Conditional Vendi Score: Prompt-Aware Diversity Evaluation for Generative AI Models and LLMs

Mohammad Jalali[1] **Azim Ospanov**[1] **Amin Gohari**[2] **Farzan Farnia**[1]
[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong
[2]Department of Information Engineering, The Chinese University of Hong Kong
{mjalali24, aospanov9, farnia}@cse.cuhk.edu.hk, agohari@ie.cuhk.edu.hk

Abstract

Generative models guided by text prompts are widely evaluated for fidelity and prompt alignment, yet their ability to produce outputs remains underexplored. Existing diversity metrics such as Vendi and RKE, which are based on the von Neumann and Rényi entropies of kernel matrices, were developed for unconditional models and cannot distinguish prompt-induced from model-induced variability. We address this gap by introducing *Conditional-Vendi* and *Conditional-RKE*, diversity measures derived from the conditional entropy of positive semidefinite matrices. These scores isolate model-induced diversity in prompt-guided generation, with Conditional-RKE enjoying an $O(1/\sqrt{n})$ convergence rate. For Conditional-Vendi, we introduce a truncated-spectrum approximation that yields scalable and consistent estimates. Experiments on text-to-image, image-captioning, and LLM tasks show that the conditional scores recover ground-truth diversity orderings and can also guide diffusion models toward more diverse samples. The codebase is available at https://github.com/mjalali/conditional-vendi.

1 Introduction

Prompt-guided generative AI systems, including large language models (LLMs) (Brown et al., 2020), text-to-image models (Rombach et al., 2022; Ramesh et al., 2022; Saharia et al., 2022), and text-to-video models (Ho et al., 2022b,a; OpenAI, 2024b), have achieved remarkable success across a wide range of applications. In these models, sample generation is conditioned on an input text prompt, with the goal of producing outputs aligned to the prompt. This conditional generation mechanism distinguishes prompt-guided models from traditional unconditional generative models (Kingma and Welling, 2013; Goodfellow et al., 2014), which aim to mimic the overall data distribution without a guiding input. Because most evaluation metrics for generative models were originally developed in the unconditional setting, the recent literature has sought to design new measures that better capture the properties of text-conditioned models.

Current evaluation metrics for prompt-guided models primarily emphasize *fidelity*: the quality of generated outputs and their consistency with the input text. A common approach is to compute similarity in a shared embedding space between text and outputs, such as ClipScore (Hessel et al., 2021) for text-to-image generation, which uses CLIP embeddings (Radford et al., 2021) to quantify alignment. Such fidelity-based measures ensure that the generated content matches the semantics of the prompt but leave open the question of how to assess the *diversity* of model outputs.

Diversity has been extensively studied in unconditional generation, with metrics such as Recall (Sajjadi et al., 2018; Kynkäänniemi et al., 2019), Coverage (Naeem et al., 2020), Vendi (Friedman and Dieng, 2023), and RKE (Jalali et al., 2023). These scores are often applied directly to prompt-guided models, but they conflate two sources of variability: the *prompt-induced diversity*, arising from differences across input prompts, and the *model-induced diversity*, reflecting randomness in outputs for similar prompts. As illustrated in Figure 1, these two components capture fundamentally different aspects of a model's behavior, yet existing diversity measures do not disentangle them. This can bias diversity comparisons across models and obscure

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

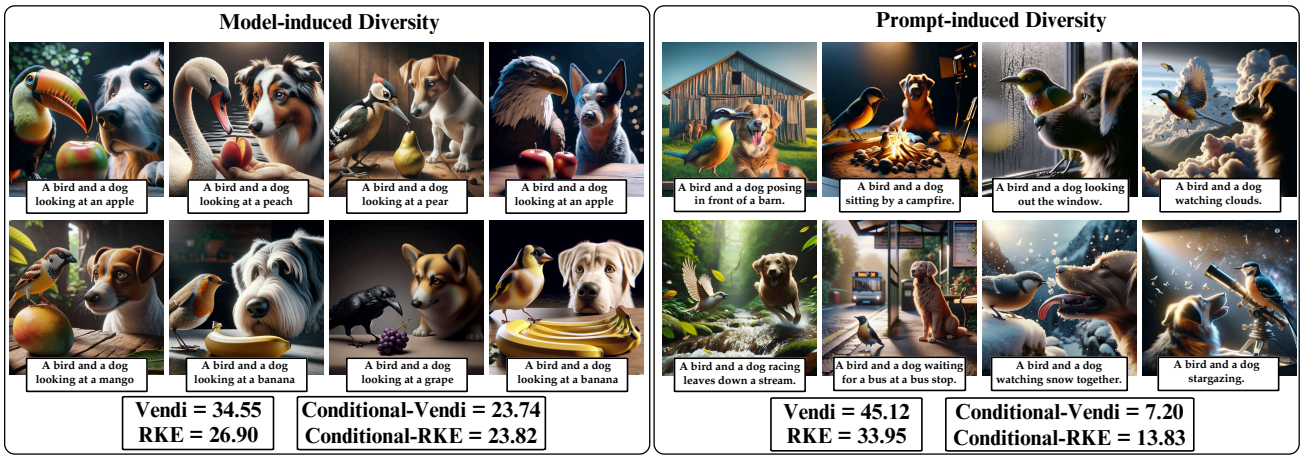


Figure 1: Illustration of *prompt-induced diversity* (left), where variation comes from different prompts, and *model-induced diversity* (right), where variation arises from the generator itself. Unconditional image diversity scores, Vendi and RKE, assign higher diversity to the right side, overlooking prompt effects. Our proposed Conditional-RKE and Conditional-Vendi account for prompts and assign higher diversity to the left side.

whether differences are due to richer model variability or merely broader prompt coverage.

To address this gap, we build on recent entropy-based approaches for unconditional models. The Vendi and RKE scores correspond respectively to the von Neumann entropy and order-2 Rényi entropy of positive semidefinite (PSD) matrices applied to kernel matrices (Friedman and Dieng, 2023; Jalali et al., 2023). We extend this framework to the *conditional entropy of PSD matrices* (Giraldo et al., 2014) and apply it to kernel matrices, yielding two prompt-aware diversity measures: the *Conditional-Vendi* and *Conditional-RKE* scores. These quantities decompose the kernel-entropy $H(X)$ of generated data X as:

$$H(X) = H(X \mid T) + I(X;T)$$

into the terms of conditional entropy $H(X|T)$ given prompts T and mutual information $I(X;T)$, where $H(X|T)$ serves as a measure of model-induced diversity for a prompt-aware diversity assessment.

We analyze the statistical behavior of these conditional scores. For Conditional-RKE, we prove an $O(n^{-1/2})$ convergence rate, enabling reliable estimation with moderate samples. For Conditional-Vendi, direct estimation is sample-inefficient due to dimension dependence, affecting its practicality for large-scale generative tasks with sample sizes bounded to several tens of thousands. To address this, we extend the eigenspectrum truncation method for unconditional Vendi score (Ospanov and Farnia, 2025) to the conditional setting, which provides scalable and consistent approximate truncated-Conditional-Vendi score.

Figure 1 illustrates the use of (truncated) Conditional-Vendi and Conditional-RKE scores in comparing the diversity of two sets of "dog and bird" samples generated by DALL · E 3. In the first set, the model produces a variety of dog and bird breeds under similar prompts, whereas in the second, the same dog and bird appear in different contextual scenes across diverse prompts. While the prompt-unaware RKE and Vendi scores assign higher diversity to the second set, the prompt-aware Conditional-Vendi and Conditional-RKE scores instead favor the first set, capturing the greater breed-level variation that is orthogonal to prompt differences.

Beyond evaluation, we also leverage prompt-aware diversity scores to guide sample generation in diffusion models. Extending the unconditional Vendi guidance method of Askari Hemmat et al. (2024), we apply Conditional-Vendi guidance to text-conditioned latent diffusion models (Rombach et al., 2021), aiming to promote prompt-aware diversity in sample generation.

Finally, we validate our framework across text-to-image, text-to-video, and language generation tasks. Using controlled experiments where ground-truth diversity rankings are available, we show that the conditional scores recover the intended rankings and remain computationally tractable at scale. We further demonstrate how Conditional-Vendi can be decomposed across different prompt modes to evaluate diversity conditioned on text categories. In summary, our contributions are as follows:

- We study prompt-aware diversity evaluation for prompt-conditioned generative models.
- We propose *Conditional-Vendi* and *Conditional-*

RKE prompt-aware diversity scores.

- We analyze the proposed scores' statistical convergence, and propose a truncation method to reduce the sample complexity of the diversity scores.
- We numerically validate the scores showing correlation with ground-truth model-induced diversity.

2 Related Work

Evaluation of deep generative models: Metrics for evaluating generative models are generally divided into reference-dependent and reference-free categories (Borji, 2022). Reference-dependent metrics compare generated and real data distributions, with common examples including FID (Heusel et al., 2017) and KID (Bińkowski et al., 2018; Wang et al., 2025). Other reference-based measures, such as the Inception Score (Salimans et al., 2016), Precision/Recall (Sajjadi et al., 2018; Kynkäänniemi et al., 2019), and Density/Coverage (Naeem et al., 2020), jointly evaluate fidelity and diversity with respect to a reference dataset.

Beyond fidelity, several works examine memorization and novelty. These include the authenticity score (Alaa et al., 2022) and Feature Likelihood Divergence (Jiralerspong et al., 2023) for assessing generalization, as well as the rarity score (Han et al., 2023) and KEN (Zhang et al., 2024, 2025) for quantifying novelty. The memorization metrics are reference-based. In contrast, reference-free evaluations assess quality and diversity directly from the generated data. Notable examples include the Vendi score (Friedman and Dieng, 2023; Pasarkar and Dieng, 2024; Ospanov et al., 2024) and RKE score (Jalali et al., 2023) for diversity, and (Nguyen and Dieng, 2024) for evaluating the quality of generated data. We also note that the diversity-aware online evaluation of generative models has been studied in the related works (Hu et al., 2025a; Rezaei et al., 2025; Hu et al., 2025b,c; Jafari and Farnia, 2026). We also note the concurrent work by Ospanov et al. (2025) on the prompt-aware diversity evaluation using the Schur complement of CLIP embeddings.

Evaluation of conditional generative models: The evaluation of prompt-based generative models, such as text-to-image and text-to-video systems, has been explored in several recent works. Most metrics focus on measuring alignment between prompts and outputs. A widely used example is CLIPScore (Hessel et al., 2021), which computes cosine similarity in the CLIP embedding space. Other efforts have introduced benchmarks and curated prompt sets to evaluate broader aspects. For instance, HEIM (Lee et al., 2023) assesses twelve criteria, including text–image alignment, image quality, and bias. Also, Kim et al. (2022b) propose the Mutual Information Divergence (MID) score, which fits multivariate Gaussian distributions to text and image representations and estimates their mutual information to quantify relevance in conditional generative models.

However, alignment- and quality-focused metrics may overlook output diversity. Astolfi et al. (2024) emphasize that metrics centered on style or aesthetics can fail to capture variability across outputs for the same prompt. They propose computing per-prompt diversity using similarity functions and then averaging across prompts. Similarly, Kannen et al. (2024) extend the Vendi score to the per-prompt setting. Both approaches require generating multiple outputs for each prompt with different seeds. In contrast, our proposed Conditional-Vendi does not require repeated generations; instead, it quantifies model-induced diversity by analyzing variability across prompt types. Our theoretical results interpret Conditional-Vendi as an aggregation of diversity scores across prompt categories.

3 Preliminaries

3.1 Generative distributions and notation

We focus on a conditional generative model that produces a random output $X \in \mathcal{X}$ given an input text prompt $T \in \mathcal{T}$ according to a conditional distribution $P_{X|T}$. For a prompt $T = t$, the model outputs a sample drawn from $P_{X|T=t}$. We consider n sample pairs $\{(t_i, x_i)\}_{i=1}^n$ where the prompts t_i are drawn independently from P_T and, conditional on t_i, the outputs x_i are drawn from $P_{X|T=t_i}$ independently across i.

3.2 Entropy-based diversity scores for unconditional generative models

Consider generated samples $x_1, \ldots, x_n \in \mathcal{X}$ drawn i.i.d. from the distribution P_X of an unconditional generative model. For a kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, the (Gram) kernel matrix $K \in \mathbb{R}^{n\times n}$ is

$$K = \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{bmatrix}. \tag{1}$$

A function k is called a positive semidefinite (PSD) kernel if the matrix K is PSD for every $n \in \mathbb{N}$ and any choice of $x_1, \ldots, x_n \in \mathcal{X}$. A widely used example is the Gaussian (RBF) kernel with bandwidth $\sigma > 0$: $k(x, x') = \exp\big(-\|x - x'\|_2^2/2\sigma^2\big)$.

Suppose k is normalized, i.e., $k(x, x) = 1$ for every $x \in \mathcal{X}$. Then the eigenvalues $\lambda_1, \ldots, \lambda_n \geq 0$ of $\frac{1}{n}K$ sum to 1, thus they form a probability distribution. The *von*

Neumann entropy of the unit-trace PSD matrix $\frac{1}{n}K$ is

$$H\big(\tfrac{1}{n}K\big) \;:=\; -\mathrm{Tr}\Big(\tfrac{1}{n}K\log\big(\tfrac{1}{n}K\big)\Big) \;=\; \sum_{i=1}^{n}\lambda_i\log\frac{1}{\lambda_i} \quad (2)$$

Friedman and Dieng (2023) propose using the von Neumann entropy of the normalized kernel matrix to define the *Vendi diversity score*:

$$\mathrm{Vendi}(x_{1:n}) := \exp\Big(H\big(\tfrac{1}{n}K\big)\Big) = \exp\Big(\sum_{i=1}^{n}\lambda_i\log\tfrac{1}{\lambda_i}\Big) \quad (3)$$

Also, Jalali et al. (2023) propose using *order-2 Rényi entropy* of the normalized kernel matrix, $H_2\big(\frac{1}{n}K\big) = \log\big(1/\sum_{i=1}^{n}\lambda_i^2\big)$, to define the *Rényi Kernel Entropy (RKE)* score. Denoting the Frobenius norm by $\|\cdot\|_F$, the following holds since for symmetric $\frac{1}{n}K$, the sum of squared eigenvalues is the squared-Frobenius norm:

$$\mathrm{RKE}(x_{1:n}) \;:=\; \exp\Big(H_2\big(\tfrac{1}{n}K\big)\Big) \;=\; \big\|\frac{1}{n}K\big\|_F^{-2} \quad (4)$$

4 Conditional Vendi and RKE Prompt-Aware Diversity Scores

To extend Vendi and RKE to prompt-aware diversity evaluation, we replace the (unconditional) entropy in these scores with the *conditional* entropy of the kernel matrix of generated samples given the kernel matrix of the corresponding prompts. We follow the matrix-based definitions of Giraldo et al. (2014), where the joint entropy of two PSD matrices A and B is defined via the von Neumann entropy of the trace-normalized Hadamard (elementwise) product $\frac{1}{\mathrm{Tr}(A\odot B)}(A\odot B)$, and the conditional entropy is $H(A\,|\,B) = H(A,B) - H(B)$. The Schur product theorem ensures $A\odot B$ is PSD, and therefore the entropy is well-defined.

To evaluate diversity of generated output X given input text prompt T, we consider a normalized kernel for outputs $k_{\mathcal{X}} : \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ with $k_{\mathcal{X}}(x,x) = 1$ for every $x\in\mathcal{X}$, and a normalized kernel for prompts $k_{\mathcal{T}} : \mathcal{T}\times\mathcal{T}\to\mathbb{R}$ with $k_{\mathcal{T}}(t,t) = 1$ for every $t\in\mathcal{T}$. Given pairs $\{(t_i,x_i)\}_{i=1}^n$, define the kernel matrices

$$K_T = \big[k_{\mathcal{T}}(t_i,t_j)\big]_{i,j=1}^n, \qquad K_X = \big[k_{\mathcal{X}}(x_i,x_j)\big]_{i,j=1}^n.$$

By normalization, the diagonal entries of K_T, K_X, and $K_T\odot K_X$ are all 1, hence $\mathrm{Tr}(K_T) = \mathrm{Tr}(K_X) = \mathrm{Tr}(K_T\odot K_X) = n$. Following the definitions of conditional entropy, the matrix-based conditional entropy specialized for these kernel matrices becomes

$$\begin{aligned} H\big(\tfrac{1}{n}K_X \,\big|\, \tfrac{1}{n}K_T\big) &:= H\big(\tfrac{1}{n}K_X, \tfrac{1}{n}K_T\big) - H\big(\tfrac{1}{n}K_T\big) \quad (5)\\ &= H\big(\tfrac{1}{n}(K_X\odot K_T)\big) - H\big(\tfrac{1}{n}K_T\big)\,. \end{aligned}$$

Conditional-Vendi and Information-Vendi. Replacing the von Neumann entropy in Vendi with the conditional entropy in equation 5, we define

$$\begin{aligned} &\text{Conditional-Vendi}\big(x_{1:n}\,\big|\,t_{1:n}\big) \quad (6)\\ &:= \exp\Big(H\big(\tfrac{1}{n}(K_X\odot K_T)\big) - H\big(\tfrac{1}{n}K_T\big)\Big). \end{aligned}$$

We also define the matrix-based information:

$$\begin{aligned} &\text{Information-Vendi}\big(x_{1:n}; t_{1:n}\big) \quad (7)\\ &:= \exp\Big(H\big(\tfrac{1}{n}K_X\big) + H\big(\tfrac{1}{n}K_T\big) - H\big(\tfrac{1}{n}(K_X\odot K_T)\big)\Big). \end{aligned}$$

These yield the following decomposition of Vendi:

$$\begin{aligned} \mathrm{Vendi}\big(x_{1:n}\big) \;=\; &\text{Conditional-Vendi}\big(x_{1:n}\,\big|\,t_{1:n}\big) \quad (8)\\ &\times \text{Information-Vendi}\big(x_{1:n}\,;\,t_{1:n}\big). \end{aligned}$$

Conditional-RKE and Information-RKE. Similarly, by using the order-2 Rényi entropy $H_2(\cdot)$ in the definitions, we define the Conditional-RKE and Information-RKE scores:

$$\begin{aligned} &\text{Conditional-RKE}\big(x_{1:n}\,\big|\,t_{1:n}\big) \quad (9)\\ &:= \exp\Big(H_2\big(\tfrac{1}{n}(K_X\odot K_T)\big) - H_2\big(\tfrac{1}{n}K_T\big)\Big), \end{aligned}$$

$$\begin{aligned} &\text{Information-RKE}\big(x_{1:n}; t_{1:n}\big) \quad (10)\\ &:= \exp\Big(H_2\big(\tfrac{1}{n}K_X\big) + H_2\big(\tfrac{1}{n}K_T\big) - H_2\big(\tfrac{1}{n}(K_X\odot K_T)\big)\Big). \end{aligned}$$

Considering that for the order-2 entropy of a symmetric PSD M, we have $H_2(M) = \log\big(1/\sum_i\lambda_i(M)^2\big)$ and $\sum_i\lambda_i(M)^2 = \|M\|_F^2$, and thus the above formulations can be written as follows using their unit diagonals:

$$\begin{aligned} \text{Conditional-RKE}(x_{1:n}\,|\,t_{1:n}) &= \frac{\big\|K_T\big\|_F^2}{\big\|K_X\odot K_T\big\|_F^2}, \quad (11)\\ \text{Information-RKE}(x_{1:n}\,;\,t_{1:n}) &= \frac{n^2\cdot\big\|K_X\odot K_T\big\|_F^2}{\big\|K_X\big\|_F^2\cdot\big\|K_T\big\|_F^2}. \end{aligned}$$

Therefore, RKE also admits the product decomposition property we discussed for the Vendi score:

$$\begin{aligned} \mathrm{RKE}\big(x_{1:n}\big) = &\text{Conditional-RKE}\big(x_{1:n}\,\big|\,t_{1:n}\big)\\ &\times \text{Information-RKE}\big(x_{1:n}; t_{1:n}\big) \end{aligned}$$

5 Statistical Convergence of Conditional Diversity Scores

We now establish finite-sample convergence guarantees for the conditional diversity scores introduced in the previous section. Let $(t_i,x_i)_{i=1}^n$ be i.i.d. samples from $P_T\times P_{X|T}$, and let $d_{\mathcal{X}}$ and $d_{\mathcal{T}}$ denote the dimensions of the kernel feature maps of $k_{\mathcal{X}}$, $k_{\mathcal{T}}$. We assume normalized kernels $k_{\mathcal{X}}, k_{\mathcal{T}}$ with unit self-kernel similarity scores and suppose the following spectral boundedness

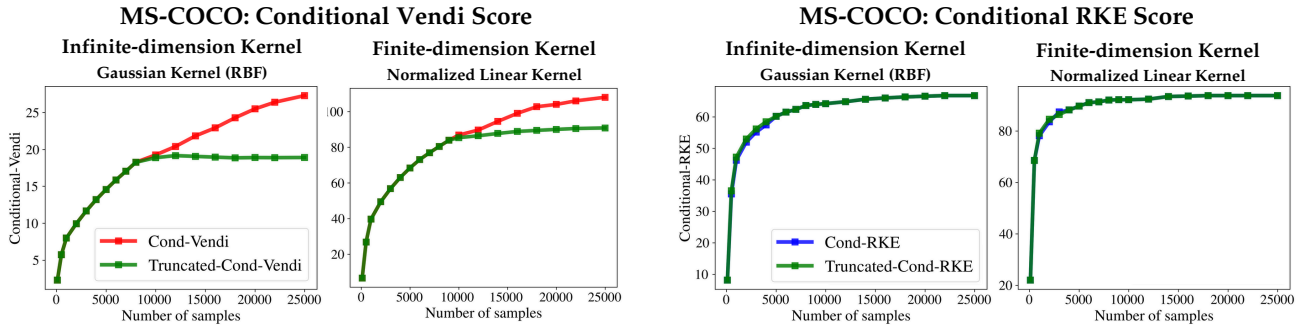


Figure 2: Statistical convergence of Conditional-Vendi and RKE scores with different sample sizes on MS-COCO validation dataset with finite-dimension (normalized) linear kernel and infinite-dimension Gaussian (RBF) kernel.

for the vector of eigenvalues of the population kernel covariance matrices $C_{\mathcal{T}}$, $C_{\mathcal{X},\mathcal{T}}$ of the population distributions P_T and $P_{X|T}$ as

$$0 < m_{\mathcal{T}} \le \|\lambda(C_{\mathcal{T}})\|_2 \le M_{\mathcal{T}},$$
$$0 < m_{\mathcal{X},\mathcal{T}} \le \|\lambda(C_{\mathcal{X},\mathcal{T}})\|_2 \le M_{\mathcal{X},\mathcal{T}}. \tag{12}$$

Convergence of Conditional-Vendi Score. The following theorem proves a convergence bound on the gap between the empirical and population Conditional-Vendi scores. This result extends the concentration bound of Ospanov and Farnia (2025) from the unconditional Vendi score to the conditional case. We defer the theorem proofs to the Appendix, where we also present theoretical results connecting the Conditional-RKE and Conditional-Vendi scores to the component-based aggregation of the unconditional Vendi and RKE scores, given a mixture text distribution with well-separated components.

