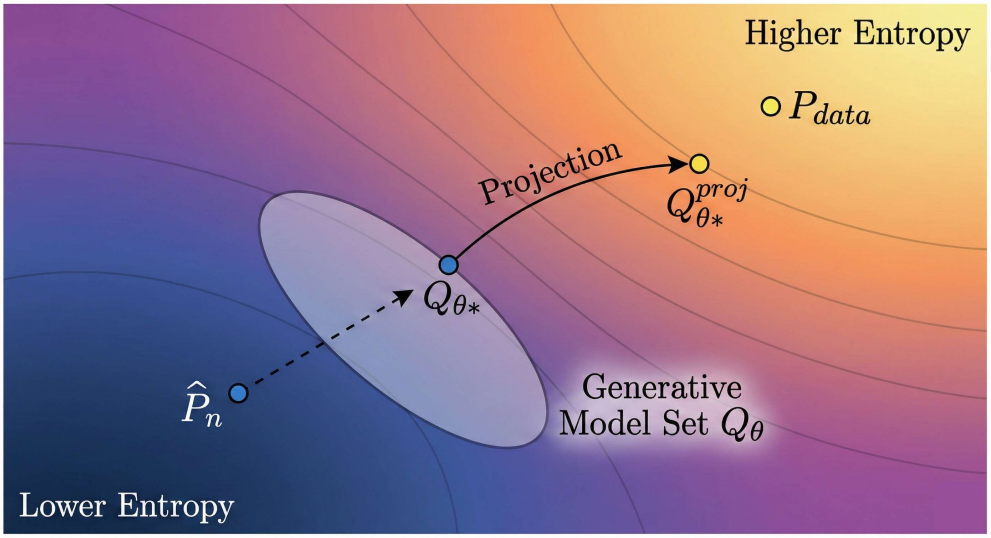


Figure 6: Entropy level sets in distribution space. $\widehat{P}_n$ and Q_{θ^*} lie in a lower-entropy region; P_{data} lies higher. Projection onto the entropy superlevel set yields $Q_{\theta^{\text{proj}}}$, closer to P_{data} within the model family.

sampling), and (ii) the Projected-Vendi reweighted distribution obtained by solving the quadratic program. Diversity is measured via Vendi and RKE scores, while distributional proximity to the distribution of test samples is quantified via kernel distance (KD) and Fréchet distance (FD) in DINOv2 feature space. In Figure 7, we observe the Vendi and RKE trends as the number of samples increases. We find that diversity scores consistently improve across different sample sizes. The quantitative results, reported in Table 1 for ProjectedGAN Sauer et al. (2021), BigGAN Brock et al. (2018), and LDM Rombach et al. (2022), demonstrate that Projected-Vendi consistently achieves considerably higher Vendi and RKE scores compared to uniform weighting, while incurring only modest increases in KD and FD, thereby confirming the projection's ability to reduce diversity bias.

E.1 Extended analysis of Vendi score convergence on empirical distributions

Building on the results in Figure 2, we extend our analysis to AFHQ-v2 Choi et al. (2020) and LSUN-Bedroom Yu et al. (2015), using parameters identical to those in the main text. Note that for AFHQ-v2, which contains only 15,803 images, we do not report a confidence interval for the final data point. We observe that regardless of the chosen dataset, the increasing trend remains.

E.2 Extended analysis of generative models

In this section we report more results on comparing generative models to the underlying training set. We provide results for a variety of architectures across different embedding models.

Table 1: Evaluated scores for comparison of Uniform sampling vs. Vendi-Projected sampling.

	Model	Sampling	Vendi ↑	RKE ↑	KD ↓	FD ↓
FFHQ	ProjectedGAN	Uniform	193.64	9.31	1.958	586.43
		Projected-Vendi	237.34	10.88	1.642	562.92
	LDM	Uniform	210.32	10.21	0.587	223.32
		Projected-Vendi	248.01	11.92	0.462	213.87
ImageNet	BigGAN	Uniform	885.68	32.56	0.539	403.57
		Projected-Vendi	947.32	33.27	0.517	386.34
	LDM	Uniform	889.44	33.83	0.102	124.31
		Projected-Vendi	919.18	34.89	0.094	105.95

