A. Appendix

B. Statistical Interpretation of the Entropy-based Scores

In this section, we aim to statistically interpret the defined conditional diversity scores as the expectation of prompt-specific entropy $H(X|T=t)$. First, we derive the statistic estimated from empirical samples by the entropy-based scores, and then we connect the conditional entropy measure to the expectation of unconditional entropy values. According to the Schur product theorem, the Hadamard product $K_X \odot K_T$ of PSD kernel matrices K_X, K_T will also be a PSD kernel matrix. We note that the kernel matrix corresponds to the following feature map $\phi_{X,T} : \mathcal{X} \times \mathcal{T} \to \mathbb{R}^{d_x d_t}$ where $\otimes$ denotes the Kronecker product:

$$\phi_{X,T}\big([x,t]\big) \;=\; \phi_X(x) \otimes \phi_T(t)$$

The above holds due to the identity $\big\langle \phi_{X,T}\big([x,t]\big)\,,\, \phi_{X,T}\big([x',t']\big)\big\rangle = k_X\big(x,x'\big)k_T\big(t,t'\big)$. The following proposition formulates kernel-based conditional entropy and mutual information using $\phi_{X,T}$.

Proposition 1. *Consider the kernel matrices K_X for samples $x_1,\ldots,x_n$ and K_T for samples $t_1,\ldots,t_n$. Then, $\frac{1}{n}K_X \odot K_T$ (used for defining joint entropy $H_\alpha(X,T)$) share the same non-zero eigenvalues with the following kernel covariance matrix:*

$$\widehat{C}_{X,T} \;:=\; \frac{1}{n}\sum_{i=1}^{n} \phi_{X,T}\big([x_i,t_i]\big)\phi_{X,T}\big([x_i,t_i]\big)^\top \;=\; \frac{1}{n}\sum_{i=1}^{n}\Big[\phi_X\big(x_i\big)\phi_X\big(x_i\big)^\top\Big] \otimes \Big[\phi_T\big(t_i\big)\phi_T\big(t_i\big)^\top\Big]$$

Proof. We defer the proof to the Appendix C.1. □

Corollary 1. *Consider the composite feature map $\phi_{X,T}$ and joint kernel covariance matrix $\widehat{C}_{X,T}$ defined above. Then, given marginal kernel covariance matrices $\widehat{C}_X = \frac{1}{n}\sum_{i=1}^{n}\phi_X(x_i)\phi_X(x_i)^\top$, $\widehat{C}_T = \frac{1}{n}\sum_{i=1}^{n}\phi_T(t_i)\phi_T(t_i)^\top$, the following holds for the defined conditional entropy and mutual information:*

$$H_\alpha(X|T) = \; H_\alpha(\widehat{C}_{X,T}) - H_\alpha(\widehat{C}_T), \quad I_\alpha(X;T) = \; H_\alpha(\widehat{C}_X) + H_\alpha(\widehat{C}_T) - H_\alpha(\widehat{C}_{X,T})$$

Proof. The proof is deferred to the Appendix C.2. □

Corollary 1 shows that given the underlying covariance matrices $C_X = \mathbb{E}_{x\sim P_X}\big[\phi_X(x)\phi_X(x)^\top\big]$, $C_T = \mathbb{E}_{t\sim P_T}\big[\phi_T(t)\phi_T(t)^\top\big]$, and $C_{X,T} = \mathbb{E}_{(x,t)\sim P_{X,T}}\big[\phi_{X,T}([x,t])\phi_{X,T}([x,t])^\top\big]$, the defined entropy-based scores converge to the following statistics when the sample size n tends to infinity:

$$\widetilde{H}_\alpha(X|T) = H_\alpha(C_{X,T}) - H_\alpha(C_T), \quad \widetilde{I}_\alpha(X;T) = H_\alpha(C_X) + H_\alpha(C_T) - H_\alpha(C_{X,T}).$$

Note that the entropy-based statistic $H_\alpha(C_X)$ represents the statistic estimated by the logarithm of the Vendi score defined in [11].

Next, we prove that for a mixture text distribution P_T where the text variable follows random mode $G \in \{1,\ldots,m\}$, the defined conditional entropy score aggregates the expectation of the unconditional entropy score $H(X|G=i)$over the m text modes $1,\ldots,m$.