Theorem 1. *Under normalized kernels and spectral boundedness in equation 12, for every $\delta \in (0,1)$, the following holds with probability at least $1-\delta$:*

$$\Big|\log\Big(\text{Cond-Vendi}(x_{1:n}|t_{1:n})\Big) - \log\Big(\text{Cond-Vendi}(P_{T,X})\Big)\Big| \le \sqrt{\tfrac{20 d_{\mathcal{X}} d_{\mathcal{T}} \log(4/\delta)}{n}} \log(n d_{\mathcal{X}} d_{\mathcal{T}})$$

The factor $d_{\mathcal{X}} d_{\mathcal{T}}$ arises because the Hadamard product $K_X \odot K_T$ corresponds to a tensorized feature space of dimension $d_{\mathcal{X}} \times d_{\mathcal{T}}$. For example, CLIP embeddings with normalized linear kernels ($k_{\text{lin}}(z,z') = \langle z, z'\rangle / \|z\|_2 \|z'\|_2$) lead to $d_{\mathcal{X}} = d_{\mathcal{T}} = 512$, which yields $d_{\mathcal{X}} d_{\mathcal{T}} \approx 2.6 \times 10^5$. This implies the inefficient sample complexity of the Conditional-Vendi score, given the $O(n^3)$ computational cost of the eigendecomposition of the $n \times n$ kernel matrices in computing the score.

Truncated Conditional-Vendi Score. To mitigate the discussed curse of dimensionality, we adopt the spectral truncation technique in (Ospanov and Farnia, 2025) for unconditional entropy measures. For an integer hyperparameter $t \in \mathbb{N}$, let $\lambda^{(t)}(M)$ denote the top-t eigenvalues of M, all shifted with the same positive constant $c = \big(1 - \sum_{i=1}^{t} \lambda_i(M)\big)/t$ to sum up to one, and then the t-truncated entropy of M is defined:

$$H^{(t)}(M) \;:=\; \sum_{i=1}^{t} \lambda_i^{(t)}(M) \log \frac{1}{\lambda_i^{(t)}(M)}$$

We define the truncated Conditional-Vendi score as

$$\text{Conditional-Vendi}^{(t)}(x_{1:n}|t_{1:n}) := \exp\Big(H^{(t)}\big(\tfrac{1}{n}(K_X \odot K_T)\big) - H^{(t)}\big(\tfrac{1}{n}K_T\big)\Big). \tag{13}$$

Theorem 2. *Under normalized kernels and spectral boundedness in (12), for any fixed $t \in \mathbb{N}$ and $\delta \in (0,1)$, the following holds with probability at least $1-\delta$:*

$$\Big|\log\Big(\text{Cond-Vendi}^{(t)}(x_{1:n}|t_{1:n})\Big) - \log\Big(\text{Cond-Vendi}^{(t)}(P_{T,X})\Big)\Big| \le \sqrt{\frac{20\, t\, \log(4/\delta)}{n}} \log(nt)$$

The truncated estimator achieves an $\tilde{O}(\sqrt{t/n})$ rate independent of $d_{\mathcal{X}}, d_{\mathcal{T}}$. In practice, choosing t between 10^3 and 10^4 captures most of the spectral mass while enabling efficient computation via partial eigendecomposition in $O(n^2 t)$ time.

Convergence of Conditional-RKE Score. In contrast, Conditional-RKE avoids dimension dependence as long as the prompt and output underlying entropies are bounded. The closed form in equation 11 depends only on Frobenius norms computable in $O(n^2)$. The following is our concentration bound for the Conditional-RKE score.

Theorem 3. *Under normalized kernels and spectral boundedness in equation 12, for every $\delta \in (0,1)$, the following holds with probability at least $1-\delta$:*

$$\big|\text{Cond-RKE}(x_{1:n}|t_{1:n}) - \text{Cond-RKE}(P_{T,X})\big| \le \frac{32\big(\frac{M_{\mathcal{T}}}{m_{\mathcal{X},\mathcal{T}}}\big)^3\big(\frac{2}{m_{\mathcal{T}}} + \frac{2M_{\mathcal{X},\mathcal{T}}}{m_{\mathcal{T}}^2}\big)}{\sqrt{n}}\Big(1 + \sqrt{2\log\tfrac{4}{\delta}}\Big).$$

Figure 2 illustrates the convergence of the original and truncated (with t=10,000) Conditional-Vendi and Conditional-RKE scores on the MS-COCO (text,image) benchmark using standard CLIP text and DINOv2 image embeddings. The truncated and Conditional-RKE scores stabilize before 15,000 samples, whereas the Conditional-Vendi does not converge within the computationally feasible size n=25,000.

6 Conditional-Vendi Guidance for Text-Guided Diffusion Models

Askari Hemmat et al. (2024) demonstrate that maximizing the Vendi score during sampling can enhance unconditional diversity in (unconditional) diffusion model sample generation. Building on our proposed formulation of Conditional-Vendi, we extend this idea to text-conditioned latent diffusion models (Rombach et al., 2022). Specifically, we employ the *truncated Conditional-Vendi* score as a guidance objective, which encourages each new latent to increase the prompt-aware entropy of the joint kernel spectrum. This approach ensures that generated outputs diversify along dimensions relevant to prompt variability.

Let Conditional-Vendi$^{(t)}(z_{1:n}|t_{1:n})$ denote the truncated Conditional-Vendi score in the latent space $\mathcal{Z}$. At step τ, we augment the classifier-free update (Ho and Salimans, 2022a) with a diversity ascent step:

$$\begin{aligned} \boldsymbol{z}_{\tau-1}^{(n)} \leftarrow \text{Sampler}\big(\boldsymbol{z}_{\tau}^{(n)},\, \hat{\epsilon}_\theta(\boldsymbol{z}_{\tau}^{(n)}, \tau, t_n)\big) \\ + \eta_\tau \nabla_{\boldsymbol{z}^{(n)}} \text{Conditional-Vendi}^{(t)}(z_{1:n} \mid t_{1:n}), \end{aligned} \tag{14}$$

where $\eta_\tau > 0$ is the guidance scale at iteration τ. Equation 14 is the prompt-aware analogue of unconditional Vendi guidance, but uses the truncated spectrum of the joint kernel for scalability. The same construction can also be applied with the Conditional-RKE score as discussed further in the Appendix.

7 Numerical Results

We numerically evaluated Conditional-Vendi and Conditional-RKE scores for four types of conditional generative models: 1) text-to-image, 2) text-to-video, 3) image-captioning, and 4) Large Language Models. For text-to-image, we tested Flux (Lab, 2024), Stable Diffusion 2.1 (Rombach et al., 2022), Stable Diffusion XL (Podell et al., 2024), GigaGAN (Kang et al., 2023), Kandinsky (Razzhigaev et al., 2023), and PixArt (Chen et al., 2023b, 2024b). For video, we used VideoCrafter1 (Chen et al., 2023a), Show-1 (Zhang et al., 2023a), and Open-Sora (Zheng et al., 2024). For image-captioning, we tested BLIP (Li et al., 2022), GIT (Wang et al., 2022), and GPT4o-mini (OpenAI, 2024a). For LLMs, we used Llama (Touvron et al., 2023) and Gemma (Team et al., 2024).

Embeddings used in the evaluation of generative models. Unlike standard embedding-based scores for text-to-image models such as CLIPScore (Hessel et al., 2021), which require the same embedding model for the text and generated image, our proposed scores allow different feature extractors for text and generated samples. In our experiments, we followed (Stein et al., 2023; Kynkäänniemi et al., 2023), to use the DINOv2 (Oquab et al., 2023) embedding for image data. For text data, we used Gemini (Team, 2024) and CLIP (Radford et al., 2021), and for video samples, following the video evaluation literature (Kim et al., 2024; Saito et al., 2020; Unterthiner et al., 2019), we used I3D (Carreira and Zisserman, 2017). To select the bandwidth parameter, we followed the previous works (Jalali et al., 2023; Ospanov et al., 2024) and we set the truncation parameter t to $10,000$ as suggested in (Ospanov and Farnia, 2025). For the detailed experimental setup, we refer to the Appendix.

Convergence Analysis of Conditional-Vendi and Conditional-RKE. To assess the convergence of the Conditional-Vendi and Conditional-RKE scores, we conducted experiments for different sample sizes on samples generated with SDXL and Kandinsky using prompts from the MS-COCO 2014 validation set. We used the cosine similarity for the finite-dimensional kernel and the Gaussian kernel for the infinite-dimensional kernel. Our results, presented in Figure 3, show that for RKE, Conditional-RKE converged, while for Conditional-Vendi, the non-truncated score did not converge; our proposed truncated Conditional-Vendi converged with 15000 samples. Additional results are provided in the Appendix.

Quantifying model-induced diversity via Conditional-Vendi. To illustrate how Conditional-Vendi correlates with the model-induced diversity, we considered an experiment were we generated 10 types of animals using Stable Diffusion XL and used two sets of prompts generated using GPT-4o (OpenAI, 2024a) where in the first set, the types of animals were not specified, while in latter, the animal was explicityly mentioned. As shown in Figure 5, increasing the number of animal types led to growth of the Vendi score, regardless of whether the animal type was mentioned in the prompt or not. However, Conditional-Vendi only increased when the types of animal was not specified in the prompts and in the second case, where the types were mentioned in the prompt, Conditional-Vendi slightly increased showing that it only focused on the diversity coming from the background or other parts of images except the animal types. We provided additional experiments

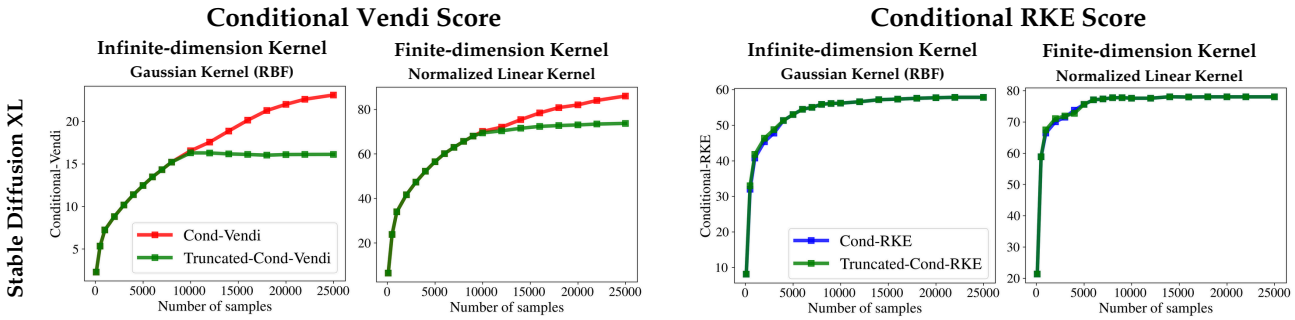


Figure 3: Statistical convergence of Conditional-Vendi scores with different sample sizes on data generated by SDXL using MS-COCO validation set prompts with finite-dimension cosine similarity and infinite-dimension Gaussian kernel. DINOv2 and CLIP embeddings are used for image and text modalities, respectively.

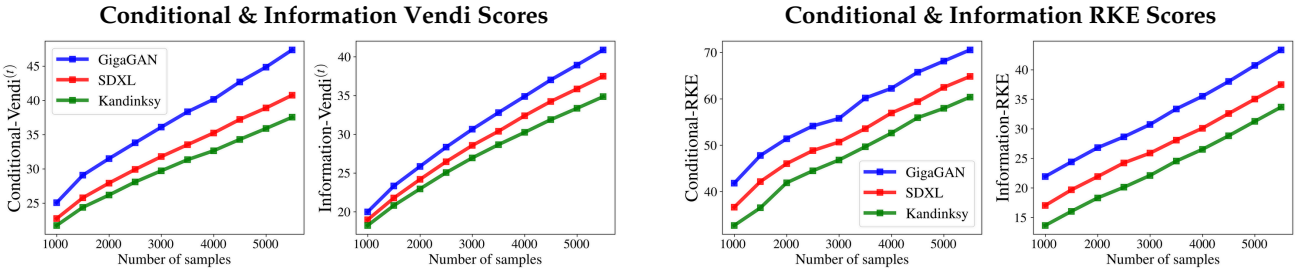


Figure 4: Conditional and Information Vendi and RKE score comparison across text-to-image models. We clustered MS-COCO prompts into k groups and generated images for each cluster center. Within each cluster, we paired prompts with identical images. The results show increasing diversity and stronger correlation as the number of clusters grows, indicating that clusters become more relevant and diverse with finer partitioning.

with different diffusion models and different concepts in the Appendix G.4.

Measuring Conditional-Vendi across prompt types. To measure Conditional-Vendi conditioned on the prompt type, we created 10,000 prompts with different categories using GPT-4o and generated the corresponding images with the text-to-image models. In Figure 6, the top 2 groups of PixArt-α in terms of conditional entropy values are shown. We observed that the "dog" text-based top 3 clusters of images looked more diverse than the image clusters for "airplane"-type prompts. Also, our evaluated Conditional-Vendi score of "dog" texts was significantly higher than that of the "airplane" class. We have reported the scores for three more generative models in the Appendix.

Text-to-Video Model Evaluation. For the experiments on video data, to ensure the fairness of our evaluation, we used VBench samples (Huang et al., 2024), which generated samples belong to the 8 content categories. In Figure 8, we used VideoCrafter-1, Show-1, and Open-Sora-1.2. We observed that VideoCrafter videos look less diverse and, in some cases, may not correlate significantly with the captions when compared to Open-Sora. Confirming this observation, the Conditional-Vendi and Information-Vendi scores were lower for VideoCrafter than those for Open-Sora.

Evaluation of LLMs and Image Captioning Models. To evaluate Conditional-Vendi and RKE on LLMs, we varied the temperature parameter and generated 20K short stories with Llama 2 for each temperature setting. The dataset covered 10 genres, each with 20 distinct subjects and themes. As shown in Table 2, both Conditional-Vendi and RKE increase with higher temperatures, indicating that the outputs become more diverse. Additional numerical results on other LLMs are presented in the Appendix. We also evaluate our proposed scores on various image captioning models in the Appendix.

Conditional-Vendi Guidance for Diverse Image Generation. To demonstrate the advantages of using prompt-aware metrics over unconditional Vendi score, we investigated how they can improve the sample diversity in latent diffusion models. Specifically, we guided the model with both Truncated-Conditional-Vendi and the standard Vendi score, following the method introduced in (Askari Hemmat et al., 2024).

In our experiments, we applied guidance to SD-XL in the latent space rather than in the ambient space. We

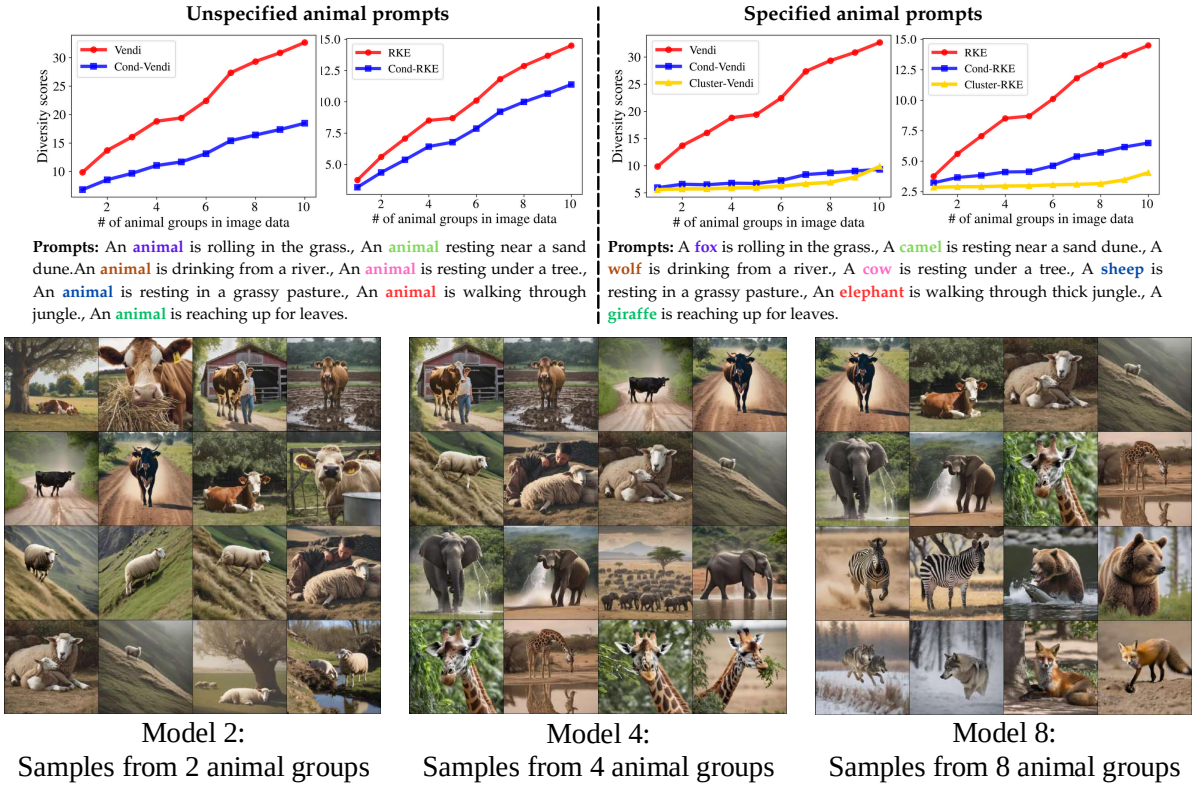


Figure 5: Evaluated Conditional-Vendi and Vendi scores on animal samples generated by SDXL. (Left Plot) We do not specify the animal types in the prompt (Right Plot) we specify the animal types in the prompt.

Table 1: Quantitative comparison of guidance methods on Stable Diffusion XL

Guidance Method	**CLIPScore** ↑	**KD**$\times 10^2$ ↓	**Cond-Vendi**$_{\text{DINOv2}}$ ↑	**Vendi**$_{\text{DINOv2}}$ ↑	**In-batch Sim.**$\times 10^2$ ↓
Vendi$_{\text{Latent}}$	30.41	35.40	29.85	309.45	80.06
Conditional-Vendi$_{\text{Latent}}$	30.47	29.14	32.45	325.20	78.02

Table 2: Conditional Vendi and RKE Scores evaluated for Llama 2 with different temperature parameters.

Method	$T = 0.4$	$T = 0.7$	$T = 1.0$	$T = 1.3$
Conditional-Vendi	45.38	46.73	48.26	49.60
Conditional-RKE	41.45	44.67	46.29	48.45

found that applying guidance in the latent space can improve image diversity and quality while also significantly reducing computational costs.

Figure 7 shows qualitative results using Stable Diffusion-XL, highlighting how prompt-aware guidance can lead to more relevant image generations. We also provide quantitative results of Vendi vs. Conditional-Vendi guidance methods on Stable Diffusion XL in Table 1, indicating that Conditional-Vendi guidance improved the sample diversity (Vendi and in-batch similarity) while maintaining text-image alignment in the CLIPScore and KD metrics.

8 Conclusion and Limitations

We introduced Conditional-Vendi and Conditional-RKE as prompt-aware diversity scores for generative models, extending Vendi and RKE to kernel matrices of prompt–output pairs to separate model-induced diversity from prompt-induced variation. Our theoretical analysis established finite-sample convergence guarantees and a truncated-spectrum approximation for scalability, while experiments on benchmark datasets showed that the proposed scores capture intuitive aspects of diversity that unconditional metrics overlook. A limitation of the approach would be its reliance on high-quality embeddings, which may introduce biases for an arbitrarily selected embedding, as also discussed in (Stein et al., 2023). Beyond evaluation, we demonstrated that Conditional-Vendi can also guide generation toward diverse outputs, and we highlight future applications in using these scores as regularizers for training or as fairness measures to assess demographic consistency in generative models.

Figure 6: Quantifying image diversity of PixArt-α–generated outputs for the top-2 prompt clusters (from 10,000 GPT-4o–generated prompts).

Figure 7: Qualitative comparison of Conditional-Vendi score guidance vs. Vendi score guidance using SD-XL.