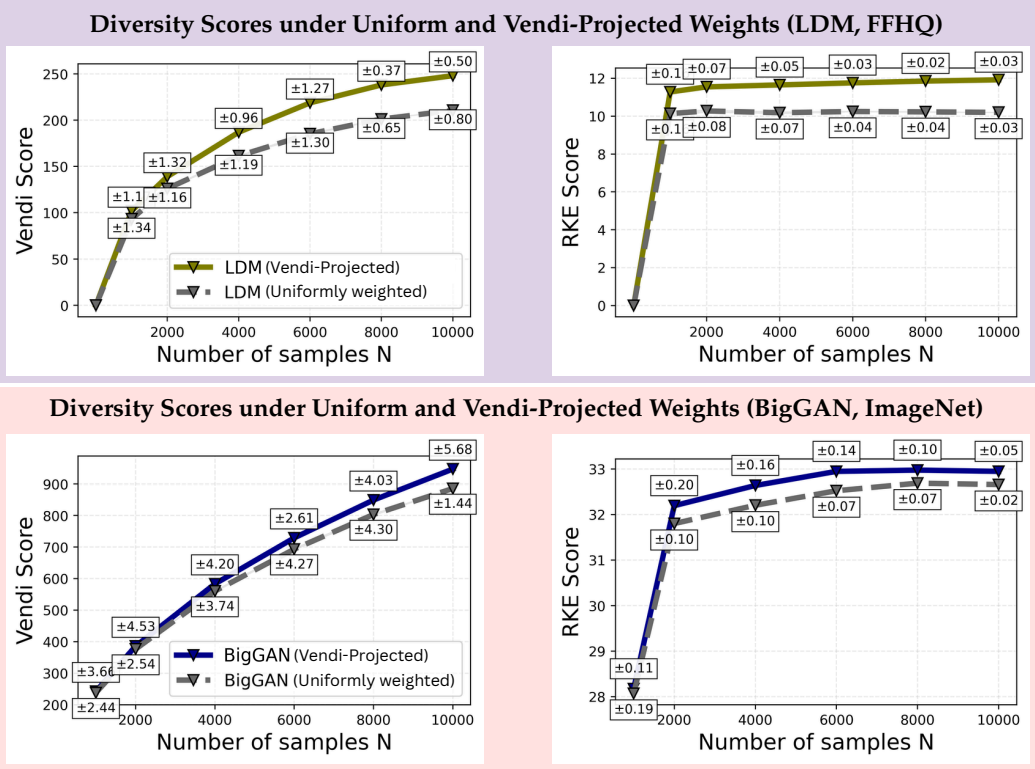


Figure 7: Vendi and RKE diversity scores vs. size N for N uniformly weighted generated samples and their post-hoc Vendi-projected reweighting. (Upper) LDM trained on FFHQ. (Lower) BigGAN trained on ImageNet. The Vendi-based projection yields higher diversity scores for the reweighted empirical distribution.

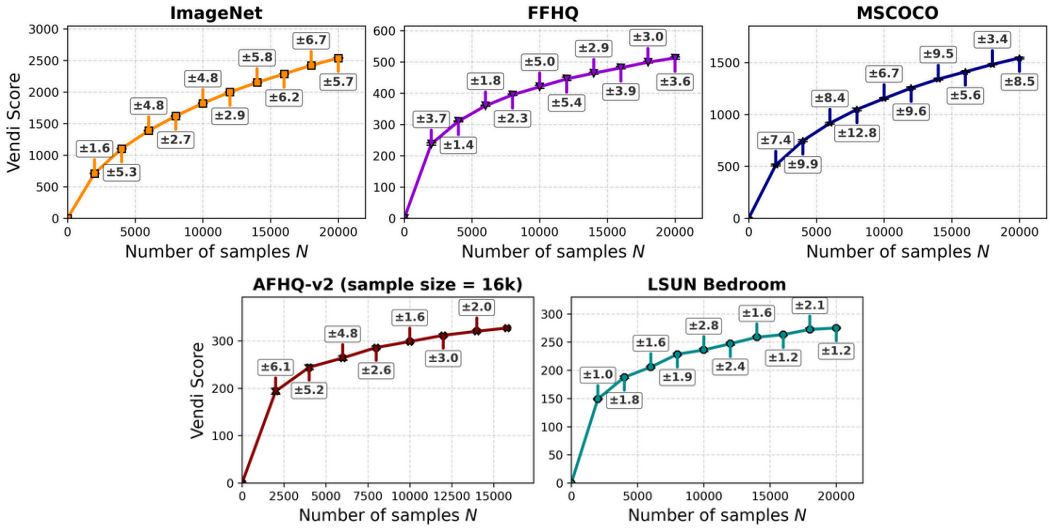


Figure 8: Vendi score curves (mean and 95% confidence intervals over 5 independent sample sets) for ImageNet, FFHQ, MSCOCO, LSUN Bedroom (sample size values of $n \leq 20$K) and AFHQ-v2 (sample size values $n \leq 16000$), computed using DINOv2 embeddings and a Gaussian (RBF) kernel with bandwidth $\sigma = 35$.

E.2.1 The effect of chosen embeddings on ranking of generative models

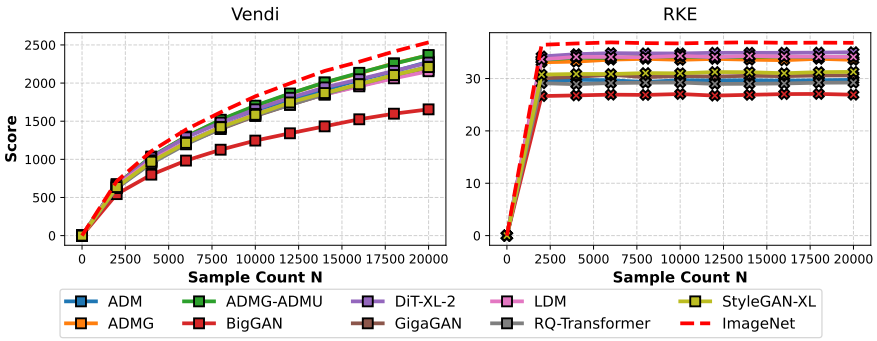
Extending the comparison in Figure 3, we evaluated the generative models across four distinct embedding architectures: DINOv2 Oquab et al. (2023), DINOv3 Siméoni et al. (2025), CLIP Radford et al. (2021), and InceptionV3 Szegedy et al. (2016). We observe that for all embeddings, the diversity of the real dataset consistently matches or exceeds that of the generated samples. These results are detailed in Figures 9–12.

E.2.2 The effect of choosing a cosine similarity kernel

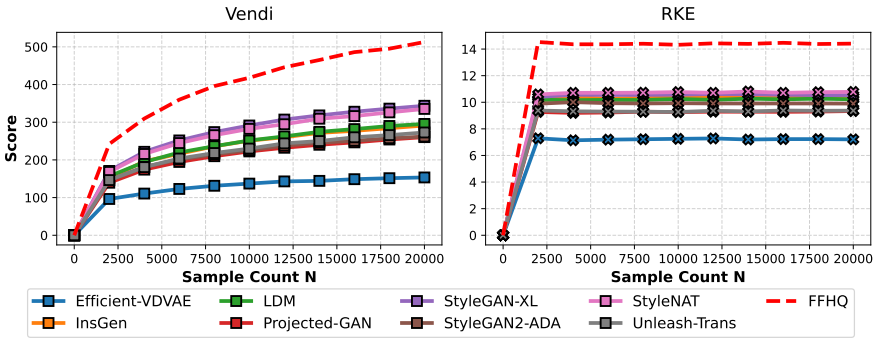
To provide a comprehensive analysis of the Vendi metric, we also evaluated diversity using the cosine similarity kernel. As established by Ospanov & Farnia (2025), the convergence of this kernel scales with the dimension of the embedding model. We confirm that this behavior holds across all four datasets. As shown with Gaussian kernel the diversity of the training set consistently exceeds that of the generative models. The downward diversity bias of the generative holds even when we swap the underlying kernel function.