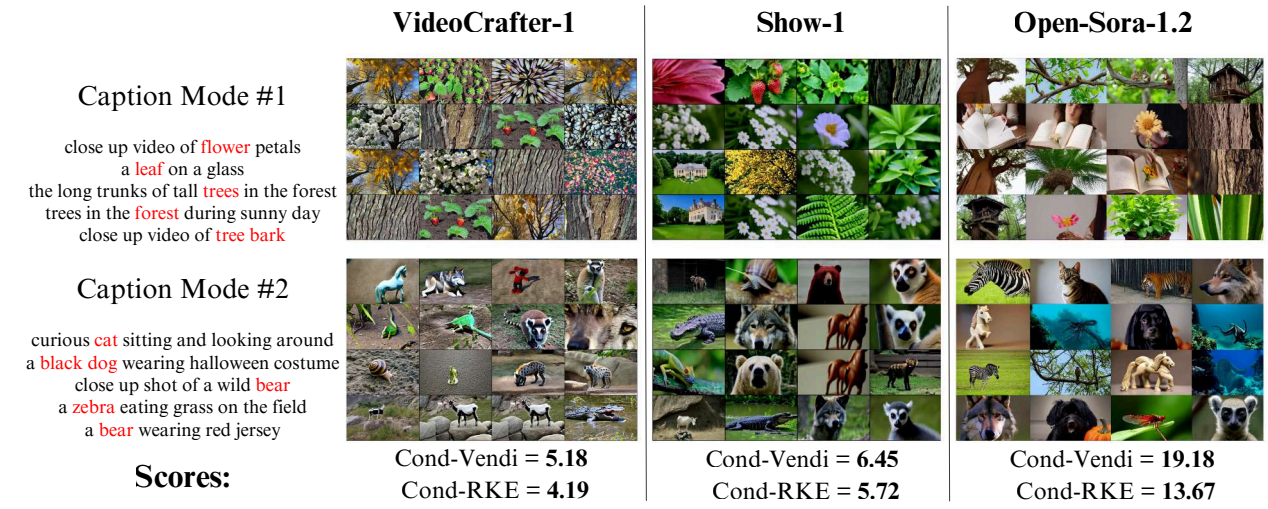


Figure 8: Measuring Conditional-Vendi and Information-Vendi for text-to-video models

Acknowledgments

The work of Farzan Farnia is partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China, Project 14209920, and is partially supported by CUHK Direct Research Grants with CUHK Project No. 4055164 and 4937054. The work of Amin Gohari is supported by the Hong Kong Research Grants Council (RGC) under grant number 14310025. The work is also supported by a grant under 1+1+1 CUHK-CUHK(SZ)-GDSTC Joint Collaboration Fund. Finally, the authors would like to sincerely thank the anonymous reviewers for their insightful feedback and constructive suggestions.

References

Alaa, A., Van Breugel, B., Saveliev, E. S., and van der Schaar, M. (2022). How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International Conference on Machine Learning*, pages 290–306. PMLR.

Askari Hemmat, R., Hall, M., Sun, A., Ross, C., Drozdzal, M., and Romero-Soriano, A. (2024). Improving geo-diversity of generated images with contextualized vendi score guidance. *arXiv e-prints*, pages arXiv–2406.

Astolfi, P., Careil, M., Hall, M., Mañas, O., Muckley, M., Verbeek, J., Soriano, A. R., and Drozdzal, M. (2024). Consistency-diversity-realism pareto fronts of conditional image generative models.

Bansal, A., Chu, H.-M., Schwarzschild, A., Sengupta, S., Goldblum, M., Geiping, J., and Goldstein, T. (2023). Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 843–852.

Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. (2018). Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*.

Borji, A. (2022). Pros and cons of gan evaluation measures: New developments. *Computer Vision and Image Understanding*, 215:103329.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

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

Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., Weng, C., and Shan, Y. (2023a). Videocrafter1: Open diffusion models for high-quality video generation.

Chen, J., Ge, C., Xie, E., Wu, Y., Yao, L., Ren, X., Wang, Z., Luo, P., Lu, H., and Li, Z. (2024a). Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation.

Chen, J., Wu, Y., Luo, S., Xie, E., Paul, S., Luo, P., Zhao, H., and Li, Z. (2024b). Pixart-δ: Fast and controllable image generation with latent consistency models.

Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., and Li, Z. (2023b). Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis.

Corso, G., Xu, Y., Bortoli, V. D., Barzilay, R., and Jaakkola, T. S. (2024). Particle Guidance: non-I.I.D. diverse sampling with diffusion models. In *The Twelfth International Conference on Learning Representations*.

Dhariwal, P. and Nichol, A. (2021). Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794.

Friedman, D. and Dieng, A. B. (2023). The vendi score: A diversity evaluation metric for machine learning. *Transactions on machine learning research*.

Gaviria Rojas, W., Diamos, S., Kini, K., Kanter, D., Janapa Reddi, V., and Coleman, C. (2022). The Dollar Street Dataset: Images representing the geographic and socioeconomic diversity of the world. In *Advances in Neural Information Processing Systems*, volume 35, pages 12979–12990. Curran Associates, Inc.

Giraldo, L. G. S., Rao, M., and Principe, J. C. (2014). Measures of entropy from data using infinitely divisible kernels. *IEEE Transactions on Information Theory*, 61(1):535–548.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680.

Han, J., Choi, H., Choi, Y., Kim, J., Ha, J.-W., and Choi, J. (2023). Rarity score : A new metric to evaluate the uncommonness of synthesized images. In *The Eleventh International Conference on Learning Representations*.

He, Y., Murata, N., Lai, C.-H., Takida, Y., Uesaka, T., Kim, D., Liao, W.-H., Mitsufuji, Y., Kolter, J. Z., Salakhutdinov, R., and Ermon, S. (2024). Manifold preserving guided diffusion. In *The Twelfth International Conference on Learning Representations*.

Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., and Choi, Y. (2021). CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., and Salimans, T. (2022a). Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., and Salimans, T. (2022b). Video diffusion models. In *Advances in Neural Information Processing Systems*, volume 35.

Ho, J. and Salimans, T. (2022a). Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Ho, J. and Salimans, T. (2022b). Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Hu, X., Leung, H.-f., and Farnia, F. (2025a). 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)*.

Hu, X., Leung, H.-f., and Farnia, F. (2025b). 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)*.

Hu, X., Pick, L., Leung, H.-f., and Farnia, F. (2025c). Promptwise: Online learning for cost-aware prompt assignment in generative models. *arXiv preprint arXiv:2505.18901*.

Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., Wang, Y., Chen, X., Wang, L., Lin, D., Qiao, Y., and Liu, Z. (2024). VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Jafari, D. and Farnia, F. (2026). Diversity-aware online prompt assignment to generative models. In *The Fourteenth International Conference on Learning Representations*.

Jalali, M., Li, C. T., and Farnia, F. (2023). An information-theoretic evaluation of generative models in learning multi-modal distributions. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jiralerspong, M., Bose, J., Gemp, I., Qin, C., Bachrach, Y., and Gidel, G. (2023). Feature likelihood score: Evaluating the generalization of generative models using samples. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Kang, M., Zhu, J.-Y., Zhang, R., Park, J., Shechtman, E., Paris, S., and Park, T. (2023). Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kannen, N., Ahmad, A., Andreetto, M., Prabhakaran, V., Prabhu, U., Dieng, A. B., Bhattacharyya, P., and Dave, S. (2024). Beyond aesthetics: Cultural competence in text-to-image models.

Kim, G., Kwon, T., and Ye, J. C. (2022a). Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2426–2435.

Kim, J.-H., Kim, Y., Lee, J., Yoo, K. M., and Lee, S.-W. (2022b). Mutual information divergence: A unified metric for multimodal generative models. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

Kim, P. J., Kim, S., and Yoo, J. (2024). STREAM: Spatio-temporal evaluation and analysis metric for video generative models. In *The Twelfth International Conference on Learning Representations*.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kynkäänniemi, T., Karras, T., Aittala, M., Aila, T., and Lehtinen, J. (2023). The role of imagenet classes in fréchet inception distance. In *The Eleventh International Conference on Learning Representations*.

Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., and Aila, T. (2019). Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 32.

Lab, B. F. (2024). FLUX: A diffusion-based text-to-image (T2I) model. `https://github.com/blackforestlab/flux`. Accessed: 2024-09.

Lee, T., Yasunaga, M., Meng, C., Mai, Y., Park, J. S., Gupta, A., Zhang, Y., Narayanan, D., Teufel, H. B., Bellagente, M., Kang, M., Park, T., Leskovec, J., Zhu, J.-Y., Fei-Fei, L., Wu, J., Ermon, S., and Liang, P. (2023). Holistic evaluation of text-to-image models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Li, J., Li, D., Xiong, C., and Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

Liu, X., Park, D. H., Azadi, S., Zhang, G., Chopikyan, A., Hu, Y., Shi, H., Rohrbach, A., and Darrell, T. (2023). More control for free! image synthesis with semantic diffusion guidance. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 289–299.

Miao, Z., Wang, J., Wang, Z., Yang, Z., Wang, L., Qiu, Q., and Liu, Z. (2024). Training diffusion models towards diverse image generation with reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10844–10853.

Microsoft, :, Abouelenin, A., Ashfaq, A., Atkinson, A., Awadalla, H., Bach, N., Bao, J., Benhaim, A., and Others (2025). Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras.

Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., and Shan, Y. (2024). T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 4296–4304.

Naeem, M. F., Oh, S. J., Uh, Y., Choi, Y., and Yoo, J. (2020). Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, pages 7176–7185. PMLR.

Nguyen, Q. and Dieng, A. B. (2024). Quality-weighted vendi scores and their application to diverse experimental design. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Nichol, A. Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., and Chen, M. (2022). GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 16784–16804. PMLR.

OpenAI (2024a). GPT-4o mini: advancing cost-efficient intelligence.

OpenAI (2024b). Sora: A text-to-video model. *arXiv preprint arXiv:2402.17177*.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.-Y., Xu, H., Sharma, V., Li, S.-W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. (2023). Dinov2: Learning robust visual features without supervision.

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

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

Ospanov, A., Zhang, J., Jalali, M., Cao, X., Bogdanov, A., and Farnia, F. (2024). Towards a scalable reference-free evaluation of generative models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Pasarkar, A. and Dieng, A. B. (2024). Cousins of the vendi score: A family of similarity-based diversity metrics for science and machine learning. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. (2024). SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision.

Ramaswamy, V. V., Lin, S. Y., Zhao, D., Adcock, A. B., van der Maaten, L., Ghadiyaram, D., and Russakovsky, O. (2023). GeoDE: a geographically diverse

evaluation dataset for object recognition. In *NeurIPS Datasets and Benchmarks*.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Razzhigaev, A., Shakhmatov, A., Maltseva, A., Arkhipkin, V., Pavlov, I., Ryabov, I., Kuts, A., Panchenko, A., Kuznetsov, A., and Dimitrov, D. (2023). Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion.

Rezaei, P., Farnia, F., and Li, C. T. (2025). 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*.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2021). High-resolution image synthesis with latent diffusion models.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. (2023). Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510.

Sadat, S., Kansy, M., Hilliges, O., and Weber, R. M. (2025). No training, no problem: Rethinking classifier-free guidance for diffusion models. In *The Thirteenth International Conference on Learning Representations*.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*.

Saito, M., Saito, S., Koyama, M., and Kobayashi, S. (2020). Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal gan. *International Journal of Computer Vision*, 128(10):2586–2606.

Sajjadi, M. S., Bachem, O., Lucic, M., Bousquet, O., and Gelly, S. (2018). Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., and Chen, X. (2016). Improved techniques for training GANs. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Sehwag, V., Hazirbas, C., Gordo, A., Ozgenel, F., and Ferrer, C. C. (2022). Generating high fidelity data from low-density regions using diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11482–11491.

Stein, G., Cresswell, J. C., Hosseinzadeh, R., Sui, Y., Ross, B. L., Villecroze, V., Liu, Z., Caterini, A. L., Taylor, J. E. T., and Loaiza-Ganem, G. (2023). Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models.

Sutherland, D. J., Strathmann, H., Arbel, M., and Gretton, A. (2018). Efficient and principled score estimation with nyström kernel exponential families. In *International Conference on Artificial Intelligence and Statistics*, pages 652–660. PMLR.

Team, G. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., Tafti, P., Hussenot, L., Sessa, P. G., Chowdhery, A., Roberts, A., Barua, A., Botev, A., Castro-Ros, A., Slone, A., Héliou, A., Tacchetti, A., Bulanova, A., Paterson, A., Tsai, B., Shahriari, B., Lan, C. L., A., C., et al. (2024). Gemma: Open models based on gemini research and technology.

Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-or, D., and Bermano, A. H. (2023). Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*.

Touvron, H., Martin, L., Stone, K. R., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., and Gelly, S. (2019). FVD: A new metric for video generation.

Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C., and Wang, L. (2022). GIT: A generative image-to-text transformer for vision and language.

Wang, Z., Farzan, F., Lin, Z., Shen, Y., and Yu, B. (2025). On the distributed evaluation of generative models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 7644–7653.

Ye, H., Lin, H., Han, J., Xu, M., Liu, S., Liang, Y., Ma, J., Zou, J. Y., and Ermon, S. (2024). Tfg: Unified training-free guidance for diffusion models. *Advances in Neural Information Processing Systems*, 37:22370–22417.

Yu, J., Wang, Y., Zhao, C., Ghanem, B., and Zhang, J. (2023). Freedom: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23174–23184.

Zhang, D. J., Wu, J. Z., Liu, J.-W., Zhao, R., Ran, L., Gu, Y., Gao, D., and Shou, M. Z. (2023a). Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*.

Zhang, J., Jalali, M., Li, C. T., and Farnia, F. (2025). Unveiling differences in generative models: A scalable differential clustering approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, J., Li, C. T., and Farnia, F. (2024). An interpretable evaluation of entropy-based novelty of generative models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 59148–59172. PMLR.

Zhang, L., Rao, A., and Agrawala, M. (2023b). Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847.

Zhao, M., Bao, F., Li, C., and Zhu, J. (2022). Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. *Advances in Neural Information Processing Systems*, 35:3609–3623.

Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., and You, Y. (2024). Open-Sora: Democratizing efficient video production for all.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials: Conditional Vendi Score: Prompt-Aware Diversity Evaluation for Text-Guided Generative AI Models

A Proofs of Theorems in Section 5

A.1 Formal Statement of Assumptions and Theorems

Assumption 1 (Normalized kernels). *$k_{\mathcal{X}}$ and $k_{\mathcal{T}}$ are normalized kernel functions satisfying $k_{\mathcal{X}}(x,x) = k_{\mathcal{T}}(t,t) = 1$ for all $x \in \mathcal{X}$ and $t \in \mathcal{T}$.*

Assumption 2 (Population kernel covariance norm bounds). *There exist positive constants $m_{\mathcal{X},\mathcal{T}}, M_{\mathcal{X},\mathcal{T}}, m_{\mathcal{T}}, M_{\mathcal{T}}$ such that, for the Hilbert-Schmidt norm of population kernel covaraince matrices $C_{\mathcal{T}} = \mathbb{E}_{t\sim P_{\mathcal{T}}}[\phi_{\mathcal{T}}(t)\phi_{\mathcal{T}}(t)^\top]$ and $C_{\mathcal{X},\mathcal{T}} = \mathbb{E}_{t\sim P_{\mathcal{T}}, x\sim P_{\mathcal{X}|\mathcal{T}t}}[(\phi_{\mathcal{T}}(t)\otimes\phi_{\mathcal{X}}(x))(\phi_{\mathcal{T}}(t)\otimes\phi_{\mathcal{X}}(x))^\top]$,*

$$0 < m_{\mathcal{X},\mathcal{T}} \le \|C_{\mathcal{X},\mathcal{T}}\|_{\text{HS}} \le M_{\mathcal{X},\mathcal{T}}, \qquad 0 < m_{\mathcal{T}} \le \|C_{\mathcal{T}}\|_{\text{HS}} \le M_{\mathcal{T}}.$$

Note that the above assumption is identical to the following inequalities for the vector of the eigenvalues of the kernel covariance matrices, denoted by $\widetilde{\boldsymbol{\lambda}}_{\mathcal{T}}$ and $\widetilde{\boldsymbol{\lambda}}_{\mathcal{X},\mathcal{T}}$,

$$0 < m_{\mathcal{X},\mathcal{T}} \le \|\widetilde{\boldsymbol{\lambda}}_{\mathcal{X},\mathcal{T}}\|_2 \le M_{\mathcal{X},\mathcal{T}}, \qquad 0 < m_{\mathcal{T}} \le \|\widetilde{\boldsymbol{\lambda}}_{\mathcal{T}}\|_2 \le M_{\mathcal{T}}.$$

Also, we define and use the condition number $L := \frac{m_{\mathcal{X},\mathcal{T}}}{M_{\mathcal{T}}}$ in our analysis.

Theorem 1 (Conditional-Vendi convergence). *Suppose Assumptions 1-2 hold with $d_{\mathcal{X}} < \infty$ and $d_{\mathcal{T}} < \infty$. For every $\delta \in (0,1)$ such that $n \ge 4e^2\big(1+\sqrt{2\log\frac{4}{\delta}}\big)^2$, then the following holds with probability at least $1-\delta$,*

$$\begin{aligned}
&\Big|\log \text{Conditional-Vendi}(x_{1:n}|t_{1:n}) - \log \text{Conditional-Vendi}(P)\Big| \\
&\le \frac{1}{\sqrt{n}}\Big(1+\sqrt{2\log\tfrac{4}{\delta}}\Big)\Big[\sqrt{d_{\mathcal{X}}d_{\mathcal{T}}}\log\big(nd_{\mathcal{X}}d_{\mathcal{T}}\big) + \sqrt{d_{\mathcal{T}}}\log\big(nd_{\mathcal{T}}\big)\Big] \\
&\le \sqrt{\tfrac{20\,d_{\mathcal{X}}d_{\mathcal{T}}\log(4/\delta)}{n}}\log\big(nd_{\mathcal{X}}d_{\mathcal{T}}\big)
\end{aligned}$$

The final inequality holds as $d_{\mathcal{X}} \ge 1$ and for every $0 < \delta < 1$, we have $(\sqrt{5}-\sqrt{2})\log(4/\delta) \ge 1$.

Theorem 2 (Truncated Conditional-Vendi convergence). *Suppose Assumptions 1-2 hold. Fix a truncation level $t \in \mathbb{N}$ and truncate both the joint and prompt spectra to their top-t components with renormalization. For every $\delta \in (0,1)$ satisfying $n \ge 4e^2\big(1+\sqrt{2\log\frac{4}{\delta}}\big)^2$, we have the following with probability at least $1-\delta$,*

$$\Big|\log \text{Conditional-Vendi}^{(t)}(x_{1:n}|t_{1:n}) - \log \text{Conditional-Vendi}^{(t)}(P)\Big| \le \sqrt{\frac{t}{n}}\Big(2+\sqrt{8\log\tfrac{4}{\delta}}\Big)\log\big(nt\big) \le \sqrt{\frac{20t\log(4/\delta)}{n}}\log(nt)$$

The final inequality holds as for every $0 < \delta < 1$, we have $(\sqrt{20}-\sqrt{8})\log(4/\delta) \ge 2$.

Theorem 3 (Conditional-RKE convergence). *Suppose Assumptions 1-2 hold. Let $C_0 = \frac{2}{m_{\mathcal{T}}} + \frac{2M_{\mathcal{X},\mathcal{T}}}{m_{\mathcal{T}}^2}$. For every $\delta > 0$ that satisfies $n \ge 16\big(1+\sqrt{2\log\frac{4}{\delta}}\big)^2 \max\big\{\frac{1}{m_{\mathcal{T}}^2}, \frac{4C_0^2}{L^2}\big\}$, then the following will hold with probability at least $1-\delta$*

$$\Big|\,\text{Conditional-RKE}(x_{1:n}|t_{1:n}) - \text{Conditional-RKE}(P)\,\Big| \;\le\; \frac{32C_0}{L^3\sqrt{n}}\Big(1+\sqrt{2\log\tfrac{4}{\delta}}\Big).$$

A.2 Auxiliary Lemmas and Propositions

The following are the propositions and lemmas we utilize to prove the theorems.

Proposition 1. *Under Assumption 1, let $K_{\mathcal{X}}, K_{\mathcal{T}} \in \mathbb{R}^{n\times n}$ denote kernel matrices on $\{x_i\}$ and $\{t_i\}$. The normalized Hadamard product $\frac{1}{n}(K_{\mathcal{X}} \odot K_{\mathcal{T}})$ and the operator $\widehat{C}_{\mathcal{X},\mathcal{T}}$ share the same multiset of nonzero eigenvalues.*

Proof. Let $\phi_{\mathcal{X}}$ and $\phi_{\mathcal{T}}$ denote feature maps with $\|\phi_{\mathcal{X}}(x)\| = \|\phi_{\mathcal{T}}(t)\| = 1$. We define $\phi_{\mathcal{X},\mathcal{T}}([x,t]) := \phi_{\mathcal{X}}(x) \otimes \phi_{\mathcal{T}}(t)$. It follows that

$$\langle \phi_{\mathcal{X},\mathcal{T}}([x,t]), \phi_{\mathcal{X},\mathcal{T}}([x',t'])\rangle = k_{\mathcal{X}}(x,x')\, k_{\mathcal{T}}(t,t').$$

Let $\Phi_{\mathcal{X},\mathcal{T}} \in \mathbb{R}^{n\times D}$ be the matrix that stacks the row vectors $\phi_{\mathcal{X},\mathcal{T}}([x_i,t_i])^\top$. We then have $\frac{1}{n}(K_{\mathcal{X}} \odot K_{\mathcal{T}}) = \frac{1}{n}\Phi_{\mathcal{X},\mathcal{T}}\Phi_{\mathcal{X},\mathcal{T}}^\top$. The nonzero eigenvalues of $\frac{1}{n}\Phi_{\mathcal{X},\mathcal{T}}\Phi_{\mathcal{X},\mathcal{T}}^\top$ and $\frac{1}{n}\Phi_{\mathcal{X},\mathcal{T}}^\top\Phi_{\mathcal{X},\mathcal{T}} = \widehat{C}_{\mathcal{X},\mathcal{T}}$ coincide, which completes the proof. □

Corollary 1. *Under Assumption 1, the conditional von Neumann entropy satisfies*

$$H(X|T) = H(\widehat{C}_{\mathcal{X},\mathcal{T}}) - H(\widehat{C}_{\mathcal{T}}),$$

where $H(\cdot)$ denotes the von Neumann entropy of the nonzero spectrum. For the RKE case, we consider the order-2 Rényi entropy $H_2(M) = \log(1/\|M\|_F^2)$, resulting in $H_2(X|T) = H_2(\widehat{C}_{\mathcal{X},\mathcal{T}}) - H_2(\widehat{C}_{\mathcal{T}})$.