E.2.3 Extending results to reference-based metrics (Recall and Coverage)

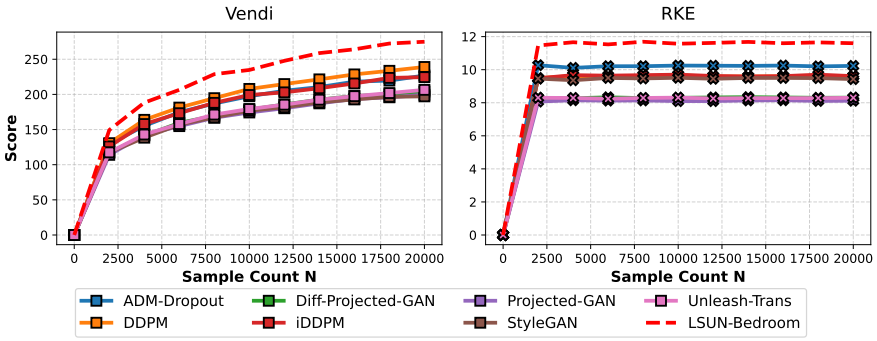
To provide a comprehensive analysis of generative model diversity, we also report the reference-based Recall Kynkäänniemi et al. (2019) and Coverage Naeem et al. (2020) metrics. As these metrics require a reference distribution, scores for the datasets themselves are omitted. For completeness, we present these results alongside the corresponding RKE and Vendi scores. We report scores in Tables 2, 3 and 4.



(a) ImageNet



(b) FFHQ



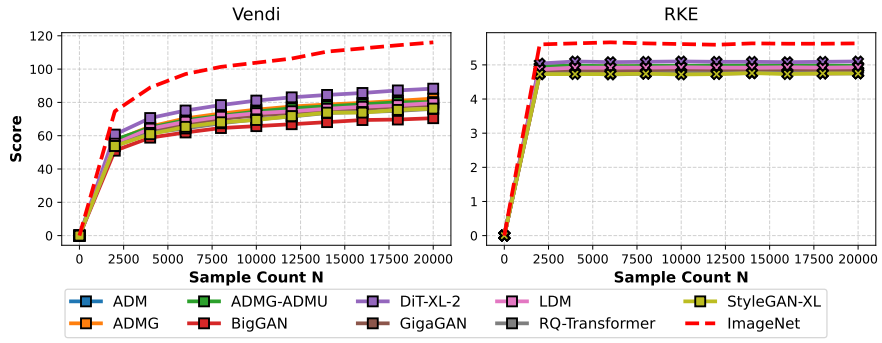
(c) LSUN Bedroom

Figure 9: Comparison of Vendi scores of the test sample set (the dashed red curve) and the generated samples by pre-trained generative models across three datasets. The backbone embedding is DINOv2 embeddings using Gaussian kernel with bandwidth $\sigma = 35$.

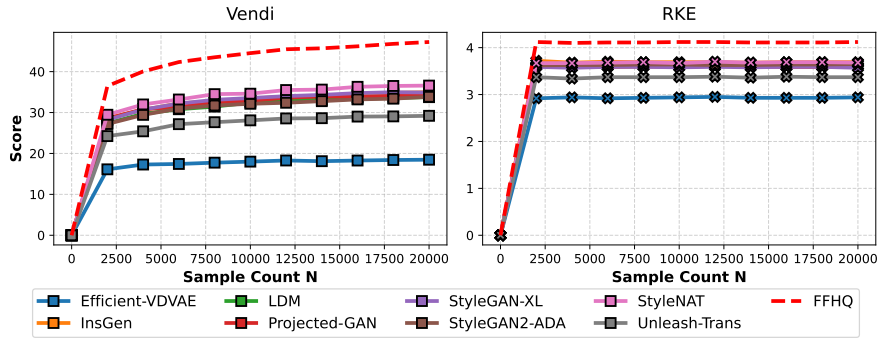
E.3 Extended analysis of generative models trained on subsets of training data

E.3.1 Additional results on LDM, StyleGAN-XL and UNet based architectures

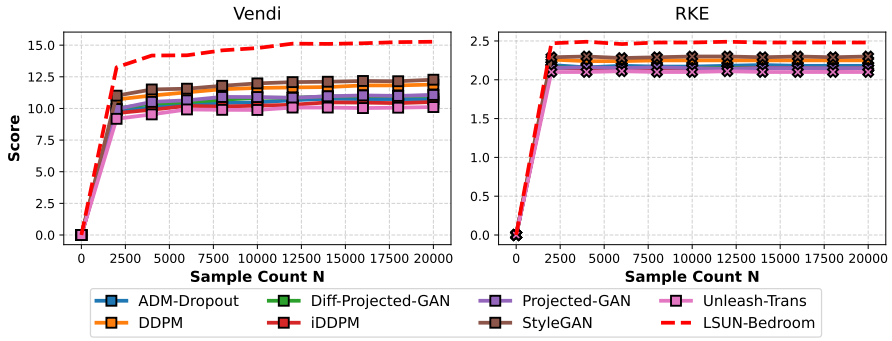
Expanding on the findings in Figure 4, we evaluated two additional architectures: a small U-Net and a Latent Diffusion Model (LDM). The U-Net was trained on 100%, 10%, 2%, and 1% subsets of ImageNet, while the LDM was trained on 100%, 50%, and 25% subsets of FFHQ, using standard technical configurations for each. Figures 14 (DINOv2) and 15 (CLIP) confirm that the downward diversity bias persists across architectures and amplifies as the training set size decreases.



(a) ImageNet



(b) FFHQ

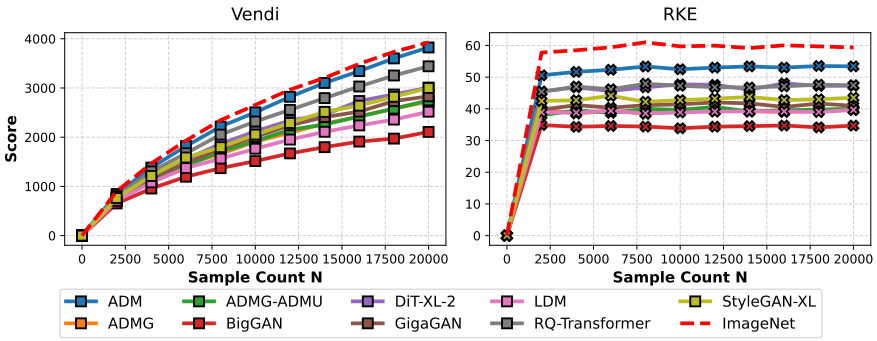


(c) LSUN Bedroom

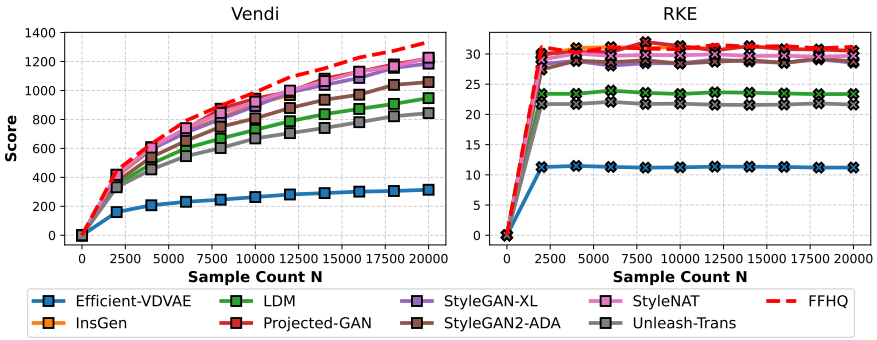
Figure 10: Comparison of Vendi scores of the test sample set (the dashed red curve) and the generated samples by pre-trained generative models across three datasets. The backbone embedding is CLIP embeddings using Gaussian kernel with bandwidth $\sigma = 8.0$.

E.3.2 Extending results to reference-based metrics (Recall and Coverage)

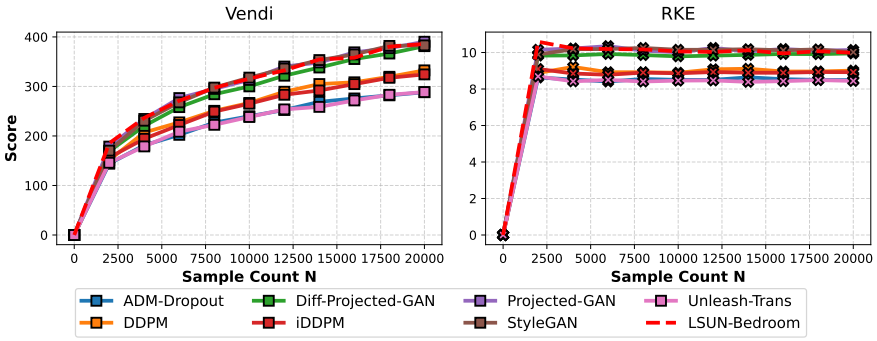
We further evaluated the LDM and StyleGAN-XL models using Recall, Coverage, RKE, and Vendi scores. The results align with our reference-free analysis, as both Recall and Coverage confirm that diversity degrades with smaller training set sizes.



(a) ImageNet



(b) FFHQ



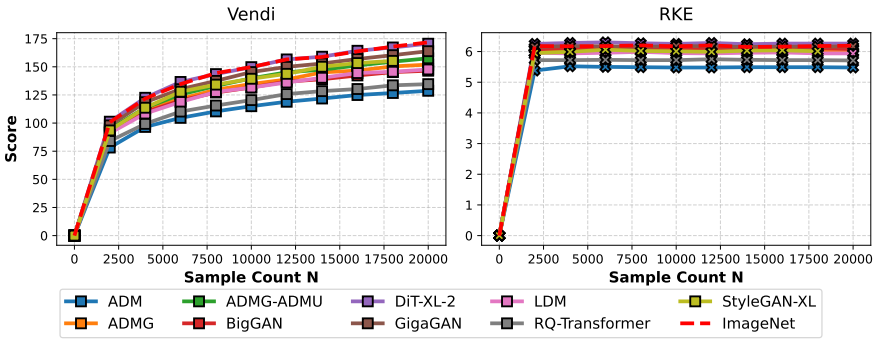
(c) LSUN Bedroom

Figure 11: Comparison of Vendi scores of the test sample set (the dashed red curve) and the generated samples by pre-trained generative models across three datasets. The backbone embedding is InceptionV3 embeddings using Gaussian kernel with bandwidth $\sigma = 9.0$.