Lemma 1. *Let $\beta < 0$ and define $f(u) = u^\beta$ on $[u_0,\infty)$ with $u_0 > 0$. If $z, w \geq u_0$ and $|z-w| \leq \epsilon$, then*

$$\big|f(z) - f(w)\big| \leq |\beta|\, u_0^{\beta-1}\, \epsilon.$$

Proof. By the mean value theorem there exists Γ between z and w such that $|f(z) - f(w)| = |f'(\Gamma)|\,|z-w| = |\beta|\,\Gamma^{\beta-1}|z-w|$. Since $\beta < 0$, we have $\Gamma^{\beta-1} \leq u_0^{\beta-1}$, which completes the proof of the lemma. □

Lemma 2 (Lemma 2 in (Ospanov and Farnia, 2025))**.** *If $a, b \in [0,1]$ satisfy $|b-a| \leq \frac{1}{e}$, then*

$$\Big|a\log\big(\frac{1}{a}\big) - b\log\big(\frac{1}{b}\big)\Big| \;\leq\; |b-a|\log\big(\frac{1}{|b-a|}\big)$$

Lemma 3 (Lemma 3 in (Ospanov and Farnia, 2025))**.** *If $\mathbf{u} \in \mathbb{R}_+^d$ satisfies $\|\mathbf{u}\|_2 \leq \epsilon \leq \frac{1}{e}$ for some $\epsilon > 0$, then*

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

A.3 Auxiliary Matrix-based Concentration Bounds

Lemma 4. *Under Assumption 1, for every $\delta \in (0,1)$, with probability at least $1-\delta$, the following hold simultaneously*

$$\big\|\widehat{C}_{\mathcal{X},\mathcal{T}} - C_{\mathcal{X},\mathcal{T}}\big\|_{\mathrm{HS}} \;\leq\; \frac{2}{\sqrt{n}}\Big(1 + \sqrt{2\log\tfrac{4}{\delta}}\Big), \qquad \big\|\widehat{C}_{\mathcal{T}} - C_{\mathcal{T}}\big\|_{\mathrm{HS}} \;\leq\; \frac{2}{\sqrt{n}}\Big(1 + \sqrt{2\log\tfrac{4}{\delta}}\Big).$$

Proof. Given i.i.d. pairs $(t_i, x_i)_{i=1}^n \sim P_T \times P_{X|T}$, we define the empirical covariance operator

$$\widehat{C}_{\mathcal{X},\mathcal{T}} = \frac{1}{n}\sum_{i=1}^{n} \phi_{\mathcal{X},\mathcal{T}}([x_i,t_i])\phi_{\mathcal{X},\mathcal{T}}([x_i,t_i])^\top$$

and its population counterpart $C_{\mathcal{X},\mathcal{T}} = \mathbb{E}[\phi_{\mathcal{X},\mathcal{T}}([X,T])\phi_{\mathcal{X},\mathcal{T}}([X,T])^\top]$. We then define the centered random operator $Z_i^{(\mathcal{X},\mathcal{T})} := \phi_{\mathcal{X},\mathcal{T}}([x_i,t_i])\phi_{\mathcal{X},\mathcal{T}}([x_i,t_i])^\top - C_{\mathcal{X},\mathcal{T}}$. We observe that $\mathbb{E}[Z_i^{(\mathcal{X},\mathcal{T})}] = 0$ and, by the normalization property, we have $\|Z_i^{(\mathcal{X},\mathcal{T})}\|_{\mathrm{HS}} \leq 2$. Applying the Hoeffding inequality for Hilbert-Schmidt operators (Sutherland et al., 2018, Lemma 11) with $L = 2$ yields the following bound with probability at least $1 - \frac{\delta}{2}$:

$$\big\|\widehat{C}_{\mathcal{X},\mathcal{T}} - C_{\mathcal{X},\mathcal{T}}\big\|_{\mathrm{HS}} \;\leq\; \frac{2}{\sqrt{n}}\Big(1 + \sqrt{2\log\tfrac{4}{\delta}}\Big)$$

The same argument applies to the prompt operator $C_{\mathcal{T}}$ and its empirical $\widehat{C}_{\mathcal{T}}$ to show with probability at least $1-\frac{\delta}{2}$:

$$\|\widehat{C}_{\mathcal{T}} - C_{\mathcal{T}}\|_{\mathrm{HS}} \;\le\; \frac{2}{\sqrt{n}}\Big(1+\sqrt{2\log\tfrac{4}{\delta}}\Big)$$

Using a union bound shows that the above inequalities will simultaneously hold with probability at least $1-\delta$. □

Theorem 4. *Under Assumptions 1-2, for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\|\widehat{\boldsymbol{\lambda}}_{\mathcal{X},\mathcal{T}} - \widetilde{\boldsymbol{\lambda}}_{\mathcal{X},\mathcal{T}}\|_2 \;\le\; \frac{2}{\sqrt{n}}\Big(1+\sqrt{2\log\tfrac{4}{\delta}}\Big), \qquad \|\widehat{\boldsymbol{\lambda}}_{\mathcal{T}} - \widetilde{\boldsymbol{\lambda}}_{\mathcal{T}}\|_2 \;\le\; \frac{2}{\sqrt{n}}\Big(1+\sqrt{2\log\tfrac{4}{\delta}}\Big).$$

Proof. The application of Hoffman–Wielandt inequality shows that $\|\widehat{\boldsymbol{\lambda}}_{\mathcal{X},\mathcal{T}} - \widetilde{\boldsymbol{\lambda}}_{\mathcal{X},\mathcal{T}}\|_2 \le \|\widehat{C}_{\mathcal{X},\mathcal{T}} - C_{\mathcal{X},\mathcal{T}}\|_{\mathrm{HS}}$and $\|\widehat{\boldsymbol{\lambda}}_{\mathcal{T}} - \widetilde{\boldsymbol{\lambda}}_{\mathcal{T}}\|_2 \le \|\widehat{C}_{\mathcal{T}} - C_{\mathcal{T}}\|_{\mathrm{HS}}$. The result follows immediately from applying Lemma 4. □

A.4 Theorem Proofs

A.4.1 Proof of Theorem 1

We define $\varepsilon'_\delta = \frac{2}{\sqrt{n}}(1+\sqrt{2\log\frac{4}{\delta}})$. The logarithm of the Conditional-Vendi score is given by

$$\log \text{Conditional-Vendi}(x_{1:n}|t_{1:n}) = H(\widehat{C}_{\mathcal{X},\mathcal{T}}) - H(\widehat{C}_{\mathcal{T}}),$$

where $H(\cdot)$ denotes the von Neumann entropy. We note that our sample size condition ensures $\varepsilon'_\delta \le 1/e$.

By Theorem 4, we have $\|\widehat{\boldsymbol{\lambda}}_{\mathcal{X},\mathcal{T}} - \widetilde{\boldsymbol{\lambda}}_{\mathcal{X},\mathcal{T}}\|_2 \le \varepsilon'_\delta$ and $\|\widehat{\boldsymbol{\lambda}}_{\mathcal{T}} - \widetilde{\boldsymbol{\lambda}}_{\mathcal{T}}\|_2 \le \varepsilon'_\delta$ with probability at least $1-\delta$. We now apply Lemma 2 coordinatewise to each eigenvalue difference, followed by Lemma 3 to bound the sum. This yields

$$|H(\widehat{C}_{\mathcal{X},\mathcal{T}}) - H(C_{\mathcal{X},\mathcal{T}})| \le \varepsilon'_\delta\sqrt{d_{\mathcal{X}}d_{\mathcal{T}}}\log\Big(\tfrac{\sqrt{d_{\mathcal{X}}d_{\mathcal{T}}}}{\varepsilon'_\delta}\Big),$$

and similarly,

$$|H(\widehat{C}_{\mathcal{T}}) - H(C_{\mathcal{T}})| \le \varepsilon'_\delta\sqrt{d_{\mathcal{T}}}\log\Big(\tfrac{\sqrt{d_{\mathcal{T}}}}{\varepsilon'_\delta}\Big).$$

The triangle inequality then yields the desired bound, noting that by definition $\epsilon'_\delta \ge \frac{2}{\sqrt{n}}$ will automatically hold and therefore $\log(\frac{C}{\epsilon'_\delta}) \le \log(\frac{C\sqrt{n}}{2})$ for every $C > 0$.

A.4.2 Proof of Theorem 2

We use the notation $\widehat{\boldsymbol{\lambda}}_{\mathcal{X},\mathcal{T}}$ and $\widetilde{\boldsymbol{\lambda}}_{\mathcal{X},\mathcal{T}}$ (resp. $\widehat{\boldsymbol{\lambda}}_{\mathcal{T}}$ and $\widetilde{\boldsymbol{\lambda}}_{\mathcal{T}}$) for the empirical and population eigenvalue vectors of the joint (resp. prompt) operator, each sorted in nonincreasing order and summing to 1. Fix a truncation level $t \in \mathbb{N}$ and define the t-truncated *and renormalized* vectors by keeping the top-t coordinates and then rescaling to unit ℓ_1 mass.

Lemma A. Let $\mathbf{v} \in [0,1]^d$ with $\mathbf{1}^\top\mathbf{v} = 1$ and let $S_t = \sum_{i=1}^t v_i$. Define $\mathbf{v}^{(t)} \in [0,1]^d$ by

$$v_i^{(t)} = \begin{cases} v_i + \frac{1-S_t}{t}, & i \le t, \\ 0, & i > t. \end{cases}$$

Then $\mathbf{v}^{(t)}$ is the (Euclidean) projection of $\mathbf{v}$ onto the convex set

$$\Delta_t := \Big\{\, \mathbf{u} \in [0,1]^d :\; u_i = 0 \;(i > t),\; \sum_{i=1}^t u_i = 1 \,\Big\}.$$

Proof of Lemma A. Consider the convex program

$$\min_{\mathbf{u}\in\mathbb{R}^t} \sum_{i=1}^t (u_i - v_i)^2 \quad \text{s.t.} \quad u_i \ge 0 \;(i \le t),\; \sum_{i=1}^t u_i = 1.$$

With Lagrangian $L(\mathbf{u},\lambda,\boldsymbol{\mu}) = \sum_{i=1}^{t}(u_i - v_i)^2 + \lambda\big(\sum_{i=1}^{t} u_i - 1\big) - \sum_{i=1}^{t}\mu_i u_i$, the KKT conditions are satisfied by $u_i^* = v_i + \frac{1-S_t}{t}$, $\lambda^* = \frac{1-S_t}{t}$, and $\mu_i^* = 0$ (primal feasibility: $u_i^* \geq 0$ and $\sum_i u_i^* = 1$; dual feasibility: $\mu_i^* \geq 0$; complementary slackness: $\mu_i^* u_i^* = 0$; stationarity: $2(u_i^* - v_i) + \lambda^* - \mu_i^* = 0$). Since the problem is convex with affine constraints, KKT optimality is sufficient. Extending to d coordinates by padding zeros yields $\mathbf{v}^{(t)} \in \Delta_t$ as the Euclidean projection of $\mathbf{v}$. □

Lemma B. For any $\mathbf{u},\mathbf{v} \in [0,1]^d$ with $\mathbf{1}^\top\mathbf{u} = \mathbf{1}^\top\mathbf{v} = 1$,

$$\big\|\mathbf{u}^{(t)} - \mathbf{v}^{(t)}\big\|_2 \;\leq\; \|\mathbf{u}-\mathbf{v}\|_2, \qquad \text{where } \mathbf{u}^{(t)},\mathbf{v}^{(t)} \text{ are as in Lemma 1.}$$

Proof of Lemma B. Euclidean projection onto a closed convex set is a nonexpansive map in ℓ_2, i.e., $\|\Pi_C(a) - \Pi_C(b)\|_2 \leq \|a-b\|_2$ for all a,b and any closed convex C. Applying this with $C = \Delta_t$ and $\Pi_C(\cdot) = (\cdot)^{(t)}$ gives the claim. □

To prove the theorem, we first analyze the eigenvalue concentration through truncation. By Theorem 4, with probability at least $1-\delta$,

$$\big\|\widehat{\boldsymbol{\lambda}}_{\mathcal{X},\mathcal{T}} - \widetilde{\boldsymbol{\lambda}}_{\mathcal{X},\mathcal{T}}\big\|_2 \leq \varepsilon'_\delta, \qquad \big\|\widehat{\boldsymbol{\lambda}}_{\mathcal{T}} - \widetilde{\boldsymbol{\lambda}}_{\mathcal{T}}\big\|_2 \leq \varepsilon'_\delta,$$

where $\varepsilon'_\delta = \frac{2}{\sqrt{n}}\big(1 + \sqrt{2\log\frac{4}{\delta}}\big)$. Applying Lemma B separately to the joint and prompt spectra,

$$\big\|\widehat{\boldsymbol{\lambda}}^{(t)}_{\mathcal{X},\mathcal{T}} - \widetilde{\boldsymbol{\lambda}}^{(t)}_{\mathcal{X},\mathcal{T}}\big\|_2 \leq \varepsilon'_\delta, \qquad \big\|\widehat{\boldsymbol{\lambda}}^{(t)}_{\mathcal{T}} - \widetilde{\boldsymbol{\lambda}}^{(t)}_{\mathcal{T}}\big\|_2 \leq \varepsilon'_\delta. \tag{15}$$

Then, we analyze entropy perturbation in the truncated size-t dimensions. Assume $n \geq 4e^2\big(1 + \sqrt{2\log\frac{4}{\delta}}\big)^2$ so that $\varepsilon'_\delta \leq 1/e$. Since the truncated (renormalized) vectors each have at most t nonzero entries and sum to 1, we can apply the coordinatewise log-difference bound (Lemma 2) followed by the ℓ_2–to–entropy control (Lemma 3) to obtain

$$\big|H(\widehat{\boldsymbol{\lambda}}^{(t)}_{\mathcal{X},\mathcal{T}}) - H(\widetilde{\boldsymbol{\lambda}}^{(t)}_{\mathcal{X},\mathcal{T}})\big| \leq \big\|\widehat{\boldsymbol{\lambda}}^{(t)}_{\mathcal{X},\mathcal{T}} - \widetilde{\boldsymbol{\lambda}}^{(t)}_{\mathcal{X},\mathcal{T}}\big\|_2 \sqrt{t}\,\log\Big(\frac{\sqrt{t}}{\big\|\widehat{\boldsymbol{\lambda}}^{(t)}_{\mathcal{X},\mathcal{T}} - \widetilde{\boldsymbol{\lambda}}^{(t)}_{\mathcal{X},\mathcal{T}}\big\|_2}\Big),$$

and the analogous bound with $\mathcal{T}$ in place of $(\mathcal{X},\mathcal{T})$. Using equation 15 and the monotonicity of $\log(\cdot)$,

$$\big|H(\widehat{\boldsymbol{\lambda}}^{(t)}_{\mathcal{X},\mathcal{T}}) - H(\widetilde{\boldsymbol{\lambda}}^{(t)}_{\mathcal{X},\mathcal{T}})\big| \leq \varepsilon'_\delta \sqrt{t}\,\log\Big(\frac{\sqrt{t}}{\varepsilon'_\delta}\Big), \qquad \big|H(\widehat{\boldsymbol{\lambda}}^{(t)}_{\mathcal{T}}) - H(\widetilde{\boldsymbol{\lambda}}^{(t)}_{\mathcal{T}})\big| \leq \varepsilon'_\delta \sqrt{t}\,\log\Big(\frac{\sqrt{t}}{\varepsilon'_\delta}\Big).$$

Next, note that by definition:

$$\log \text{Conditional-Vendi}^{(t)}(x_{1:n}|t_{1:n}) = H(\widehat{\boldsymbol{\lambda}}^{(t)}_{\mathcal{X},\mathcal{T}}) - H(\widehat{\boldsymbol{\lambda}}^{(t)}_{\mathcal{T}}),$$

and the same for the population quantity. The triangle inequality therefore gives

$$\Big|\log \text{Conditional-Vendi}^{(t)}(x_{1:n}|t_{1:n}) - \log \text{Conditional-Vendi}^{(t)}(P)\Big| \leq 2\,\varepsilon'_\delta \sqrt{t}\,\log\Big(\frac{\sqrt{t}}{\varepsilon'_\delta}\Big).$$

Finally, $\log\big(\frac{\sqrt{t}}{\varepsilon'_\delta}\big) \leq \log(nt)$ for $n,t \geq 2$, and substituting $\varepsilon'_\delta = \frac{2}{\sqrt{n}}\big(1 + \sqrt{2\log\frac{4}{\delta}}\big)$ yields the stated bound.

A.4.3 Proof of Theorem 3

We define $\varepsilon_\delta = \frac{2}{\sqrt{n}}(1 + \sqrt{2\log\frac{4}{\delta}})$. The Conditional-RKE score is defined as

$$\text{Conditional-RKE}(x_{1:n}|t_{1:n}) = \Big(\frac{\|\widehat{\boldsymbol{\lambda}}_{\mathcal{X},\mathcal{T}}\|_2}{\|\widehat{\boldsymbol{\lambda}}_{\mathcal{T}}\|_2}\Big)^{-2}.$$

We denote by R the ratio $\|\widehat{\boldsymbol{\lambda}}_{\mathcal{X},\mathcal{T}}\|_2/\|\widehat{\boldsymbol{\lambda}}_{\mathcal{T}}\|_2$ and by R_* its population counterpart $\|\widetilde{\boldsymbol{\lambda}}_{\mathcal{X},\mathcal{T}}\|_2/\|\widetilde{\boldsymbol{\lambda}}_{\mathcal{T}}\|_2$. We observe that the assumption on the lower-bound on sample size n can be rewritten to show that $\varepsilon_\delta \leq \min\{\frac{1}{2}m_{\mathcal{T}}, \frac{L}{2C_0}\}$.

To prove the theorem, we first bound the ratio deviation. To do this, we denote $A = \|\widehat{\boldsymbol{\lambda}}_{\mathcal{X},\mathcal{T}}\|_2$, $B = \|\widehat{\boldsymbol{\lambda}}_{\mathcal{T}}\|_2$, and their population counterparts $A_* = \|\widetilde{\boldsymbol{\lambda}}_{\mathcal{X},\mathcal{T}}\|_2$, $B_* = \|\widetilde{\boldsymbol{\lambda}}_{\mathcal{T}}\|_2$. From Theorem 4, we have $|A - A_*| \le \varepsilon_\delta$ and $|B - B_*| \le \varepsilon_\delta$. We can then write

$$|R - R_*| = \Big|\frac{AB_* - A_*B}{BB_*}\Big| \le \frac{|A - A_*|}{B} + \frac{A_*|B_* - B|}{BB_*}.$$

Using the fact that $B \ge m_{\mathcal{T}} - \varepsilon_\delta \ge \frac{1}{2}m_{\mathcal{T}}$ (which holds under our sample size condition) and $A_* \le M_{\mathcal{X},\mathcal{T}}$ by assumption, we obtain

$$|R - R_*| \le \frac{2\varepsilon_\delta}{m_{\mathcal{T}}} + \frac{2M_{\mathcal{X},\mathcal{T}}\varepsilon_\delta}{m_{\mathcal{T}}^2} = C_0\varepsilon_\delta.$$

Then, we analyze the deviation when we change the power. We note that $R_* \ge L$ by Assumption 2. Since we have shown $|R - R_*| \le C_0\varepsilon_\delta$ and our sample size condition ensures $C_0\varepsilon_\delta \le \frac{1}{2}L$, we conclude that $R \ge \frac{1}{2}L$. Applying Lemma 1 with $\beta = -2$, we obtain

$$|R^{-2} - R_*^{-2}| \le 2(\tfrac{1}{2}L)^{-3}|R - R_*| = \frac{16}{L^3}C_0\varepsilon_\delta.$$

This completes the proof.

B Theoretical Interpretation of the Conditional-Entropy Score

Theorem 5. *Consider the Gaussian kernel with bandwidth σ. Suppose T follows a mixture distribution $\sum_{i=1}^m \omega_i P_{T,i}$ with component means μ_i and within-component variances $\sigma_i^2 = \mathbb{E}_{T\sim P_{T,i}}[\|T - \mu_i\|_2^2]$. Define the error quantity term Γ as*

$$\Gamma \;=\; 32\sum_{i=1}^m \omega_i\left[\frac{\sigma_i^2}{\sigma^2} \;+\; (i-1)\sum_{j=1}^{i-1}\exp\Big(-\tfrac{\|\mu_i-\mu_j\|_2^2}{\sigma^2}\Big)\right]. \tag{16}$$

Then, the matrix-based order-2 conditional entropy satisfies the following for $g(z) = 2\log\big(\frac{1}{1-z/\|\omega\|_2}\big)$

$$\Big|\,\widetilde{H}_2(X|T) \;-\; \log\Big(1/\mathbb{E}_{I\sim\omega^2}\big[\exp(-\widetilde{H}_2(X|G=I))\big]\Big)\Big| \;\le\; 2\,g(\Gamma).$$

In the above, ω^2 represents the probability model $p_i = \omega_i^2/\big(\sum_{j=1}^m \omega_j^2\big)$, whose probability values are proportional to the square of the probability weight ω_i's. Also, note that the above is equivalent to what follows in terms of the Conditional-RKE score: $\text{Conditional-RKE}(X|T) = \exp(\widetilde{H}_2(X|T))$,

$$\exp\big(-2g(\Gamma)\big)\cdot\left(\mathbb{E}_{I\sim\omega^2}\Big[\tfrac{1}{\text{RKE}(X|G=I)}\Big]\right)^{-1} \le \text{Conditional-RKE}(X|T) \le \exp\big(2g(\Gamma)\big)\cdot\left(\mathbb{E}_{I\sim\omega^2}\Big[\tfrac{1}{\text{RKE}(X|G=I)}\Big]\right)^{-1}.$$

To prove the above theorem, we prove a more general result that applies to every matrix-based order-α Rényi conditional entropy of a unit-trace PSD matrix $M \in \mathbb{R}^{n\times n}$, defined as $H_\alpha(M) = \frac{1}{1-\alpha}\log\big(\sum_{i=1}^n \lambda_i^\alpha\big)$, on how it relates to the aggregation of text-instance entropy values for every $\alpha \ge 2$. Note that Theorem 5 is the direct corollary of the next theorem with $\alpha = 2$.

Theorem 6. *Consider the Gaussian kernel with bandwidth σ. Suppose T follows a mixture distribution $\sum_{i=1}^m \omega_i P_{T,i}$ where ω_i denotes the weight of the ith component $P_{T,i}$ with mean vector μ_i and total variance $\mathbb{E}_{T\sim P_{T,i}}[\|T - \mu_i\|_2^2] = \sigma_i^2$. Given the aggregation map $f(z) = \exp((1-\alpha)z)$, for every order $\alpha \ge 2$, the matrix-based order-α conditional entropy satisfies the following inequality with Γ defined in equation 16 where $g(z) = \frac{\alpha}{\alpha-1}\log\big(\frac{1}{1-z/\|\boldsymbol{\omega}\|_\alpha}\big)$ is an increasing scalar function with $g(0) = 0$:*

$$\left|\widetilde{H}_\alpha(X|T) - f^{-1}\Big(\mathbb{E}_{I\sim\boldsymbol{\omega}^\alpha}\Big[f\big(\widetilde{H}_\alpha(X|G=I)\big)\Big]\Big)\right| \;\le\; 2g\big(\Gamma\big)$$

Proof. To prove Theorem 6, we begin by showing the following lemma.