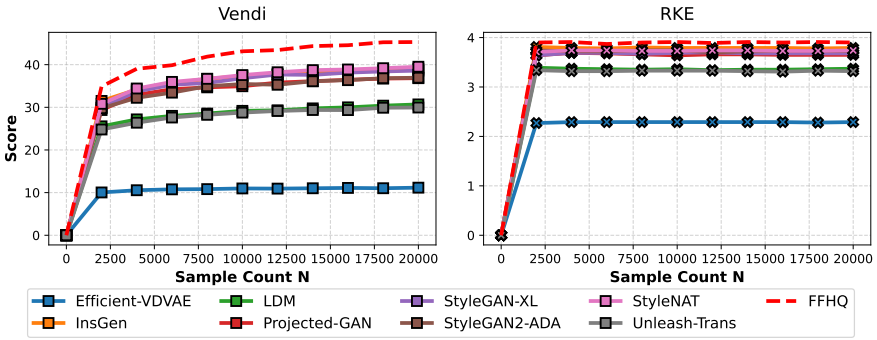
E.4 Additional guidance results

E.4.1 Hyperparameters and Experimental settings

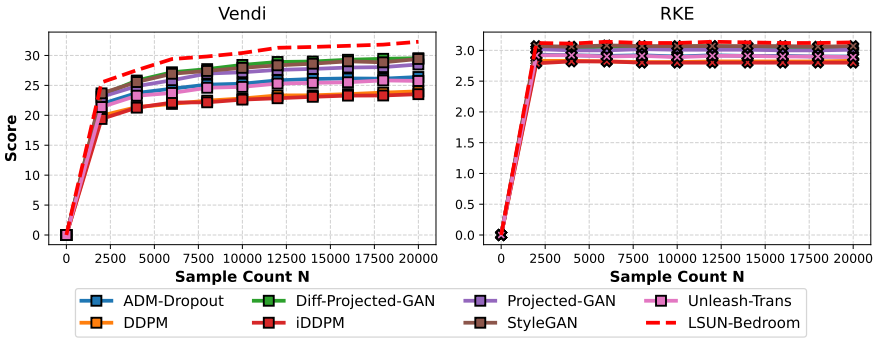
In the SPARKE and Vendi Guidance experiment, we considered a Gaussian kernel, which consistently led to higher output scores in comparison to the other standard cosine similarity kernel. We used the same Gaussian kernel bandwidth σ in the RKE and Vendi experiments, and the bandwidth parameter choice matches the selected value in Jalali et al. (2026); Friedman & Dieng (2023). The numerical experiments were conducted on 4×NVIDIA GeForce RTX 4090 GPUs, each of which has 22.5 GB of memory.



(a) ImageNet



(b) FFHQ



(c) LSUN Bedroom

Figure 12: Comparison of Vendi scores of the test sample set (the dashed red curve) and the generated samples by pre-trained generative models across three datasets. The backbone embedding is DINOv3-ViTS16 embeddings using Gaussian kernel with bandwidth $\sigma = 8.0$.

E.4.2 Experimental Configuration for Table 8

Vendi Guidance. We used a Gaussian kernel with bandwidth $\sigma_{img} = 0.8$ and used $\eta = 0.03$ as the weight of Vendi guidance. To balance the effects of the diversity guidance in sample generation, the Vendi guidance update was applied every 10 reverse-diffusion steps in the diffusion process, which is similar to the implementation of Vendi score guidance in Askari Hemmat et al. (2024).

SPARKE (RKE Guidance). We considered the same Gaussian kernel for the image generation with bandwidth $\sigma_{img} = 0.8$. The guidance hyperparameter was set to $\eta = 0.03$, as in Vendi guidance. The SPARKE diversity guidance was applied every 10 reverse-diffusion steps.

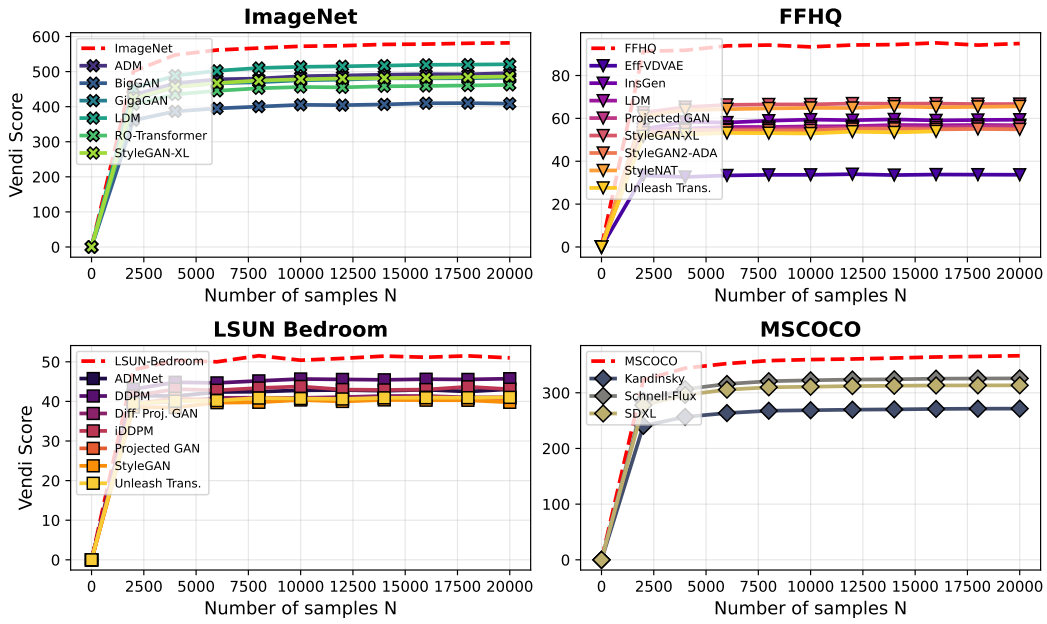


Figure 13: Comparison of generative model architectures across four datasets. We evaluate models on ImageNet, FFHQ, LSUN Bedroom, and MSCOCO using DINOv2 embeddings with cosine similarity kernel.