Lemma 5. *Suppose that the kernel function k and variable T satisfy the assumptions in Theorem 6. Then, the following Frobenius norm bound holds for $C_i = \mathbb{E}\big[\phi_X(x)\phi_X(x)^\top | G = i\big]$ where $G \in \{1, \ldots, m\}$ is the cluster random variable for text T:*

$$\Big\| C_{\mathcal{X},\mathcal{T}} - \sum_{i=1}^{m} \omega_i C_i \otimes \phi_T(\mu_i)\phi_T(\mu_i)^\top \Big\|_F^2 \le \frac{\sum_{i=1}^m 2\omega_i \sigma_i^2}{\sigma^2}.$$

Proof. To show this lemma, we define T_i as a variable distributed as $P_{T|G=i}$. Then,

$$\begin{aligned}
&\Big\| C_{\mathcal{X},\mathcal{T}} - \sum_{i=1}^{m} \omega_i C_i \otimes \phi(\mu_i)\phi(\mu_i)^\top \Big\|_F^2 \\
&= \Big\| \mathbb{E}\big[\phi_X(x)\phi_X(x)^\top \otimes \phi_T(t)\phi_T(t)^\top\big] - \sum_{i=1}^{m} \omega_i C_i \otimes \phi(\mu_i)\phi(\mu_i)^\top \Big\|_F^2 \\
&= \Big\| \sum_{i=1}^{m} \omega_i \mathbb{E}\big[\phi_X(x)\phi_X(x)^\top \otimes \phi_T(t)\phi_T(t)^\top | G = i\big] - \sum_{i=1}^{m} \omega_i C_i \otimes \phi(\mu_i)\phi(\mu_i)^\top \Big\|_F^2 \\
&= \Big\| \sum_{i=1}^{m} \omega_i \mathbb{E}\big[\phi_X(x)\phi_X(x)^\top \otimes \phi_T(t)\phi_T(t)^\top | G = i\big] - \sum_{i=1}^{m} \omega_i \mathbb{E}\big[\phi_X(x)\phi_X(x)^\top \otimes \phi_T(\mu_i)\phi_T(\mu_i)^\top | G = i\big] \Big\|_F^2 \\
&= \Big\| \sum_{i=1}^{m} \omega_i \mathbb{E}\big[\phi_X(x)\phi_X(x)^\top \otimes \Big(\phi_T(t)\phi_T(t)^\top - \phi_T(\mu_i)\phi_T(\mu_i)^\top\Big) | G = i\big] \Big\|_F^2 \\
&\overset{(a)}{\le} \sum_{i=1}^{m} \omega_i \mathbb{E}\Big[\Big\| \phi_X(x)\phi_X(x)^\top \otimes \Big(\phi_T(t)\phi_T(t)^\top - \phi_T(\mu_i)\phi_T(\mu_i)^\top\Big) \Big\|_F^2 | G = i\Big] \\
&\overset{(b)}{=} \sum_{i=1}^{m} \omega_i \mathbb{E}\Big[\Big\| \phi_X(x)\phi_X(x)^\top \Big\|_F^2 \Big\| \phi_T(t)\phi_T(t)^\top - \phi_T(\mu_i)\phi_T(\mu_i)^\top \Big\|_F^2 | G = i\Big] \\
&\overset{(c)}{=} \sum_{i=1}^{m} \omega_i \mathbb{E}\Big[\Big\| \phi_T(t)\phi_T(t)^\top - \phi_T(\mu_i)\phi_T(\mu_i)^\top \Big\|_F^2 | G = i\Big] \\
&\overset{(d)}{=} \sum_{i=1}^{m} \omega_i \mathbb{E}\Big[2 - 2\exp\Big(\frac{-\|t-\mu_i\|_2^2}{\sigma^2}\Big) | G = i\Big] \\
&\overset{(e)}{\le} \sum_{i=1}^{m} \omega_i \Big[2 - 2\exp\Big(\frac{-\mathbb{E}\big[\|t-\mu_i\|_2^2 | G = i\big]}{\sigma^2}\Big)\Big] \\
&\overset{(f)}{\le} \sum_{i=1}^{m} \omega_i \Big[2 - 2\exp\Big(\frac{-\sigma_i^2}{\sigma^2}\Big)\Big] \\
&\overset{(g)}{\le} \sum_{i=1}^{m} 2\omega_i \frac{\sigma_i^2}{\sigma^2}
\end{aligned}$$

In the above, (a) follows from Jensen's inequality for the convex Frobenius-norm-squared function. (b) holds because $\|A \otimes B\|_F^2 = \|A\|_F^2 \|B\|_F^2$ for every matrices A, B. (c) comes from the normalized Gaussian kernel satisfying $\langle \phi_T(t), \phi_T(t) \rangle = k(t,t) = 1$, resulting in $\|\phi_T(t)\phi_T(t)^\top\|_F^2 = \mathrm{Tr}\big(\phi_T(t)\phi_T(t)^\top \phi_T(t)\phi_T(t)^\top\big) = \mathrm{Tr}\big(\phi_T(t)\phi_T(t)^\top\big) = 1$. (d) follows from the Gaussian kernel definition, proving that $\phi_T(t)^\top \phi_T(\mu_i) = \exp\big(-\|t-\mu_i\|_2^2/2\sigma^2\big)$. (e) shows the application of Jensen's inequality to the concave $s(z) = 1 - \exp(-z)$. (f) holds because $s(z) = 1 - \exp(-z)$ is a monotonically increasing function. Finally, (g) follows from the inequality $1 - \exp(-z) \le z$ for every scalar z. Therefore, the proof is complete.

□

Next, we apply the Gram–Schmidt process to $\phi_T(\mu_1), \ldots, \phi_T(\mu_m)$ to find orthogonal vectors $u_1, \ldots, u_m$. We let

$u_1 = \phi_T(\mu_1)$. Then, for every $2 \le i \le m$, we define

$$v_i := \phi(\mu_i) - \sum_{j=1}^{i-1} \langle \phi(\mu_i), u_j \rangle u_j, \quad u_i = v_i / \|v_i\|_2$$

As a result, the following holds

$$\begin{aligned}
&\Big\| \sum_{i=1}^m \omega_i C_i \otimes \phi(\mu_i)\phi(\mu_i)^\top - \sum_{i=1}^m \omega_i C_i \otimes u_i u_i^\top \Big\|_F^2 \\
&= \Big\| \sum_{i=1}^m \omega_i C_i \otimes \Big(\phi(\mu_i)\phi(\mu_i)^\top - u_i u_i^\top \Big) \Big\|_F^2 \\
&\overset{(h)}{\le} \sum_{i=1}^m \omega_i \Big\| C_i \otimes \Big(\phi(\mu_i)\phi(\mu_i)^\top - u_i u_i^\top \Big) \Big\|_F^2 \\
&= \sum_{i=1}^m \omega_i \Big\| C_i \Big\|_F^2 \Big\| \phi(\mu_i)\phi(\mu_i)^\top - u_i u_i^\top \Big\|_F^2 \\
&\overset{(i)}{\le} \sum_{i=1}^m \omega_i \Big\| \phi(\mu_i)\phi(\mu_i)^\top - u_i u_i^\top \Big\|_F^2 \\
&\overset{(j)}{=} \sum_{i=1}^m \omega_i \Big(2 - 2\big(u_i^\top \phi_T(\mu_i) \big)^2 \Big) \\
&\overset{(k)}{=} 2 \sum_{i=1}^m \omega_i \sum_{j=1}^{i-1} \big(u_j^\top \phi_T(\mu_i) \big)^2
\end{aligned}$$

Here, (h) follows from the application of Jensen's inequality for the convex Frobenius-norm-squared. (i) holds since the text kernel is normalized and $\langle \phi_X(x), \phi_X(x) \rangle = k_X(x,x) = 1$, and therefore $\|C_i\|_F \le \mathbb{E}[\|\phi_X(x)\|_2^2] = 1$. (j) follows from the expansion $\|uu^\top - vv^\top\|_F^2 = \|u\|_2^4 + \|v\|_2^4 - 2\langle u, v \rangle^2$. Next, we bound the inner products $\langle u_j, \phi_T(\mu_i) \rangle$ that appear in step (k). Recall that each u_j is obtained from the Gram–Schmidt process applied to $\{\phi_T(\mu_\ell)\}_{\ell < i}$, so we can write

$$u_j \;=\; \sum_{\ell \le j} r_{j\ell}\, \phi_T(\mu_\ell), \qquad \|r_j\|_2 \le 1,$$

where $r_j = (r_{j1}, \ldots, r_{jj})^\top$ is the coefficient vector. Therefore,

$$\big(u_j^\top \phi_T(\mu_i) \big)^2 = \Big(\sum_{\ell \le j} r_{j\ell}\, k_T(\mu_\ell, \mu_i) \Big)^2 \;\le\; \sum_{\ell \le j} k_T(\mu_\ell, \mu_i)^2,$$

where we used $\|r_j\|_2 \le 1$ and Cauchy–Schwarz. Summing over $j < i$ yields

$$\sum_{j<i} \big(u_j^\top \phi_T(\mu_i) \big)^2 \;\le\; \sum_{j<i} \sum_{\ell \le j} k_T(\mu_\ell, \mu_i)^2 \;=\; \sum_{\ell < i} (i - \ell)\, k_T(\mu_\ell, \mu_i)^2 \;\le\; (i-1) \sum_{\ell < i} k_T(\mu_\ell, \mu_i)^2.$$

For the Gaussian kernel $k_T(t, t') = \exp(-\|t - t'\|^2/(2\sigma^2))$, we have $k_T(\mu_\ell, \mu_i)^2 = \exp(-\|\mu_\ell - \mu_i\|^2/\sigma^2)$. Hence, inequality (k) becomes

$$\sum_{i=1}^m \omega_i \Big\| \phi_T(\mu_i)\phi_T(\mu_i)^\top - u_i u_i^\top \Big\|_F^2 \le 2 \sum_{i=1}^m \omega_i (i-1) \sum_{\ell < i} \exp\Big(-\tfrac{\|\mu_i - \mu_\ell\|^2}{\sigma^2} \Big). \tag{17}$$

Combining equation 17 with the variance bound from Lemma 1, and applying $\|A + B\|_F^2 \le 2\|A\|_F^2 + 2\|B\|_F^2$, we obtain

$$\Big\| C_{\mathcal{X},\mathcal{T}} - \sum_{i=1}^m \omega_i C_i \otimes u_i u_i^\top \Big\|_F^2 \le 4 \sum_{i=1}^m \omega_i \frac{\sigma_i^2}{\sigma^2} + 8 \sum_{i=1}^m \omega_i (i-1) \sum_{\ell < i} \exp\Big(-\tfrac{\|\mu_i - \mu_\ell\|^2}{\sigma^2} \Big). \tag{18}$$

Since $u_1, \ldots, u_m$ are orthogonal vectors, the definition of Kronecker product implies that the eigenvalues of $\sum_{i=1}^m \omega_i C_i \otimes u_i u_i^\top$ will be the union of the eigenvalues of $\omega_i C_i \otimes u_i u_i^\top$ over $i \in \{1, \ldots, m\}$. On the other hand, we know that the non-zero eigenvalues of $\omega_i C_i \otimes u_i u_i^\top$ will be equal to the factor $\omega_i \|u_i\|_2^2 = \omega_i$ times the eigenvalues of C_i. Consequently, we can show that for vector $\widehat{\lambda}_{x\otimes t} = \text{Union}\big(\omega_i \,\text{Eigs}(C_i) : i \in \{1, \ldots, m\}\big)$, we have the following for every $\alpha \geq 2$ and defined increasing function g in Theorem 6

$$\begin{aligned}
\Big|\widetilde{H}_\alpha(X,T) - \frac{1}{1-\alpha}\log\big(\|\widehat{\lambda}_{x\otimes t}\|_\alpha^\alpha\big)\Big| &\leq g\big(\|\widetilde{\lambda}_{x\otimes t}\|_\alpha - \|\widehat{\lambda}_{x\otimes t}\|_\alpha\big) \\
&\leq g\big(\|\text{sort}\big(\widetilde{\lambda}_{x\otimes t}\big) - \text{sort}\big(\widehat{\lambda}_{x\otimes t}\big)\|_\alpha\big) \\
&\leq g\big(\|\text{sort}\big(\widetilde{\lambda}_{x\otimes t}\big) - \text{sort}\big(\widehat{\lambda}_{x\otimes t}\big)\|_2\big) \\
&\leq g\Big(\sum_{i=1}^m 4\omega_i \frac{\sigma_i^2}{\sigma^2} + \sum_{i=2}^m \sum_{j=1}^{i-1} 8\omega_i(i-1)\exp\Big(\frac{-\|\mu_i - \mu_j\|_2^2}{\sigma^2}\Big)\Big).
\end{aligned}$$

Note that the above proof holds for every marginal distribution on X, and we choose a deterministic constant $X = \mathbf{0}$, then the joint entropy reduces to the marginal entropy and the above inequality also shows the following:

$$\Big|\widetilde{H}_\alpha(T) - \frac{1}{1-\alpha}\log\big(\|[\omega_1, \ldots, \omega_m]\|_\alpha^\alpha\big)\Big| \leq g\Big(\sum_{i=1}^m 4\omega_i \frac{\sigma_i^2}{\sigma^2} + \sum_{i=2}^m \sum_{j=1}^{i-1} 8(i-1)\omega_i \exp\Big(\frac{-\|\mu_i - \mu_j\|_2^2}{\sigma^2}\Big)\Big).$$

Therefore, following the Triangle inequality and the definition $\widetilde{H}_\alpha(X|T) = \widetilde{H}_\alpha(X,T) - \widetilde{H}_\alpha(T)$, the previous two inequalities prove that

$$\begin{aligned}
&\Big|\widetilde{H}_\alpha(X|T) - \Big(\frac{1}{1-\alpha}\log\big(\|\widehat{\lambda}_{x\otimes t}\|_\alpha^\alpha\big) - \frac{1}{1-\alpha}\log\big(\|[\omega_1, \ldots, \omega_m]\|_\alpha^\alpha\big)\Big)\Big| \\
\leq\, & 2g\Big(\sum_{i=1}^m 4\omega_i \frac{\sigma_i^2}{\sigma^2} + \sum_{i=2}^m \sum_{j=1}^{i-1} 8(i-1)\omega_i \exp\Big(\frac{-\|\mu_i - \mu_j\|_2^2}{\sigma^2}\Big)\Big) \;=\; 2g(\Gamma)
\end{aligned}$$

On the other hand, we can simplify the above expression as

$$\begin{aligned}
&\frac{1}{1-\alpha}\log\big(\|\widehat{\lambda}_{x\otimes t}\|_\alpha^\alpha\big) - \frac{1}{1-\alpha}\log\big(\|[\omega_1, \ldots, \omega_m]\|_\alpha^\alpha\big) \\
=\, &\frac{1}{1-\alpha}\log\big(\sum_{i=1}^m \omega_i^\alpha \|\lambda_{C_i}\|_\alpha^\alpha\big) - \frac{1}{1-\alpha}\log\big(\sum_{i=1}^m \omega_i^\alpha\big) \\
=\, &\frac{1}{1-\alpha}\log\big(\sum_{i=1}^m \frac{\omega_i^\alpha}{\sum_{j=1}^m \omega_j^\alpha}\|\lambda_{C_i}\|_\alpha^\alpha\big)
\end{aligned}$$

Note that the definition $f_\alpha(t) = \exp((1-\alpha)t)$ implies that $f_\alpha^{-1}(z) = \frac{1}{1-\alpha}\log(z)$, which connects to the entropy definition as $H(X|G=i) = f_\alpha^{-1}(\|\lambda_{C_i}\|_\alpha^\alpha)$. Hence, we combine the previous two equations to complete the proof:

$$\Big|\widetilde{H}_\alpha(X|T) - f_\alpha^{-1}\Big(\sum_{i=1}^m \frac{\omega_i^\alpha}{\sum_{j=1}^m \omega_j^\alpha} f_\alpha\big(\widetilde{H}_\alpha(X|G=i)\big)\Big)\Big| \;\leq\; 2g(\Gamma)$$

□

Theorem 7 (Truncated Conditional-Vendi Interpretation)**.** *Let $T \sim \sum_{i=1}^m \omega_i P_{T,i}$ with means μ_i and within-component variances $\sigma_i^2 = \mathbb{E}\big[\|T - \mu_i\|_2^2 | G = i\big]$. Assume Gaussian kernel bandwidth σ and normalized feature maps. Fix a truncation level $t \in \mathbb{N}$ and consider Γ defined in equation 16. Let $\widetilde{\boldsymbol{\lambda}}_{\mathcal{X},\mathcal{T}}^{(t)}$ and $\widetilde{\boldsymbol{\lambda}}_{\mathcal{T}}^{(t)}$ be the top-t truncated, renormalized eigenvalue vectors of the joint and prompt operators, and define*

$$H^{(t)}(X|T) := H\big(\widetilde{\boldsymbol{\lambda}}_{\mathcal{X},\mathcal{T}}^{(t)}\big) - H\big(\widetilde{\boldsymbol{\lambda}}_{\mathcal{T}}^{(t)}\big).$$

Then, we have the following:

$$\Big|\, H^{(t)}(X|T) \;-\; \sum_{i=1}^m \omega_i\, H^{(t)}(X|G=i)\Big| \;\leq\; \sqrt{t\Gamma}\,\log\big(\tfrac{t}{\Gamma}\big). \tag{19}$$

Equivalently, for the truncated Conditional-Vendi score $\text{Vendi}^{(t)} = \exp(H^{(t)})$,

$$\begin{aligned}\big(\frac{\Gamma}{t}\big)^{2\sqrt{t\Gamma}} \prod_{i=1}^{m} \Big(\text{Vendi}^{(t)}(X|G=i)\Big)^{\omega_i} \;&\le\; \text{Conditional-Vendi}^{(t)}(X|T) \\ &\le\; \big(\frac{t}{\Gamma}\big)^{2\sqrt{t\Gamma}} \prod_{i=1}^{m} \Big(\text{Vendi}^{(t)}(X|G=i)\Big)^{\omega_i}. \end{aligned} \tag{20}$$

Proof. By the Frobenius bound in equation 18 and the Hoffman–Wielandt inequality,

$$\big\|\boldsymbol{\lambda}(C_{\mathcal{X},\mathcal{T}}) - \boldsymbol{\lambda}(\textstyle\sum_i \omega_i C_i \otimes u_i u_i^\top)\big\|_2^2 \;\le\; \Gamma.$$

Since $\{u_i\}$ are orthonormal, the nonzero eigenvalues of $\sum_i \omega_i C_i \otimes u_i u_i^\top$ are the union of the eigenvalues of $\omega_i C_i$, so $\boldsymbol{\lambda}(\sum_i \omega_i C_i \otimes u_i u_i^\top)$ is obtained by stacking $\omega_i \,\text{Eigs}(C_i)$ over i. The same argument with X fixed (or $C_X = 0$) yields the prompt-side spectral approximation. Projecting both joint and prompt spectra onto the t-truncated, renormalized simplex Δ_t (Lemma A) and using nonexpansiveness of Euclidean projection (Lemma B) gives

$$\big\|\widetilde{\boldsymbol{\lambda}}^{(t)}_{\mathcal{X},\mathcal{T}} - \widehat{\boldsymbol{\lambda}}^{(t)}_{\mathcal{X},\mathcal{T}}\big\|_2 \;\le\; \sqrt{\Gamma}, \qquad \big\|\widetilde{\boldsymbol{\lambda}}^{(t)}_{\mathcal{T}} - \widehat{\boldsymbol{\lambda}}^{(t)}_{\mathcal{T}}\big\|_2 \;\le\; \sqrt{\Gamma},$$

where $\widehat{\boldsymbol{\lambda}}^{(t)}_{\mathcal{X},\mathcal{T}}$ (resp. $\widehat{\boldsymbol{\lambda}}^{(t)}_{\mathcal{T}}$) denotes the t-truncated, renormalized vector formed from the stacked union of $\omega_i \,\text{Eigs}(C_i)$ (resp. of the mixture weights (ω_i)). On the truncated simplex space, we apply Lemma 2 coordinatewise and then Lemma 3 to obtain,

$$\big|H(\widetilde{\boldsymbol{\lambda}}^{(t)}_{\mathcal{X},\mathcal{T}}) - H(\widehat{\boldsymbol{\lambda}}^{(t)}_{\mathcal{X},\mathcal{T}})\big| \;\le\; \sqrt{t\Gamma}\,\log\Big(\sqrt{\frac{t}{\Gamma}}\Big), \qquad \big|H(\widetilde{\boldsymbol{\lambda}}^{(t)}_{\mathcal{T}}) - H(\widehat{\boldsymbol{\lambda}}^{(t)}_{\mathcal{T}})\big| \;\le\; \sqrt{t\Gamma}\,\log\Big(\sqrt{\frac{t}{\Gamma}}\Big).$$

Subtracting the two displays and using the triangle inequality yields

$$\Big|\, H^{(t)}(X|T) - \Big(H(\widehat{\boldsymbol{\lambda}}^{(t)}_{\mathcal{X},\mathcal{T}}) - H(\widehat{\boldsymbol{\lambda}}^{(t)}_{\mathcal{T}})\Big)\Big| \;\le\; 2\,\sqrt{t\Gamma}\,\log\Big(\sqrt{\tfrac{t}{\Gamma}}\Big).$$

Finally, by construction of the stacked union and the truncation on each block, $H(\widehat{\boldsymbol{\lambda}}^{(t)}_{\mathcal{X},\mathcal{T}}) - H(\widehat{\boldsymbol{\lambda}}^{(t)}_{\mathcal{T}}) = \sum_{i=1}^m \omega_i\, H_i^{(t)}$, which proves equation 19. Exponentiating both sides results in equation 20 since $\text{Vendi}^{(t)} = \exp(H^{(t)})$ and $\exp(\sum_i \omega_i H_i^{(t)}) = \prod_i (\text{Vendi}^{(t)}(X|G=i))^{\omega_i}$. □

C Details of Truncated Conditional-Vendi Guidance in Section 6

In the guidance setting, let $K_Z = [k_{\mathcal{Z}}(\boldsymbol{z}^{(i)}, \boldsymbol{z}^{(j)})]_{i,j=1}^n$ and $K_T = [k_{\mathcal{T}}(t_i, t_j)]_{i,j=1}^n$ be normalized latent variable and prompt kernel matrices with unit diagonal entries. Define the joint unit-trace PSD matrix

$$A \;=\; \tfrac{1}{n}\,(K_Z \odot K_T), \qquad \text{Tr}(A) = 1. \tag{21}$$

We use the truncated Conditional-Vendi score introduced in the main text:

$$\text{Conditional-Vendi}^{(t)}(x_{1:n} \mid t_{1:n}) \;=\; \exp\Big(H^{(t)}(A) - H^{(t)}\big(\tfrac{1}{n}K_T\big)\Big). \tag{22}$$

Given a unit-trace PSD M with eigenvalues $\lambda_1 \ge \cdots \ge \lambda_n \ge 0$ and $S_t = \sum_{i=1}^t \lambda_i$ providing $c(\boldsymbol{\lambda}) = \frac{1-S_t}{t}$, we recall the following definition of the t-truncated entropy of the matrix M:

$$H^{(t)}(M) \;=\; -\sum_{i=1}^{t} (\lambda_i + c(\boldsymbol{\lambda}))\,\log(\lambda_i + c(\boldsymbol{\lambda})). \tag{23}$$

Here we describe how to compute the gradient of $H^{(t)}(M)$ in the guidance process. Consider the eigendecomposition of symmteric matrix M as $M = V\Lambda V^\top$ where $V_t = [v_1, \ldots, v_t]$, $\Lambda_t = \text{diag}(\lambda_1, \ldots, \lambda_t)$ are the matrix of top-t eigenvectors and vector of top t eigenvalues. Using the notation $\bar{\ell}_t = \frac{1}{t}\sum_{i=1}^t \log(\lambda_i + c(\boldsymbol{\lambda}))$, the gradient is

$$\nabla_M H^{(t)}(M) \;=\; -\,V_t \log(\Lambda_t + cI_t)\, V_t^\top \;+\; \bar{\ell}_t\, V_t V_t^\top. \tag{24}$$

Since $H^{(t)}\big(\frac{1}{n}K_T\big)$ does not depend on variable Z, we will obtain

$$\nabla_{\boldsymbol{z}^{(n)}}\text{Conditional-Vendi}^{(t)}(z_{1:n};t_{1:n}) \;=\; \text{Conditional-Vendi}^{(t)}(z_{1:n};t_{1:n})\cdot\nabla_{\boldsymbol{z}^{(n)}}H^{(t)}(A). \tag{25}$$

Using equation 21, the Hadamard chain rule gives the following where $\nabla_A H^{(t)}(A)$ is given by equation 24:

$$\nabla_{K_Z}H^{(t)}(A) \;=\; \tfrac{1}{n}\left(\nabla_A H^{(t)}(A)\right)\odot K_T. \tag{26}$$

Note that only the n-th row and column of K_Z depends on $\boldsymbol{z}^{(n)}$, and thus we have

$$\nabla_{\boldsymbol{z}^{(n)}}H^{(t)}(A) \;=\; \sum_{i=1}^{n-1}\Big((\nabla_{K_Z}H^{(t)}(A))_{in}+(\nabla_{K_Z}H^{(t)}(A))_{ni}\Big)\,\nabla_{\boldsymbol{z}^{(n)}}k_{\mathcal{Z}}(\boldsymbol{z}^{(i)},\boldsymbol{z}^{(n)}). \tag{27}$$

Combining equation 25 and equation 27 results in the guidance direction as follows:

$$\boldsymbol{z}^{(n)}_{\tau-1}\leftarrow\text{Sampler}\big(\boldsymbol{z}^{(n)}_{\tau},\hat{\epsilon}_\theta(\boldsymbol{z}^{(n)}_{\tau},\tau,t_n)\big)+\eta\,\nabla_{\boldsymbol{z}^{(n)}}\text{Conditional-Vendi}^{(t)}(z_{1:n};t_{1:n}).$$

Extension to Conditional-RKE Guidance. Given the unit value diagonal entries of the kernel matrix, the Conditional-RKE score takes the following form:

$$\text{Conditional-RKE}(x_{1:n}\mid t_{1:n})=\frac{\|K_T\|_F^2}{\|K_Z\odot K_T\|_F^2}.$$

Let $S=K_Z\odot K_T$. Taking the derivative with respect to S gives the following

$$\nabla_S\,\text{Conditional-RKE}(z_{1:n};t_{1:n})=-\,2\,\|K_T\|_F^2\,\frac{S}{\|S\|_F^4},\;\nabla_{K_Z}\,\text{Conditional-RKE}(z_{1:n};t_{1:n})=\nabla_S\,\text{Conditional-RKE}\odot K_T,$$

and the latent gradient follows via the same accumulation as equation 27. This variant avoids a full spectral decomposition and can be implemented using $O(n^2)$ computations per iteration.

D Additional Related Works

Evaluation of deep generative models. Metrics for evaluating generative models are generally divided into reference-dependent and reference-free categories (Borji, 2022). Reference-dependent metrics compare generated and real data distributions, with common examples including FID (Heusel et al., 2017) and KID (Bińkowski et al., 2018). Other reference-based measures, such as the Inception Score (Salimans et al., 2016), Precision/Recall (Sajjadi et al., 2018; Kynkäänniemi et al., 2019), and Density/Coverage (Naeem et al., 2020), jointly evaluate fidelity and diversity with respect to a reference dataset.

Beyond fidelity, several works examine memorization and novelty. These include the authenticity score (Alaa et al., 2022) and Feature Likelihood Divergence (Jiralerspong et al., 2023) for assessing generalization, as well as the rarity score (Han et al., 2023) and KEN (Zhang et al., 2024) for quantifying novelty. The memorization metrics are reference-based. In contrast, reference-free evaluations assess quality and diversity directly from the generated data. Notable examples include the Vendi score (Friedman and Dieng, 2023; Pasarkar and Dieng, 2024) and RKE score (Jalali et al., 2023) for diversity, and (Nguyen and Dieng, 2024) for evaluating the quality of generated data.

Evaluation of conditional generative models. The evaluation of prompt-based generative models, such as text-to-image and text-to-video systems, has been explored in several recent works. Most metrics focus on measuring alignment between prompts and outputs. A widely used example is CLIPScore (Hessel et al., 2021), which computes cosine similarity in the CLIP embedding space. Other efforts have introduced benchmarks and curated prompt sets to evaluate broader aspects. For instance, HEIM (Lee et al., 2023) assesses twelve criteria, including text–image alignment, image quality, and bias.

However, alignment- and quality-focused metrics may overlook output diversity. Astolfi et al. (2024) emphasize that metrics centered on style or aesthetics can fail to capture variability across outputs for the same prompt. They propose computing per-prompt diversity using similarity functions and then averaging across prompts. Similarly, Kannen et al. (2024) extend the Vendi score to the per-prompt setting. Both approaches require

generating multiple outputs for each prompt with different seeds. In contrast, our proposed Conditional-Vendi does not require repeated generations; instead, it quantifies model-induced diversity by analyzing variability across prompt types. Our theoretical results interpret Conditional-Vendi as an aggregation of diversity scores across prompt categories.

Information measures for evaluating conditional generative models: Kim et al. (2022b) propose the Mutual Information Divergence (MID) score, which fits multivariate Gaussian distributions to text and image representations and estimates their mutual information to quantify relevance in conditional generative models. In contrast, our score builds on the standard PSD matrix-based entropy measures applied to kernel matrices. Unlike MID, which relies on mutual information between Gaussian-fitted embeddings, the proposed diversity operates on kernel similarity values.

Conditional Generation with Guidance. Controlling generative processes using specific conditions is increasingly important for practical applications, relying on inputs such as text prompts (Kim et al., 2022a; Nichol et al., 2022; Liu et al., 2023), class labels (Dhariwal and Nichol, 2021), style images (Mou et al., 2024; Zhang et al., 2023b), or human motions (Tevet et al., 2023), among others. Approaches to conditional generation with guidance can be divided into training-based and training-free methods. Training-based strategies either learn a time-dependent classifier to steer the noisy sample $\boldsymbol{x}_t$ toward the target condition $\boldsymbol{y}$ (Dhariwal and Nichol, 2021; Nichol et al., 2022; Zhao et al., 2022; Liu et al., 2023), or directly train the conditional denoising model $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, \boldsymbol{y})$ through few-shot adaptation (Mou et al., 2024; Ruiz et al., 2023). In contrast, training-free guidance enables zero-shot conditional generation by using a pre-trained differentiable predictor, such as a classifier, loss function, or energy function, which measures how well a generated sample aligns with the target condition (He et al., 2024; Bansal et al., 2023; Yu et al., 2023; Ye et al., 2024). Our Conditional-Vendi and Conditional-RKE guidance methods fall into this category, leveraging conditional entropy score guidance to improve the diversity of generated samples.

Guidance for Improving Diversity.

A common approach in diffusion-based generative models involves using guidance mechanisms to manage the trade-off between sample quality and diversity (Sadat et al., 2025; Ho and Salimans, 2022b). For instance, classifier-free guidance (Ho and Salimans, 2022b) significantly improves alignment with prompts and overall image quality, but can reduce diversity due to its strongly deterministic conditioning. Recent studies have sought to mitigate this diversity issue. (Sehwag et al., 2022) proposed a method that promotes diversity by explicitly sampling from low-density regions of the data manifold, though their approach works directly in pixel space, making it challenging to adapt to latent diffusion models. Another approach involves fine-tuning: (Miao et al., 2024) present a reinforcement learning-based finetuning strategy that enhances diversity by optimizing an image-set-based diversity reward function.

(Askari Hemmat et al., 2024) proposes contextualized Vendi Score Guidance (c-VSG) to boost generative diversity via the Vendi Score (Friedman and Dieng, 2023; Ospanov et al., 2024), but it relies on identical prompts, limiting its generality. In contrast, our approach uses Conditional-Vendi and Conditional-RKE Scores for prompt-aware guidance, enabling adaptive soft-clustering and effective conditioning on diverse prompts. Unlike c-VSG, which operates on latent features of reference images, our method directly guides the diffusion model's latent space, reducing computational cost while enhancing prompt-aware diversity.

E Implementation Details and Hyperparameters

E.1 Conditional-Vendi and Conditional-RKE Evaluation Hyperparameters

To select the bandwidth parameter σ, Similar to (Jalali et al., 2023; Ospanov et al., 2024), we chose the Gaussian kernel bandwidth for each type of data as the smallest σ that ensures a variance below 0.01 in the evaluated score over independent evaluations. We observed that for image data, $\sigma \in [20, 30]$; for text data, $\sigma \in [0.1, 0.8]$; and for video data, $\sigma \in [10, 20]$ can satisfy this requirement. Note that by selecting an overly large σ value for text embeddings, Conditional-Vendi simplifies to the expectation of unconditional Vendi per prompt. For Truncated Conditional-Vendi and Information-Vendi, we set the truncation parameter t to $10,000$ as suggested in (Ospanov and Farnia, 2025) in all the experiments.

E.2 Conditional-Vendi and RKE Guidance Details

In the kernel-based guidance experiments of Conditional-Vendi and Conditional-RKE, 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 (Friedman and Dieng, 2023; Jalali et al., 2023). The numerical experiments were conducted on 4×NVIDIA GeForce RTX 4090 GPUs, each of which has 22.5 GB of memory.

E.3 Experimental Configuration for Table 1

We used Stable Diffusion XL with a resolution of 1024×1024, a fixed classifier-free guidance scale of $W_{\text{CFG}} = 7.5$, and 50 inference steps using the DPM solver. We used 10,000 randomly selected prompts of the MS-COCO 2014 validation set and fixed the generation seed to be able to compare the effect of the methods. We used the following configuration to generate the results reported in Table 1.

The hyperparameter tuning was performed by performing cross-validation on the in-batch similarity score, selecting the hyperparameter values that optimized this alignment-based metric. Note that the in-batch similarity score accounts for both text-image consistency and inter-sample diversity as discussed in (Corso et al., 2024).

c-VSG. We note that the reference (Askari Hemmat et al., 2024) considered GeoDE (Ramaswamy et al., 2023) and DollarStreet (Gaviria Rojas et al., 2022) datasets, in which multiple samples exist per input prompt. On the other hand, in our experiments, we considered the standard MSCOCO prompt set where for each prompt corresponds we access a single image, making the contextualized Vendi guidance baseline in (Askari Hemmat et al., 2024) not directly applicable. Therefore, we simulated the non-contextualized version of VSG. For selecting the Vendi score guidance scale, we performed validation over the set $\{0, 0.04, 0.05, 0.06, 0.07\}$, following the procedure in (Askari Hemmat et al., 2024). A guidance frequency of 5 was used, consistent with the original implementation. To maintain stable gradient computation for the Vendi score, we implemented a sliding window of 300 most recently generated samples, as gradient calculations became numerically unstable for some steps beyond this threshold.

latent Vendi Guidance. We used a Gaussian kernel with bandwidth $\sigma_{img} = 0.8$ and used $\eta = 0.03$ as the weight of RKE 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).

latent RKE Guidance. We used a Gaussian kernel with bandwidth $\sigma_{img} = 0.8$ and used $\eta = 0.03$ as the weight of RKE guidance. To balance the effects of the diversity guidance in sample generation, the RKE 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).

latent Conditional-Vendi Guidance. We used a Gaussian kernel with bandwidth $\sigma_{img} = 0.8$ and used $\eta = 0.03$ as the weight of Conditional-Vendi guidance. To balance the effects of the diversity guidance in sample generation, the RKE 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).

latent Conditional-RKE Guidance. We considered the same Gaussian kernel for the image generation with bandwidth $\sigma_{img} = 0.8$ and used bandwidth parameter $\sigma_{text} = 0.3$ for the text kernel. The guidance hyperparameter was set to $\eta = 0.03$, as in RKE guidance. Similar to the RKE and Vendi guidance, the Conditional-RKE diversity guidance was applied every 10 reverse-diffusion steps. 3

F Additional Numerical Results on Conditional-Vendi and Conditional-RKE Guidance

To demonstrate the advantages of prompt-aware metrics over unconditional Vendi score, we explored their potential to enhance sample diversity in PixArt. We guided the model using both Truncated-Conditional-Vendi and standard Vendi score, following the methodology from (Askari Hemmat et al., 2024).

In our implementation, we applied guidance to PixArt in the latent space rather than the ambient space. This approach substantially reduced memory requirements from over 50 GB to approximately 20 GB while maintaining

performance. We observed that latent-space guidance not only improves image diversity and quality but also offers significant computational efficiency gains. A comprehensive comparison between latent and ambient-space guidance is included in the Appendix.

Figure 10 presents qualitative results using PixArt, demonstrating how prompt-aware guidance generates more relevant and contextually diverse images. Quantitative comparisons between Vendi and Conditional-Vendi guidance methods on PixArt are provided in Table 3, showing that Conditional-Vendi guidance enhances sample diversity (as measured by Vendi score and in-batch similarity) while preserving text-image alignment through competitive CLIPScore and KD metrics.

Table 3: Quantitative comparison of guidance methods on PixArt

Guidance Method	**CLIPScore** ↑	**KD**$\times 10^2$ ↓	**Cond-Vendi$_{\text{DINOv2}}$** ↑	**Vendi$_{\text{DINOv2}}$** ↑	**In-batch Sim.**$\times 10^2$ ↓
Vendi$_{\text{CLIP}}$	29.63	38.14	26.28	261.73	83.26
Vendi$_{\text{Latent}}$	30.39	36.20	28.95	298.15	81.50
Conditional-Vendi$_{\text{Latent}}$	30.44	29.80	31.50	312.80	79.45

G Additional Numerical Results on the Evaluation of Generative Models

G.1 Ablation Studies

Toy example on Gaussian Mixture Models. To validate that Conditional-Vendi and Information-Vendi accurately quantify model-induced diversity and prompt correlation, we evaluate these metrics on multiple Gaussian Mixture datasets. As illustrated in Figure 11, we generated separate 2D Gaussian distributions to represent text and image modalities, which we then paired through minimum weight bipartite graph matching. In the first row, we fix the number of image Gaussian distributions (X) while increasing the number of text modes from 1 (less correlated) to 4 (highly correlated). As shown in the figure, Information-Vendi increases from 2.97 to 4.61, whereas Conditional-Vendi decreases from 2.21 to 1.34. These results indicate that conditional sample diversity is higher when paired with a single text mode compared to scenarios where images are fully aligned with prompts. The correlation between text and images is maximized when there is one cluster of images for each group of prompts. In the second row, we used two Gaussian distributions for text (T) while varying the number of image modes (X) from 2 to 6. The results show that Conditional-Vendi score increases from 1.55 to 2.46, while Information-Vendi decreases from 4.01 to 3.42. This suggests that when the model generates more modes for a group of prompts, it produces greater model-induced diversity.

Correlation between prompts and generated output. To measure the correlation between text and image using Information-Vendi, we used MS-COCO captions to generate images with Stable Diffusion XL and Flux. We gradually substituted the generated images with random ones for the same prompts at different substitution rates. As the substitution rate increased, the correlation between the text and image pairs decreased. In Figure 12, we measured Information-Vendi at various substitution rates and observed that as the substitution rate increased, Information-Vendi decreased, demonstrating that our score can successfully measure the correlation between text and image. Unlike other correlation metrics, such as CLIPScore, which require the same embedding for both text and image, our method places no such restriction. This allows for the use of different embeddings for text and image. Furthermore, our approach can be easily generalized to other conditional models, such as text-to-text or text-to-video generation.

Correlation between GroundTruth-Cluster-Vendi and Conditional-Vendi Scores. To validate the theoretical connection between the Vendi and Conditional-Vendi scores, we performed an experiment and evaluated a baseline metric called GroundTruth-Cluster-Vendi score. To measure the GroundTruth-Cluster-Vendi score, we utilize the side knowledge of the ground-truth clusters of the input prompts and then compute and average the regular Vendi scores for the data generated within each cluster. Mathematically, given t sample cluster sets in $\mathcal{S} = \{S_1, \ldots, S_t\}$, which partition the input text indices $\{1, \ldots, n\}$, we define the Cluster-Vendi score as follows, where $|S_j|$ denotes the cardinality of subset S_j:

$$\text{Cluster-Vendi}\big(x_1, \ldots, x_n \,|\, \mathcal{S}\big) := \sum_{i=1}^{t} \frac{|S_i|}{n} \cdot \text{Vendi}\Big(\{x_j : j \in S_i\}\Big),$$

$$\text{Cluster-RKE}\big(x_1,\ldots,x_n \,|\, \mathcal{S}\big) := \sum_{i=1}^{t} \frac{|S_i|}{n} \cdot \text{RKE}\Big(\{x_j : j \in S_i\}\Big).$$

Note that the above definition requires the knowledge of the clusters, which could be given by an oracle in the case of the GroundTruth-Cluster-Vendi score, or computed by a clustering algorithm such as K-Means to obtain the KMeans-Cluster-Vendi score. Observe that given the knowledge of the clusters revealed by an oracle, the GroundTruth-Cluster-Vendi score is a sensible definition of internal model diversity, which, as shown in Theorem 6, is expected to correlate with our defined Conditional-Vendi score.

In the numerical settings of Section G.2, where we know the ground-truth clusters based on the type of animal or fruit in the texts, we computed the value of the GroundTruth-Cluster-Vendi and GroundTruth-Cluster-RKE score and compared it with the evaluated Conditional-Vendi and RKE scores. As demonstrated in Figures 5,16, 17, 18, the two diversity scores, Conditional-Vendi, Cluster-Vendi and Conditional-RK,E and Cluster-RKE, highly correlate for the 4 simulated generative models in the experiments.

However, note that in a real-world scenario, we do not have access to the ground-truth clusters. To estimate the score, we should use a clustering algorithm such as K-Means to find the clusters and compute the Cluster-Vendi score. We note that the optimization problem addressed by standard clustering algorithms represents a challenging non-convex optimization, which, depending on the algorithm's initial point, could converge to different solutions.

Measuring Conditional-Vendi across prompt types In this section, we conducted additional experiments similar to those in Figure 6. We created 10,000 prompts across different categories using GPT-4o and generated corresponding images with text-to-image models. We reported Conditional-Vendi and RKE for the top 3 groups in the text data on PixArt-α, Stable Diffusion XL and FLUX text-to-image generative models.

As shown in Figure 13, Figure 14, and Figure 15, we observed the same behavior during these experiments: the Conditional-Vendi score for "dog" prompts was significantly higher than for the "airplane" and "sofa" categories. This observation suggests that the outputs of generative models are unbalanced when presented with different groups of text prompts.

G.2 Quantifying model-induced diversity via Conditional-Vendi and RKE.

To examine Conditional-Vendi in text-to-image models, we considered 10 types of animals generated by Stable Diffusion XL, as shown in Figure 5. We found that Conditional-Vendi and RKE increased more rapidly when the prompts did not specify the type of animal, indicating that the model-generated diversity was driven by its internal variability. In contrast, when the animal types were specified in the prompts, the increase in Conditional-Vendi and RKE was minimal, suggesting that the diversity in the outputs largely followed the constraints imposed by the text prompts. This demonstrates that Conditional-Vendi and Conditional-RKE effectively captures the difference between intrinsic model diversity and prompt-driven diversity. We further extended this experiment to different types of fruits and using a different generative model, PixArt-Σ (Figures 16, 17, 18), and observed the same trend. Additionally, we evaluated Cluster-Vendi and Cluster-RKE as ground-truth measures and observed the same pattern of Conditional-Vendi and RKE, confirming that Conditional-Vendi effectively captures model-intrinsic versus prompt-driven diversity.