Table 2: Comparison of RKE, Coverage, and Recall across different model architectures trained on ImageNet. Coverage and Recall metrics were computed using Stein et al. (2023) implementation. Best model scores are underlined.

Method	Recall ↑	Coverage ↑	RKE ↑	Vendi ↑
Dataset	-	-	**36.82**	**2533.93**
ADM Dhariwal & Nichol (2021)	<u>0.79</u>	0.89	29.74	<u>2279.73</u>
BigGAN Brock et al. (2018)	0.44	0.57	26.91	1655.41
GigaGAN Kang et al. (2023)	0.74	0.70	30.59	2174.02
LDM Rombach et al. (2022)	0.76	0.93	<u>34.13</u>	2155.19
RQ-Transformer Lee et al. (2022)	0.76	0.59	29.23	2214.47
StyleGAN-XL Sauer et al. (2022)	0.74	<u>0.96</u>	31.27	2209.83

E.4.3 SPARKE AND VENDI GUIDANCE

As discussed in the main text, one effective strategy to mitigate downward diversity bias is to apply regularization during training. We extend the analysis to a circular and grid pattern configurations. Our observations remain consistent: regularization significantly improves the mode coverage of the JS-GAN model. Furthermore, Table 7 shows that this improvement extends to reference-based metrics, enhancing both quality and diversity scores.

Moreover, to mitigate the downward diversity bias, we demonstrate the efficacy of inference-time guidance on Diffusion models. By leveraging Vendi and RKE-based guidance, we actively encourage the model to explore underrepresented regions of the learned manifold. Table 8 reports that applying Vendi or RKE guidance yields consistent improvements across Vendi, RKE, Recall, and Coverage scores. Crucially, this gain in diversity does not come at the cost of quality. Instead, we observe a concurrent improvement in sample quality—indicated by lower FD and KD, and higher Precision and Density—suggesting that the guidance effectively corrects.

Table 3: Comparison of RKE, Coverage, and Recall across different model architectures trained on FFHQ. Coverage and Recall metrics were computed using Stein et al. (2023) implementation. Best model scores are underlined.

Method	Recall ↑	Coverage ↑	RKE ↑	Vendi ↑
Dataset	-	-	**14.46**	**512.84**
LDM Rombach et al. (2022)	<u>0.44</u>	<u>0.74</u>	10.21	295.47
Eff. vdVAE Hazami et al. (2022)	0.14	0.52	7.21	153.45
InsGen Yang et al. (2021)	0.14	0.51	10.38	291.21
Projected GAN Sauer et al. (2021)	0.07	0.30	9.34	260.41
StyleGAN XL Sauer et al. (2022)	0.42	0.61	10.53	<u>343.63</u>
StyleGAN2-Ada Karras et al. (2020)	0.04	0.41	9.90	261.83
StyleNAT Walton et al. (2022)	0.42	0.71	<u>10.78</u>	335.26
Unleashing Trans. Bond-Taylor et al. (2022)	0.24	0.54	9.37	272.79

Table 4: Comparison of RKE, Coverage, and Recall across different model architectures trained on LSUN-Bedroom. Coverage and Recall metrics were computed using Stein et al. (2023) implementation. Dataset scores are bolded, best model scores are underlined.

Method	Recall ↑	Coverage ↑	RKE ↑	Vendi ↑
Dataset	-	-	**11.60**	**274.90**
ADMNet Dropout Dhariwal & Nichol (2021)	<u>0.75</u>	<u>0.90</u>	<u>10.23</u>	226.85
DDPM Ho et al. (2020)	0.61	0.68	9.61	<u>239.00</u>
Diffusion-Projected GAN Wang et al. (2022)	0.28	0.28	8.31	204.06
iDDPM Nichol & Dhariwal (2021)	0.64	0.76	9.61	224.55
Projected GAN Sauer et al. (2021)	0.22	0.24	8.12	199.45
StyleGAN Karras et al. (2019)	0.41	0.69	9.43	197.21
Unleasing Trans. Bond-Taylor et al. (2022)	0.41	0.42	8.30	206.47

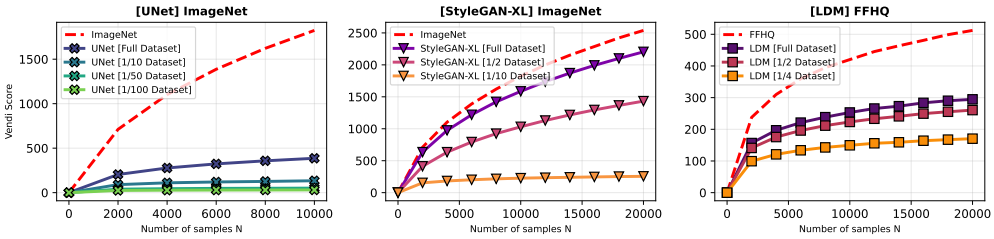


Figure 14: Vendi Score comparison for UNet, StyleGAN-XL, LDM. The models were trained on varying subsets of ImageNet, ImageNet, FFHQ respectively (indicated by fraction of the original size). Dashed red line represents score of the corresponding training dataset. The evaluation results are done using DINOv2 with $\sigma = 35$.

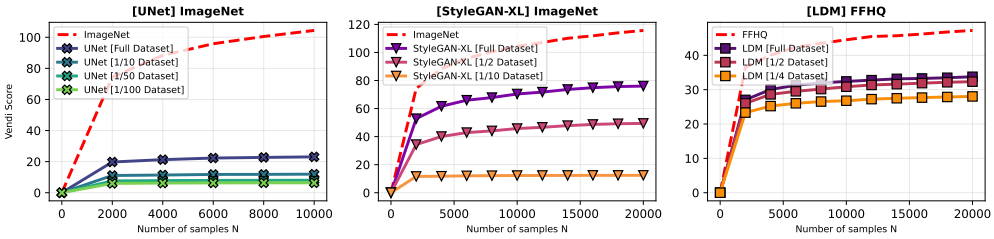


Figure 15: Vendi Score comparison for UNet, StyleGAN-XL, LDM. The models were trained on varying subsets of ImageNet, ImageNet, FFHQ respectively (indicated by fraction of the original size). Dashed red line represents score of the corresponding training dataset. The evaluation results are done using CLIP with $\sigma = 8.0$.

Table 5: Comparison of RKE, Coverage, and Recall across StyleGAN-XL models trained on varying sizes of ImageNet.

Method	Recall ↑	Coverage ↑	RKE ↑	Vendi ↑
Dataset	-	-	**36.82**	**2533.93**
StyleGAN-XL: full dataset	0.74	0.96	31.27	2203.39
StyleGAN-XL: 1/2 dataset	0.59	0.94	20.80	1432.20
StyleGAN-XL: 1/10 dataset	0.42	0.93	14.95	255.44

Table 6: Comparison of RKE, Coverage, and Recall across LDM models trained on varying sizes of FFHQ.

Method	Recall ↑	Coverage ↑	RKE ↑	Vendi ↑
Dataset	-	-	**14.46**	**514.02**
LDM: full dataset	0.44	0.74	10.21	294.60
LDM: 1/2 dataset	0.22	0.74	9.17	260.64
LDM: 1/4 dataset	0.03	0.52	7.06	170.44

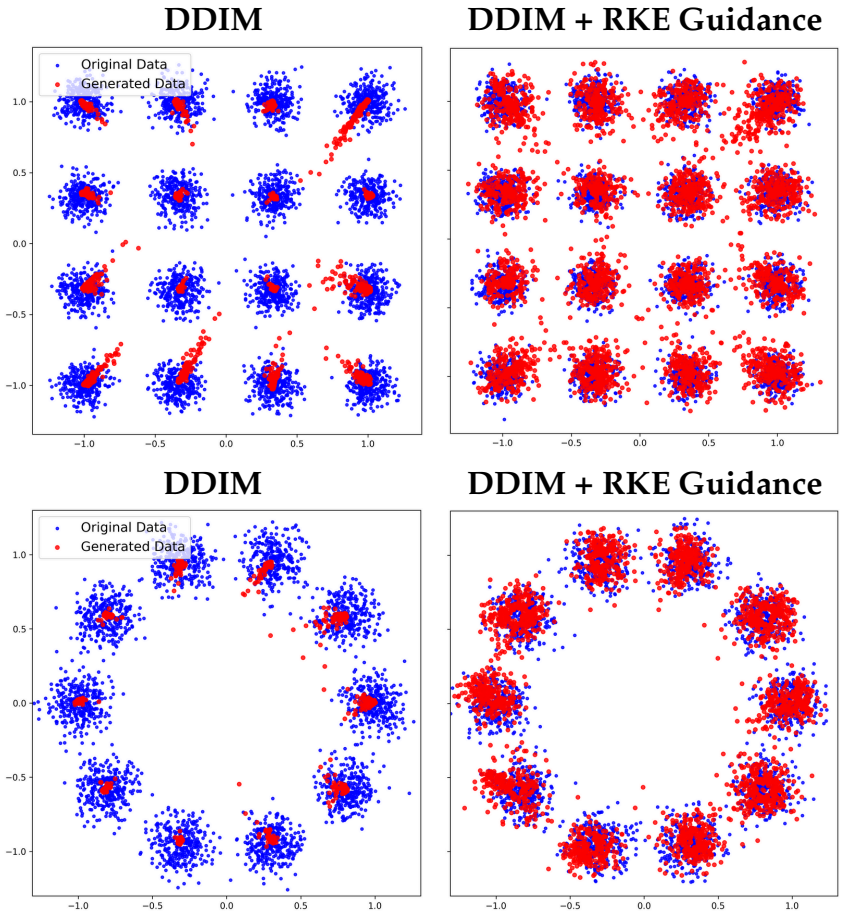


Figure 16: Comparison of DDIM with and without RKE regularization in two 2D GMM distributions: (1) modes are placed in a circular pattern, (2) modes are placed in a grid.

Table 7: Comparison of Vendi scores, RKE, KD, and FD across different sampling settings. on 2D Gaussian Mixtures with 10 components.

Method	Vendi	RKE	KD $\times 10^2$	FD	Precision	Recall	Density	Coverage
Dataset	16.49	12.86	-	-	-	-	-	-
JS-GAN+RKE Reg.	15.20	12.18	0.098	0.081	0.98	0.89	1.02	0.78
JS-GAN	12.61	10.92	0.767	0.654	0.98	0.83	0.99	0.49