To examine Conditional-Vendi in text-to-image models, we performed experiments quantifying diversity scores for unspecified and type-specified prompts across multiple categories and models. We considered nine experimental combinations consisting of two category types: animals and fruits and two state-of-the-art text-to-image models: Stable Diffusion XL (SDXL), and PixArt-Σ (Chen et al., 2024a). For each combination, we generated prompts for 10 different types within the category and created image samples by inputting the prompts into the respective model. In each experiment, we simulated 10 prompt-based generative models by considering image samples from j types for $j \in 1,\ldots,10$.

For animals generated by SDXL, as shown in Figure 5, Conditional-Vendi and RKE increased more rapidly when the prompts did not specify the type of animal, indicating that model-generated diversity was driven by intrinsic variability. In contrast, when the animal types were specified in the prompts, the increase in Conditional-Vendi and RKE was minimal, suggesting that diversity largely followed the constraints imposed by the text prompts.

We further extended this analysis to fruits and objects and to the PixArt-Σ model (Figures 16, 17, 18). Across all

categories and models, we observed the same trend: Conditional-Vendi and RKE increased rapidly for unspecified prompts but grew slowly when the type was specified, validating the correlation between Conditional-Vendi and intrinsic model diversity.

To further validate these results, we evaluated Cluster-Vendi and Cluster-RKE as ground-truth measures of non-prompt-induced diversity. The observed patterns mirrored those of Conditional-Vendi and RKE, confirming that Conditional-Vendi effectively captures the difference between intrinsic model diversity and prompt-driven diversity.

G.3 Convergence Analysis of Conditional-Vendi Score

To assess the convergence of the Conditional-Vendi and Conditional-RKE scores, we conducted experiments for different sample sizes on samples generated with SDXL and Kandinsky using prompts from the MS-COCO 2014 validation set. We used the cosine similarity for the finite-dimensional kernel and the Gaussian kernel for the infinite-dimensional kernel. Our results, presented in Figure 19, show that for RKE, Conditional-RKE converged, while for Conditional-Vendi, the non-truncated score did not converge; our proposed truncated Conditional-Vendi converged with 15000 samples.

G.4 Additional Numerical Evaluation of the Conditional-Vendi Score

Text-to-Video Model Evaluation. For the experiments on video data, to ensure the fairness of our evaluation, we used VBench samples (Huang et al., 2024), which generated samples belonging to the 8 content categories. In Figure 22, we used VideoCrafter-1, Show-1, and Open-Sora-1.2. We observed that VideoCrafter videos look less diverse and, in some cases, may not correlate significantly with the captions when compared to Open-Sora. Confirming this observation, the Conditional-Vendi and Information-Vendi scores were lower for VideoCrafter than those for Open-Sora.

Image-Captioning Evaluation. For image captioning, we used 10 classes from the ImageNet dataset as input for BLIP-2, GIT and GPT4o-mini. In Figure 20, we compared captions for the top three groups of images: gas pump, church, and cassette player. GIT generated more diverse captions compared to BLIP, which was confirmed by the Conditional-Vendi scores. On the other hand, GPT4o-mini generated longer and more detailed captions compared to GIT, which was also reflected in the evaluated Conditional-Vendi and Information-Vendi scores.

Large Language Models Evaluation. To evaluate Conditional-Vendi and RKE on LLMs, we varied the temperature parameter and generated 20K short stories with Llama 2 for each temperature setting as shown in Table 2. We also provided a comparison of the generated prompts in Figure 21. The dataset covered 10 genres, each with 20 distinct subjects and themes. We further tested Conditional-Vendi and Conditional-RKE scores on Gemma 3 and Phi 4 Mini Microsoft et al. (2025). As shown in Tables 4 and 5, both Conditional-Vendi and Conditional-RKE increase with higher temperatures, indicating that the outputs become more diverse.

Table 4: Conditional Vendi and RKE Scores evaluated for Gemma 3 with different temperature parameters.

Method	$T = 0.4$	$T = 0.7$	$T = 1.0$	$T = 1.3$
Conditional-Vendi	40.82	42.82	44.03	48.23
Conditional-RKE	38.42	41.82	43.16	45.93

Table 5: Conditional Vendi and RKE Scores evaluated for Phi 4 Mini with different temperature parameters.

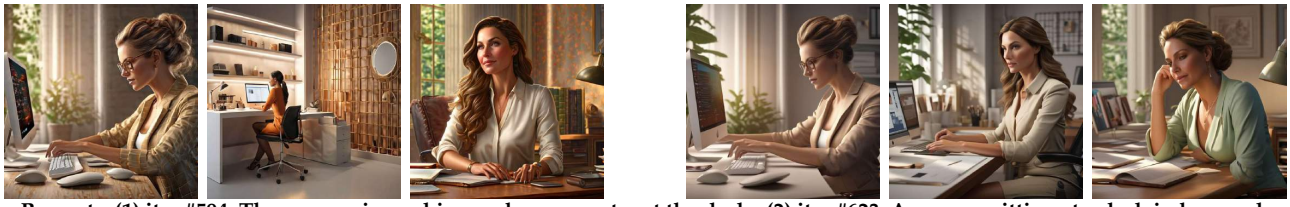
Method	$T = 0.4$	$T = 0.7$	$T = 1.0$	$T = 1.3$
Conditional-Vendi	41.02	45.93	49.93	51.60
Conditional-RKE	39.43	41.74	47.27	49.82

Qualitative results for generative models trained on MS-COCO dataset In this section, we provide images and prompts corresponding to Figure 23. Figure 24 illustrates three clusters obtained by applying KMeans

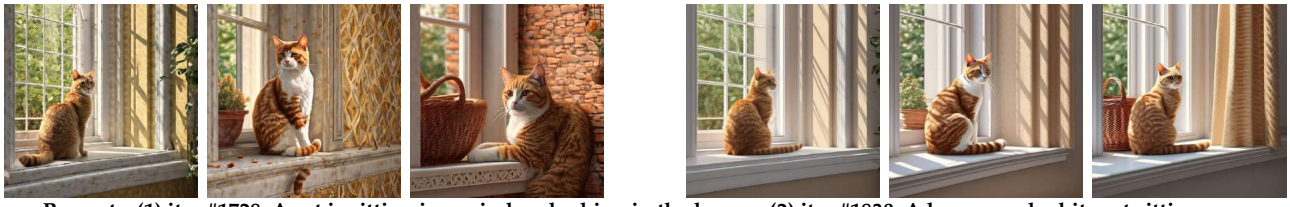
to cluster MS-COCO validation set prompts into 1000 clusters. The images are presented for four generative models. Comparing the prompts with the generated images reveals that FLUX exhibits the highest alignment between text and image, while GigaGAN demonstrates greater diversity but misses some features of the prompts. These observations are further supported by the Conditional-Vendi and Information-Vendi metrics.

Stable Diffusion XL

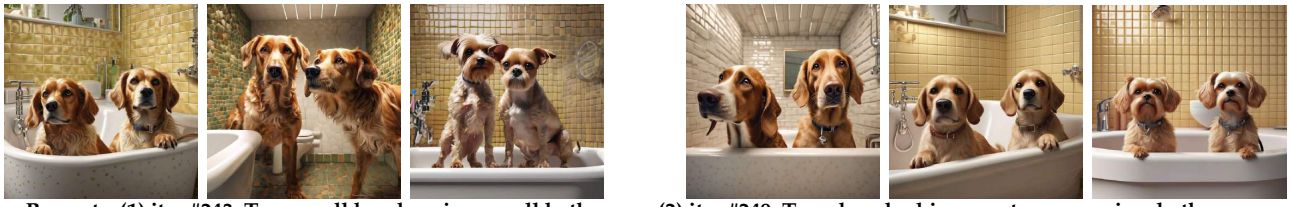
Conditional Vendi Score Guidance | **Vendi Score Guidance**



Prompts: (1) iter #594: The woman is working on her computer at the desk., (2) iter #623: A woman sitting at a desk in her work station., (3) iter #734: a woman at her desk sits intently and happily.



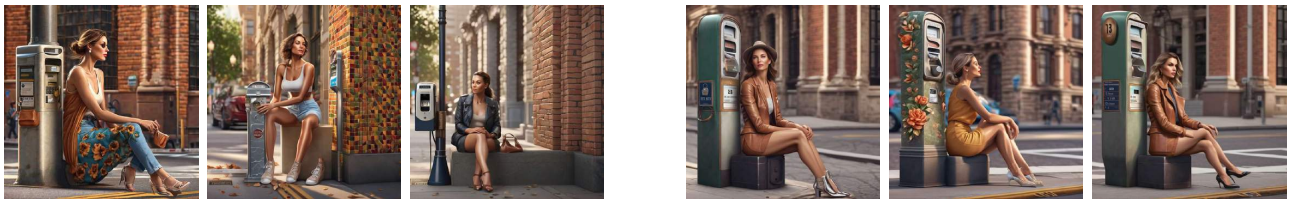
Prompts: (1) iter #1728: A cat is sitting in a window looking in the house., (2) iter #1830: A brown and white cat sitting on a window sill., (3) iter #2456: A cat sitting on a window sill near a basket.



Prompts: (1) iter #243: Two small lap dogs in a small bathroom., (2) iter #249: Two dogs looking up at a camera in a bathroom., (3) iter #436: Two small dogs stand together in a bathroom.

Prompts: (1) iter #1778: A cat is standing on a table next to the television., (2) iter #1895: A cat looking at a dog on a TV., (3) iter #2024: Cat looking around television with dog on it.

Prompts: (1) iter #1802: An orange and white cat standing in front of a flat screen TV., (2) iter #1848: A cat climbing on top of a shelf., (3) iter #196: A cat perches on top of an entertainment center in front of a TV.



Prompts: (1) iter #1557: A woman sitting next to a parking meter., (2) iter #1559: A woman is sitting on the curb with a decorated parking meter., (3) iter #1675: a woman sits in front of a parking meter.

Figure 9: Qualitative comparison of Conditional-Vendi score guidance vs. Vendi score guidance using SD-XL.

Figure 10: Qualitative comparison of Conditional-Vendi score guidance vs. Vendi score guidance using PixArt-Σ.

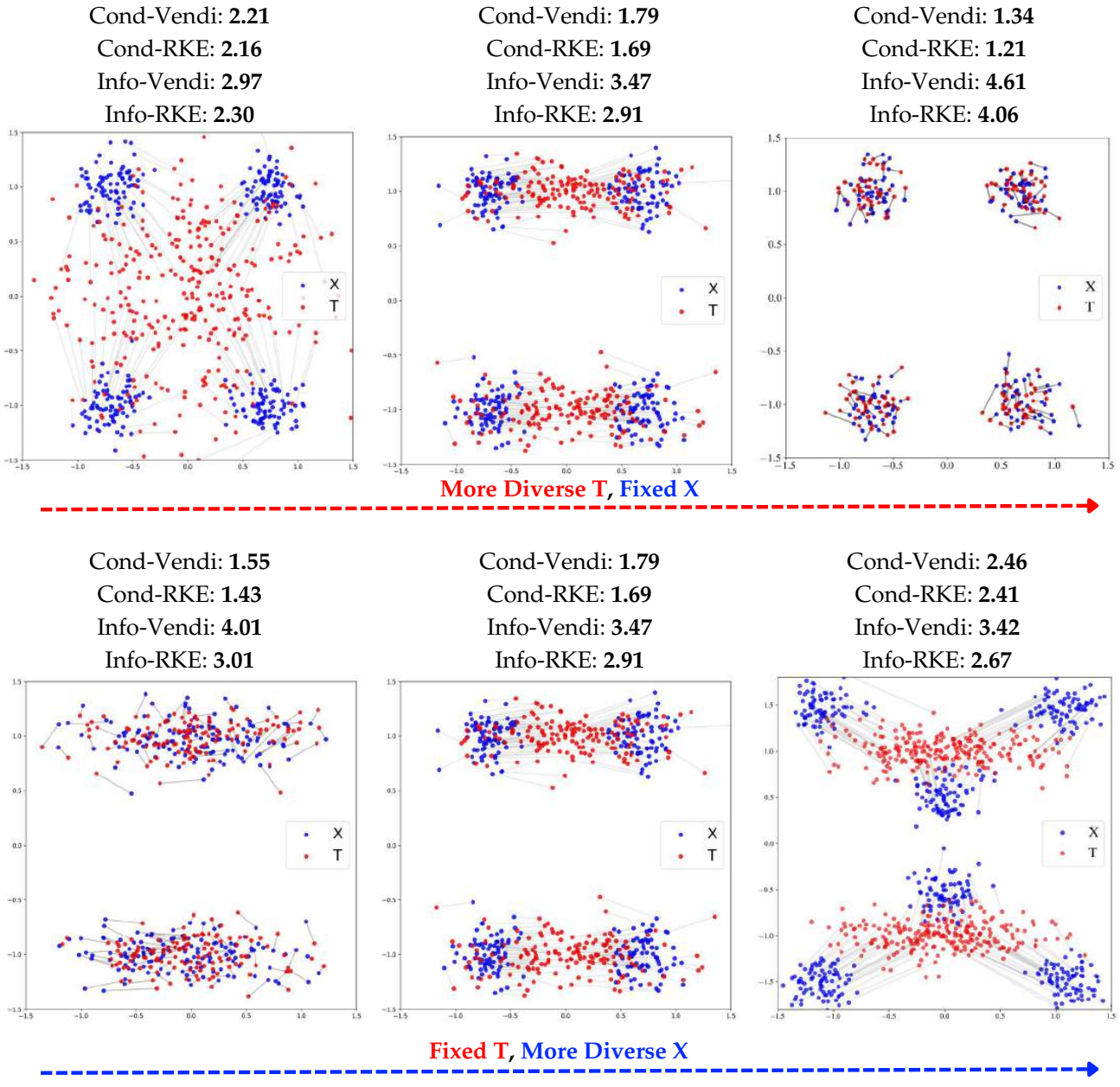


Figure 11: Comparing Conditional-Vendi and Information-Vendi on 2-D Gaussian Distribution. We used 1000 pair of points and used a Gaussian Kernel with bandwidth $\sigma = 0.6$.

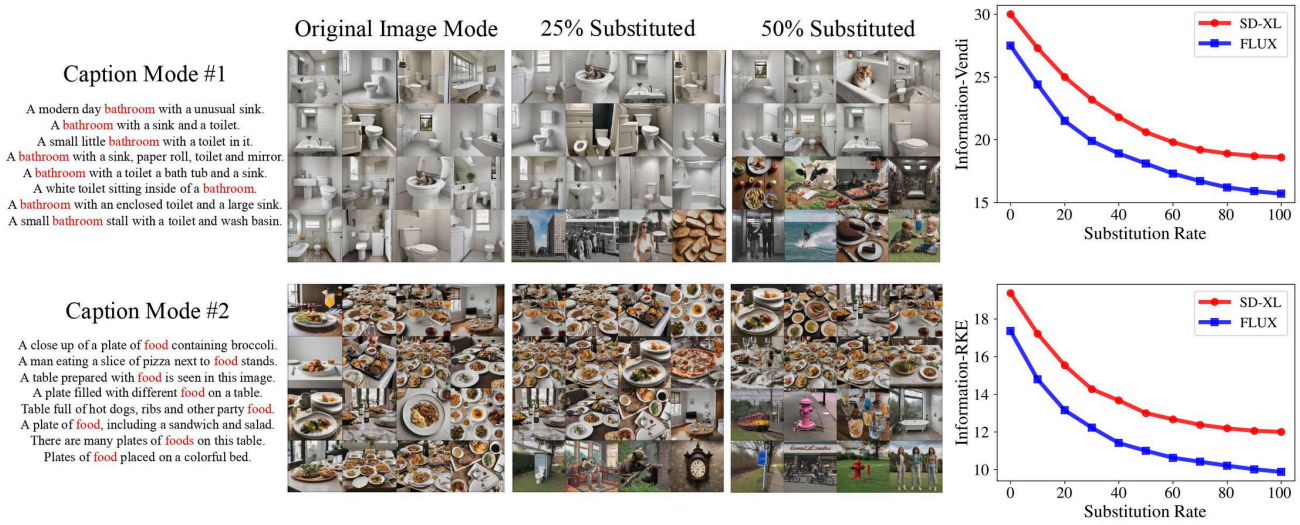


Figure 12: Substituting images generated from models trained on MS- dataset.

Figure 13: Quantifying image diversity for different clusters of text prompts. Images are generated using the Stable Diffusion XL model.

Figure 14: Quantifying image diversity for different clusters of text prompts. Images are generated using the PixArt-α model.

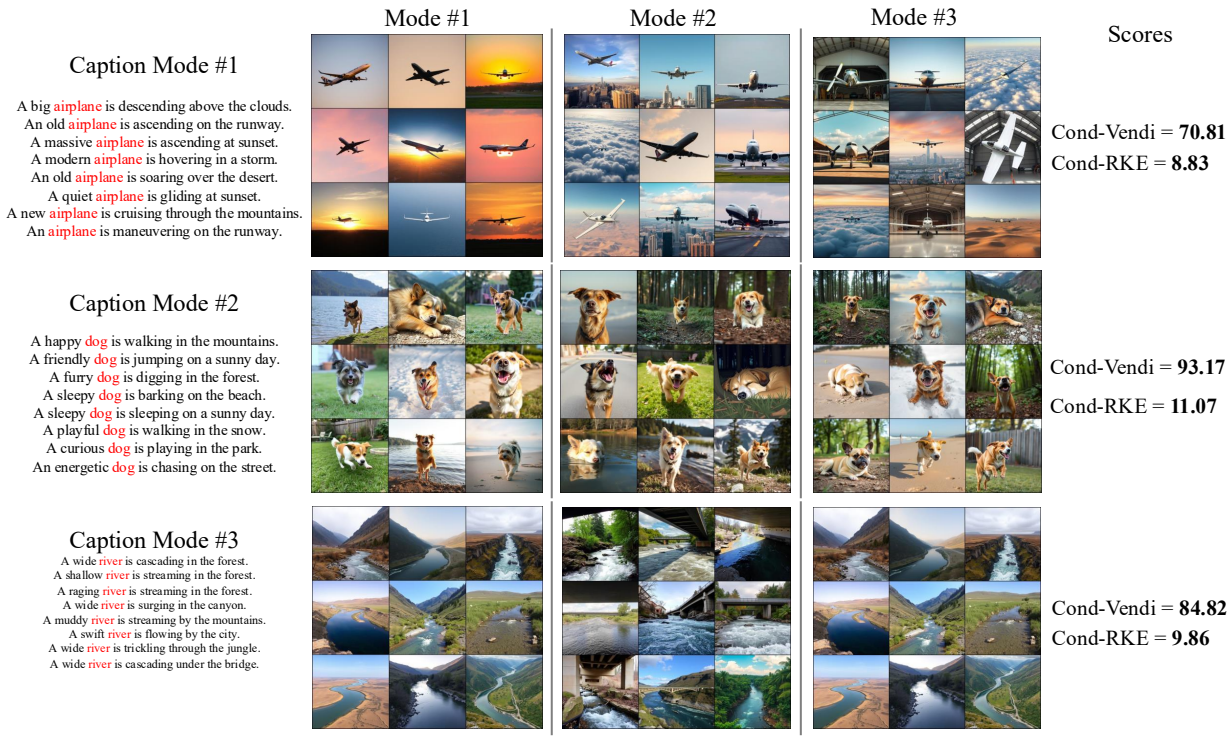
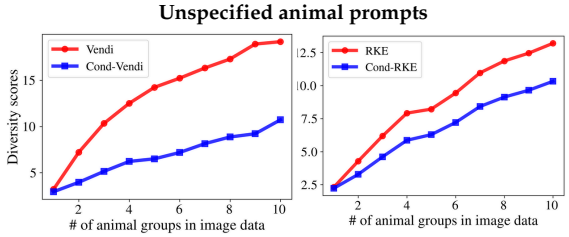
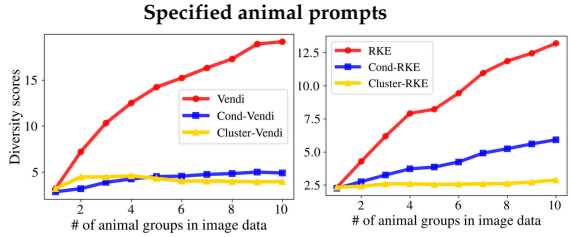


Figure 15: Quantifying image diversity for different clusters of text prompts. Images are generated using the Flux model.



Prompts: An **animal** is rolling in the grass., An **animal** resting near a sand dune.An **animal** is drinking from a river., An **animal** is resting under a tree., An **animal** is resting in a grassy pasture., An **animal** is walking through jungle., An **animal** is reaching up for leaves.



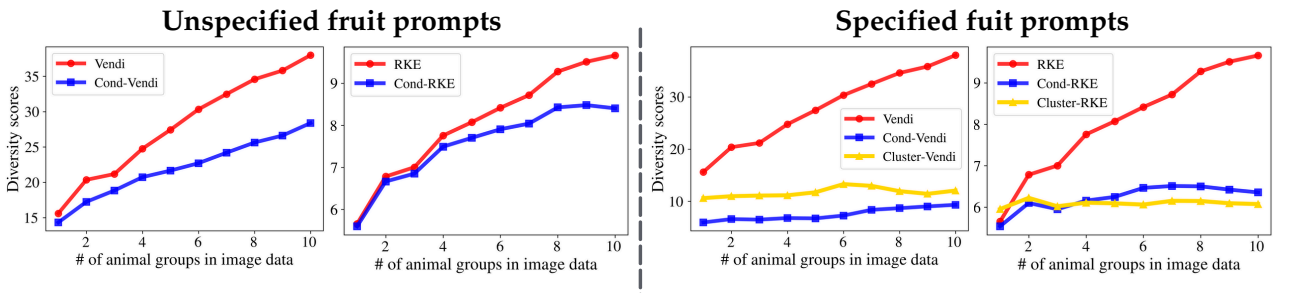
Prompts: A **fox** is rolling in the grass., A **camel** is resting near a sand dune., A **wolf** is drinking from a river., A **cow** is resting under a tree., A **sheep** is resting in a grassy pasture., An **elephant** is walking through thick jungle., A **giraffe** is reaching up for leaves.