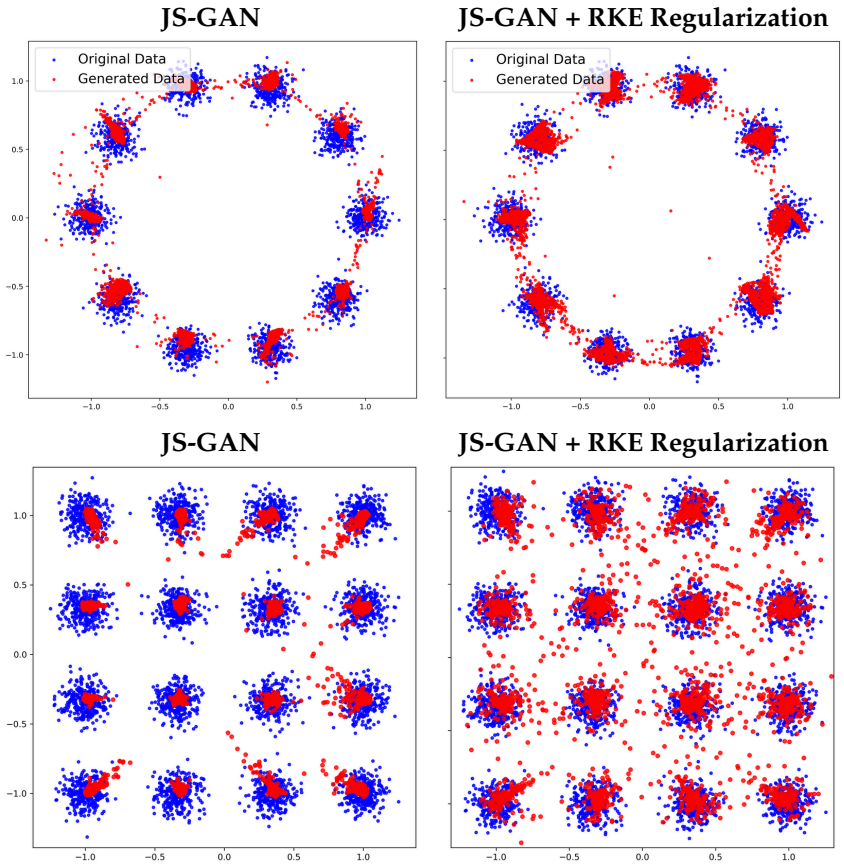


Figure 17: Comparison of JS-GAN generated GMM modes with and without diversity regularization during training. We show two mode placement examples: (1) modes are in a circular pattern, (2) modes are in the 4×4 grid pattern.

Table 8: Comparison of Vendi scores, RKE, KD, and FD across different sampling settings. on DiT (trained on ImageNet) using DINOv2 feature embeddings. ($\sigma = 35$)

Method	Vendi ↑	RKE ↑	KD ↓	FD ↓	Precision ↑	Recall ↑	Density↑	Coverage↑
Dataset	1747.60	36.37	-	-	-	-	-	-
Vendi guidance	672.24	33.94	0.1628	328.16	0.95	0.15	1.63	0.73
RKE guidance	751.09	34.50	0.1629	319.56	0.96	0.16	1.54	0.74
No guidance	603.68	33.71	0.2174	422.15	0.94	0.11	1.50	0.64

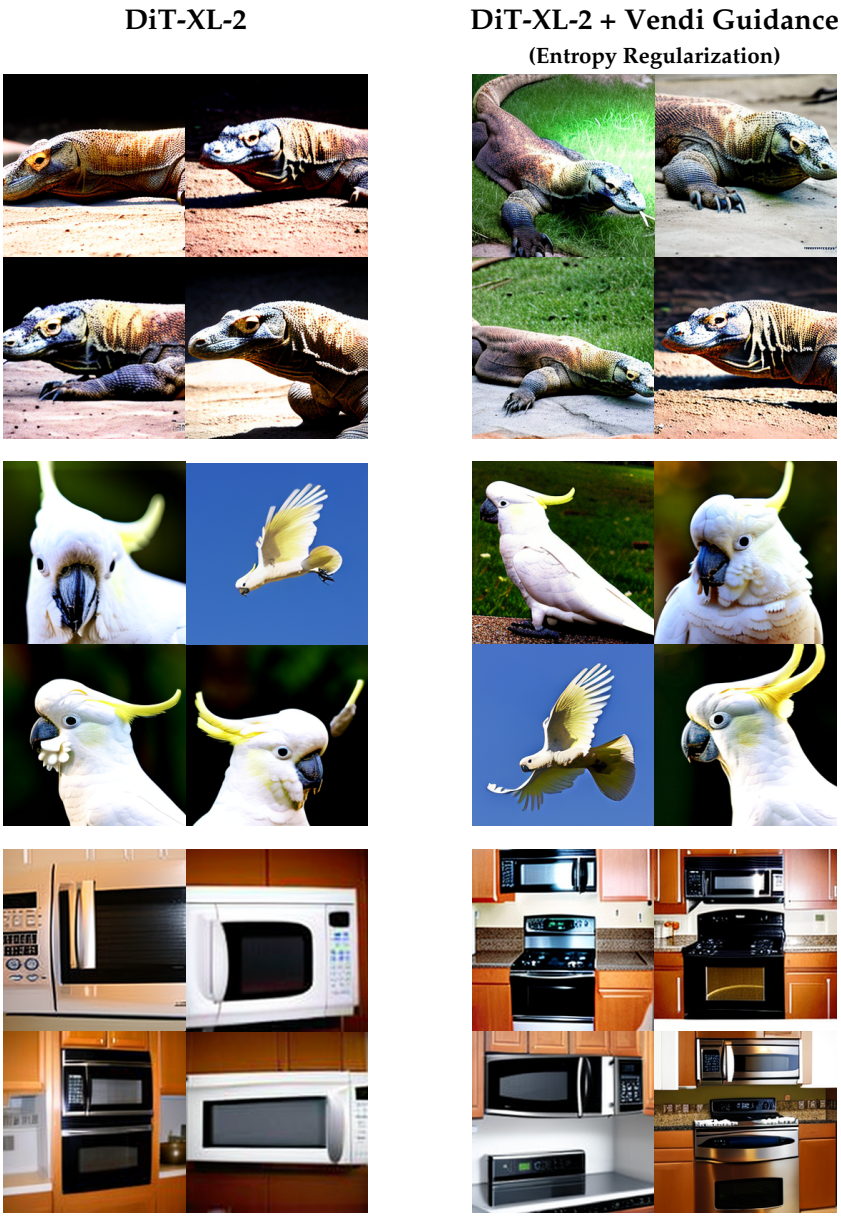


Figure 18: Qualitative Comparison of samples generated by DiT-XL-2 and DiT-XL-2 with Vendi guidance (entropy regularization).