Model 2:
Samples from 2 animal groups

Model 4:
Samples from 4 animal groups

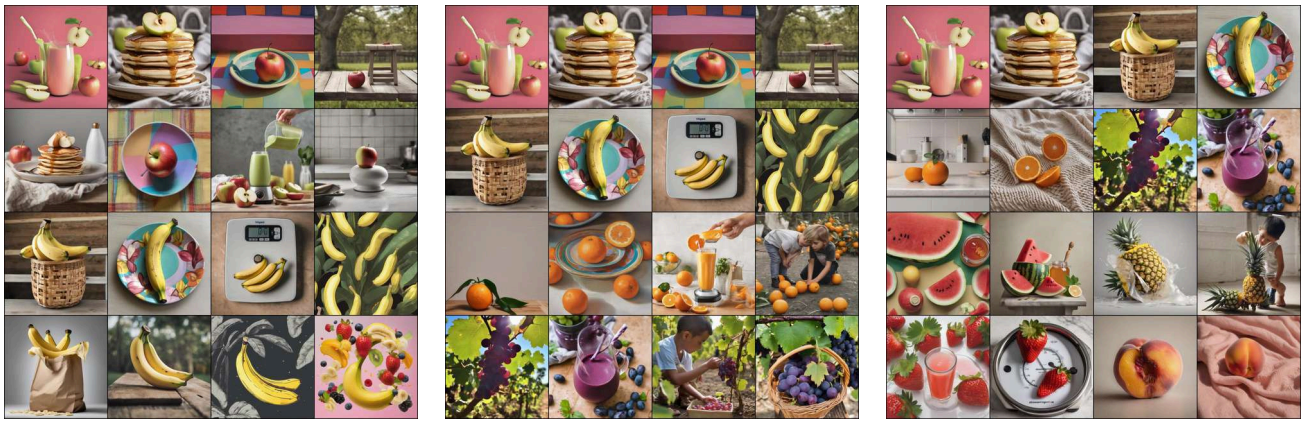
Model 8:
Samples from 8 animal groups

Figure 16: Evaluated Conditional-Vendi, Vendi, Conditional-RKE, and RKE scores on animal samples generated by PixArtΣ. (Left Plot) We do not specify the animal types in the prompt (Right Plot) we specify the animal types in the prompt.



Prompts: A **fruit** is next to a cold glass of fresh juice. A **fruit** is being sliced with a sharp knife. The **fruit** is blended into a smoothie. The **fruit** is falling out of a grocery bag. A **fruit** is being washed. The **fruit** is sitting on a kitchen countertop.

Prompts: An **apple** is next to a cold glass of fresh juice. A **strawberry** is being sliced with a sharp knife. The **peach** is blended into a smoothie. The **cherry** is falling out of a grocery bag. A **mango** is being washed. The **fig** is sitting on a kitchen.

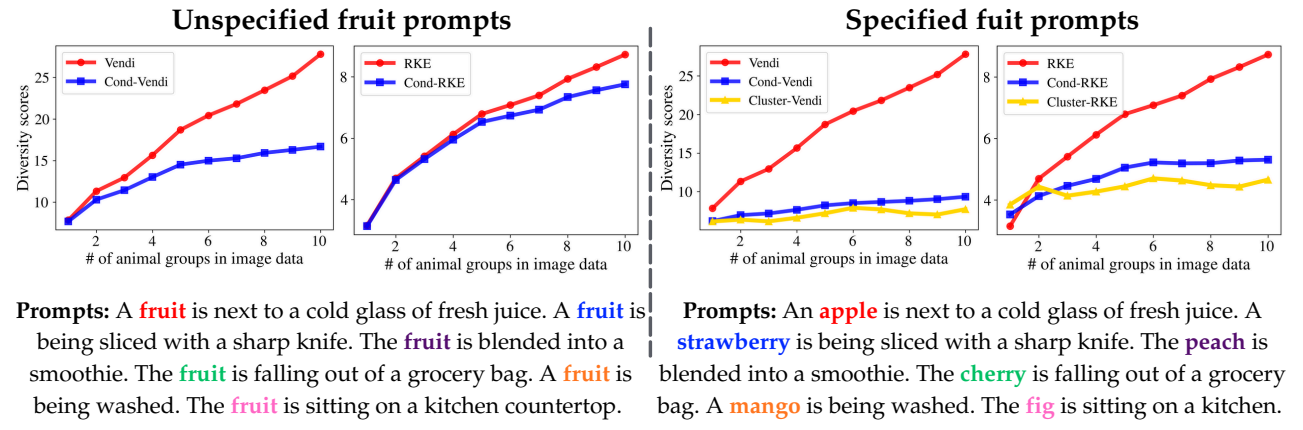


Model 2:
Samples from 2 fruit types

Model 4:
Samples from 4 fruit types

Model 8:
Samples from 8 fruit types

Figure 17: Comparing Conditional-Vendi with Vendi on different fruit types generated by Stable Diffusion-XL.



Model 2: Samples from 2 fruit types

Model 4: Samples from 4 fruit types

Model 8: Samples from 8 fruit types

Figure 18: Comparing Conditional-Vendi with Vendi on different fruit types generated by PixArt-Σ.

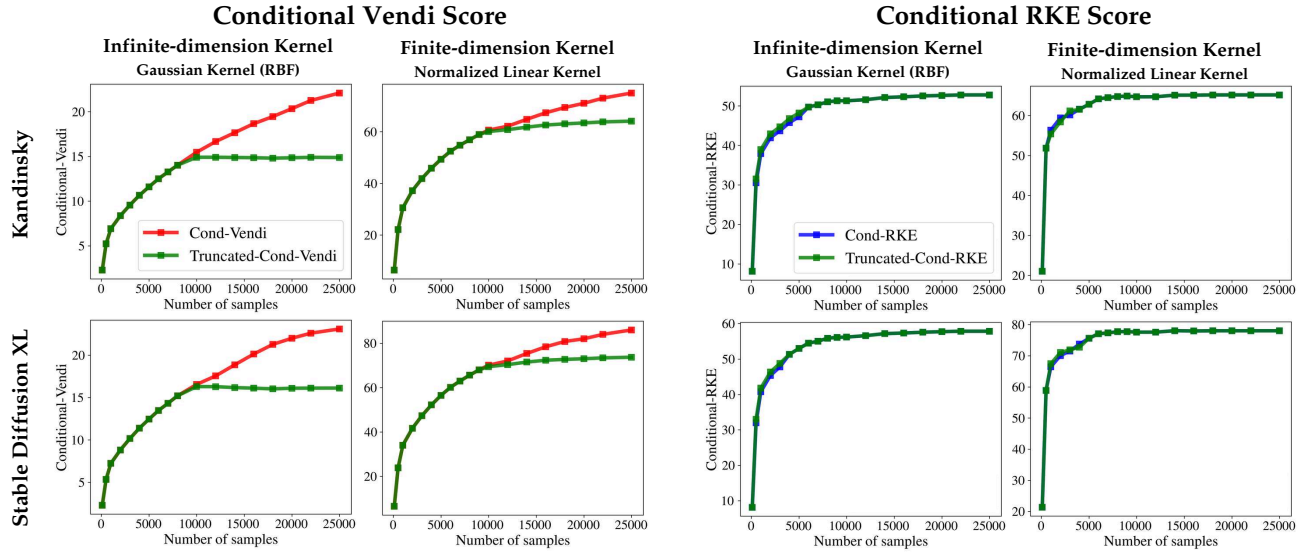


Figure 19: Statistical convergence of Conditional-Vendi scores with different sample sizes on data generated by Stable Diffusion-XL and Kandinsky using MS-COCO validation set prompts with finite-dimension cosine similarity and infinite-dimension Gaussian kernel. DINOv2 and CLIP embeddings are used for image and text modalities, respectively.

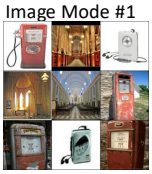
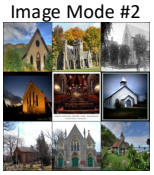
	BLIP-2	GIT	GPT4o-mini
Image Mode #1 	a gas pump with three different colors. a gas pump with two different types of fuel. a man filling up his car at a gas station. a woman filling up her car at a gas station. an old red gas pump in front of a barn. an old gas pump in front of a house.	a gas pump at a gas station with a red gas pump. a gas pump at a gas station in the middle of a road. a woman in a white dress is filling up a car. a vintage gas station with a man filling up a gas pump. an old gas pump sitting in a field. an old gas pump sitting in front of a building.	A vintage gas pump with a faded sign, displaying old pricing and a weathered metallic finish. A man stands at a gas pump, holding a nozzle and filling two metal containers on a cart. He wears a black vest over a blue shirt and denim jeans, with a focused expression. In the background. A vintage green gas pump stands in a grassy area, surrounded by trees and a white building in the background.
Image Mode #2 	the inside of a church with wooden benches. the interior of a church with wooden pews. a church with a steeple and a graveyard a church with a graveyard and a clock tower a white church with two steeple towers a white church with a red door and a fence	the interior of the cathedral of the holy trinity. the interior of a church with stained windows. a small church with a steeple and a sign that says "catholic church ". a church with a stone roof and a graveyard. a white church with a steeple and a blue sky. a white church with a steeple and steps.	Interior of a spacious church featuring wooden pews, stained glass windows, and a decorative altar at end. a tall stone church with a pointed steeple and multiple spires, surrounded by a graveyard with weathered tombstones. a white church with a steeple, surrounded by dark clouds, with light one side of the building.
Image Mode #3 	a stereo system with a record player. a stereo system with a cassette player and a radio. a blue cassette player with headphones and a microphone. a silver portable cassette player with earphones. the interior of a car with a radio and dash a car radio with a digital display and buttons	the stereo system is in the floor of the computer room. the radio is a stereo system that can be used to record or record a player. the boombox is a compact cassette player that can be used as a microphone. sony's original sony audio cassette player. the radio is a compact car that can be used as a radio.	a stereo system with two large speakers, a display panel in the center, and a remote control beside it. a vintage boombox with a cassette tape deck and speakers, featuring various buttons and a handle on top. car dashboard displaying radio and climate control buttons, air vents above, and a textured surface around the controls
Evaluated Scores:	Cond-Vendi = **4.09** Info-Vendi = **2.28** Cond-RKE= **2.53** Info-RKE= **1.80**	Cond-Vendi = **10.9** Info-Vendi = **3.28** Cond-RKE= **6.47** Info-RKE= **2.41**	Cond-Vendi = **13.10** Info-Vendi = **3.98** Cond-RKE= **7.82** Info-RKE= **2.93**

Figure 20: Conditional-Vendi and Information-Vendi of image-captioning models for 3 image types

Prompt: Write a short story about A lost letter:

T=0.4

1. It was a typical Wednesday morning when Sarah received a mysterious package in the mail. She had been expecting an important document from her bank, but instead, she found a small, unassuming envelope with no return address.....

2. It was a typical Wednesday morning when Sarah received an unexpected package in the mail. She had been expecting a delivery from her favorite clothing store, but instead, she found a small, unassuming envelope with no return address....

3. It was a typical Wednesday morning when Sarah received a mysterious package in the mail. She had not ordered anything, and there was no return address on the package....

T = 1.3

1.It was a typical Wednesday evening when Sarah walked home from work, Lost in thought as she navigated the busy streets of the city, She had just received some devastating news at work and her mind was racing with emotions.....

2. It was a beautiful spring morning when Sarah received an unexpected package in the mail. Inside, she found a letter that wasn't there a moment ago. The letter was addressed to her, but there was no return address and no indication of who had sent it....

3. It was a cold and rainy day when Emily received the news that her grandmother had passed away. She had been incommunicado for weeks, and Emily had been worried sick about her....

Prompt: Write a short story about A suspicious neighbor:

T=0.4

1. Mrs. Johnson had lived in her cozy little house on Elm Street for over 20 years, and she thought she knew every person on the block like family. But lately, she had grown suspicious of her new neighbor, Mr. Smith.

2. Mrs. Johnson had lived in her cozy little house on Elm Street for over 20 years. She was a kind and friendly person, always willing to lend a helping hand to those in need. However, there was one neighbor who made her feel uneasy.

3. Mrs. Johnson had lived in her cozy suburban home for over 20 years, and she thought she knew every person on her street. However, one day a new family moved in next door...

T = 1.3

1. It was a typical Wednesday evening when I first noticed my neighbor, Mr. Johnson, behaving strangely. I had just finished cooking dinner and was taking out the trash when I saw him peering through his living room window, ...

2. The sun was setting over the small suburban town, casting a warm orange glow over the neatly manicured lawns and tidy homes. It was a peaceful evening but in one house on Elm Street, an air of unease hung heavy in the air.....

3. Mrs. Johnson had lived in her cozy suburban home for over 20 years, and she knew every face on her street. However, one day she noticed a new neighbor move in next door.....

Figure 21: Llama 2 Samples generated with different temperature parameters.

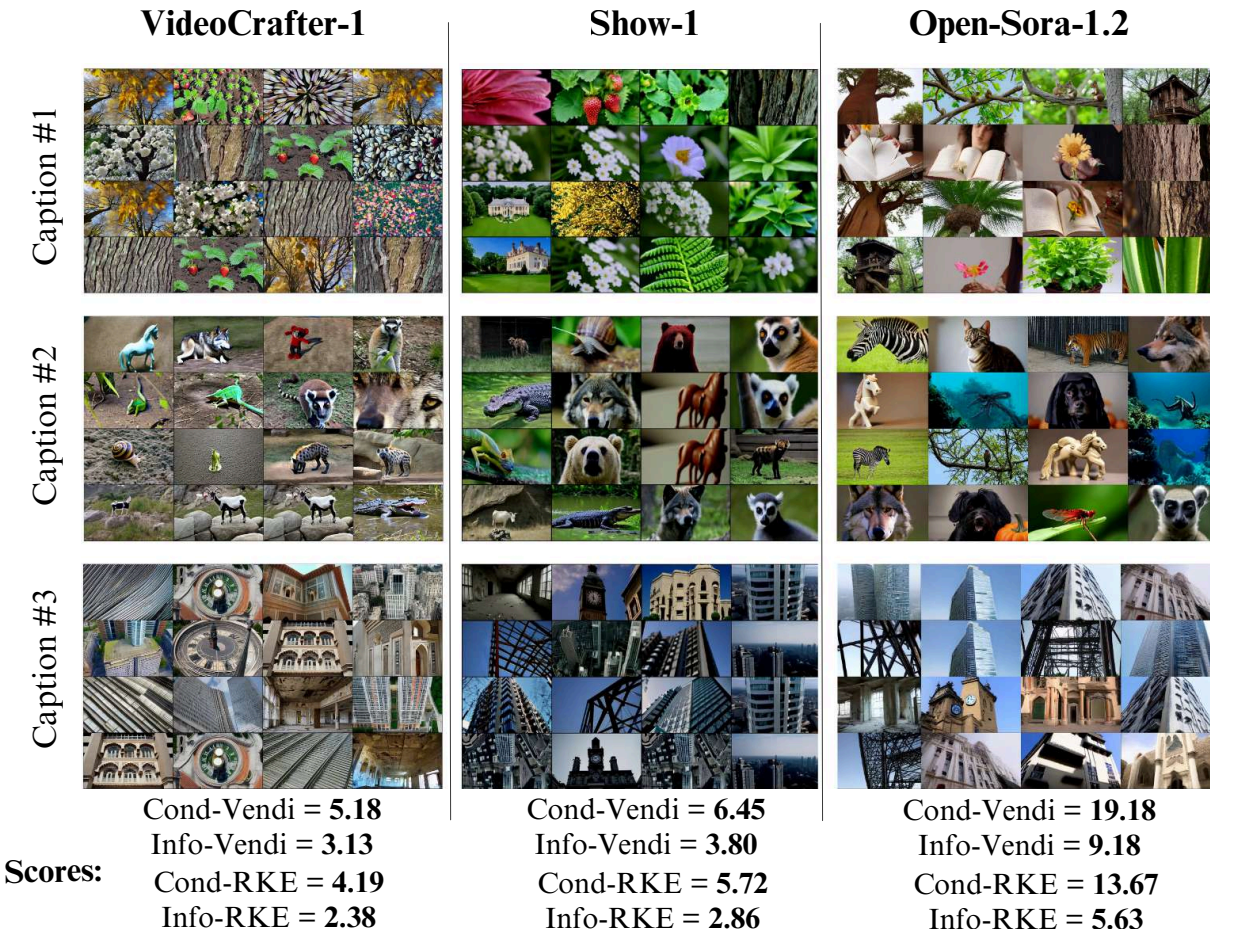


Figure 22: Measuring Conditional-Vendi and Information-Vendi for text-to-video models

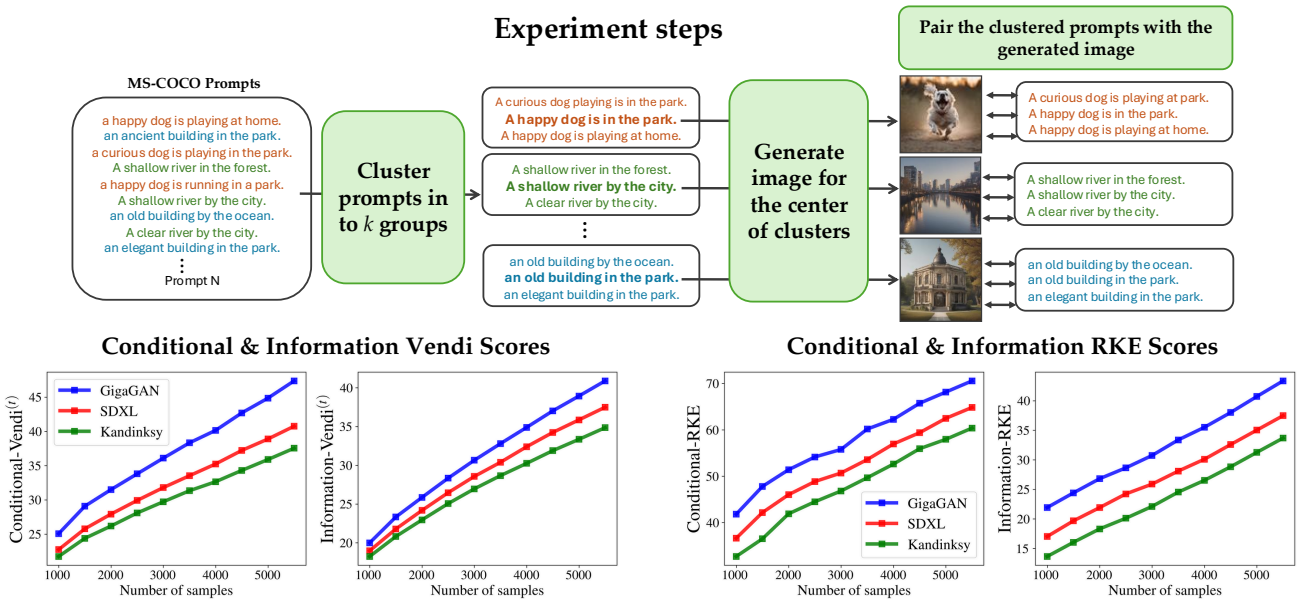
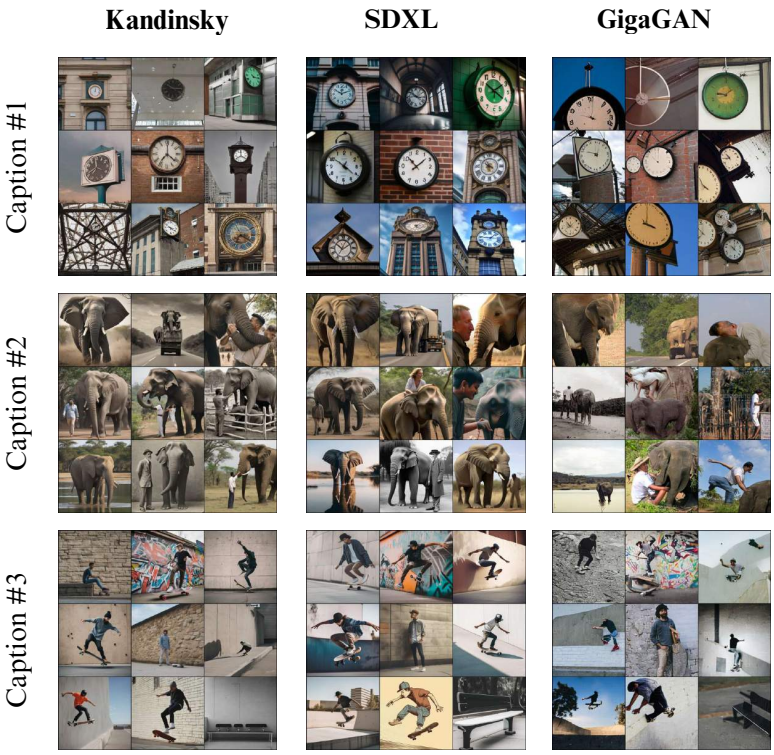


Figure 23: Conditional and Information Vendi and RKE score comparison across text-to-image models. We clustered MS-COCO prompts into k groups and generated images for each cluster center. Within each cluster, we paired prompts with identical images. The results show increasing diversity and stronger correlation as the number of clusters grows, indicating that clusters become more relevant and diverse with finer partitioning.



Caption Mode #1

A building displaying a clock showing the time to be 6 oclock.
A clock hanging from the ceiling of a building.
A large metal green clock hanging from the side of a building.
A clock that is on top of a sign.
A large clock mounted to a brick wall.
A large clock hanging off the side of a tall building.
A clock in near the triangular roof of a large building.
A large clock and a sign on top of a building.
A large clock mounted to the side of a building.

Caption Mode #2

The elephant has a large white spot on its abdomen.
The truck driver hauls an elephant down the highway.
A man getting a kiss on the neck from an elephant's trunk
A large elephant walking next to a man
A woman in white shirt climbing onto an elephant.
A man is leaning over a fence offering food to an elephant/
A large elephant standing on the side of a lake.
A man standing next to an elephant who stole his hat with it's trunk.
A man standing near an elephant with its trunk outstretched.

Caption Mode #3

A young man riding a skateboard on a stone wall.
A man balancing on a skateboard in front of a graffiti covered wall.
A man doing a trick on a wall with a skateboard.
Bearded skateboarder maintains balance while skating up wall.
A man standing next to a stone wall while holding a skateboard.
There is a man skateboard on the side of a wall.
a guy skate boarding on the edge of a wall
A man on a skateboard is trying to jump over a wall.
two black and white skate boards under a black steel bench

Figure 24: 3 clusters of MS-COCO generated samples