Theorem 1. *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 $\alpha \geq 2$, the matrix-based order-α conditional entropy satisfies the following inequality where $g(z) = \frac{\alpha}{\alpha-1}\log\big(\frac{1}{1-z/\|\boldsymbol{\omega}\|_\alpha}\big)$ is an increasing scalar function with $g(0)=0$:*

$$\Big|\widetilde{H}_\alpha(X|T) - f^{-1}\Big(\mathbb{E}_{I\sim\boldsymbol{\omega}^\alpha}\Big[f\big(\widetilde{H}_\alpha(X|G=I)\big)\Big]\Big)\Big| \leq$$
$$2\,g\Big(32\sum_{i=1}^{k}\omega_i\Big[\frac{\sigma_i^2}{\sigma^2} + \sum_{j=1}^{i-1}\exp\Big(\frac{-\|\mu_i-\mu_j\|_2^2}{\sigma^2}\Big)\Big]\Big)$$

Proof. The proof is deferred to the Appendix C.3. □

The above theorem shows that if the text samples come from m distinct modes satisfying $\frac{\|\mu_i-\mu_j\|_2}{\sigma} \gg 1$ for every $i \neq j$ and $\frac{\mathbb{E}_{P_{T,i}}[\|T-\mu_i\|_2^2]}{\sigma^2} \ll 1$, then the defined conditional entropy score $H(X|T)$ aggregates the unconditional entropy score $H(X|G=i)$ given the prompt group. Therefore, this result extends the expectation-based interpretation of Shannon conditional entropy to the matrix-based conditional entropy defined in [12].

Based on Theorem 1, we propose a text-type-based diversity evaluation, where we restrict the evaluation of the prompt-based generative model to the prompts in the same group, i.e. the same mode in the mixture text distribution. To do this, we find the eigendirections corresponding to the text clusters by performing an eigendecomposition of text kernel matrix $\frac{1}{n}K_T = \sum_{i=1}^n \lambda_i v_i v_i^\top$ where $\lambda_1 \geq \cdots \lambda_n$ are the sorted eigenvalues and $v_1, \ldots, v_n$ are the sorted eigenvectors. Then, we note that the Hadamard product $\frac{1}{n}K_X \odot K_T$ used for joint entropy can be decomposed as: $\frac{1}{n}K_X \odot K_T = \sum_{i=1}^n \lambda_i \big(K_X \odot v_i v_i^\top\big)$. As a result, to evaluate the diversity of the generation model, we apply eigendocomposition to each $K_X \odot v_i v_i^\top$, where v_i marks the samples in group i, and find the sample indices with the maximum entries on the principal eigenvectors of $K_X \odot v_i v_i^\top$.

C. Proofs

C.1. Proof of Proposition 1

First, we observe that for every $x, x' \in \mathcal{X}$ and $t, t' \in \mathcal{T}$, the following holds:

$$\begin{aligned}
\phi_{X,T}([x,t])^\top \phi_{X,T}([x',t']) &= \big(\phi_X(x) \otimes \phi_T(t)\big)^\top \big(\phi_X(x') \otimes \phi_T(t')\big) \\
&= \big(\phi_X(x)^\top \phi_X(x')\big) \otimes \big(\phi_T(t)^\top \phi_T(t')\big) \\
&= k_X(x,x') \otimes k_T(t,t') \\
&= k_X(x,x') k_T(t,t')
\end{aligned}$$

Therefore, the Hadamard product of kernel matrices $\frac{1}{n}K_X \odot K_T$ can be written as

$$\frac{1}{n}K_X \odot K_T = \frac{1}{n}\Phi_{X,T}\Phi_{X,T}^\top$$

in terms of the matrix of samples' feature maps $\Phi_{X,T} \in \mathbb{R}^{n \times d_x d_t}$ with its ith row being $\phi_{X,T}([x_i,t_i])$. We observe that the matrices $\frac{1}{n}\Phi_{X,T}\Phi_{X,T}^\top$ and $\frac{1}{n}\Phi_{X,T}^\top\Phi_{X,T}$ share the same non-zero eigenvalues, that are the square of the singular values of $\Phi_{X,T}$. Therefore, $\frac{1}{n}K_X \odot K_T$ has the same non-zero eigenvalues as the following matrix

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

which is the defined matrix $\widehat{C}_{X,T}$. Therefore, the proof of the proposition is complete.

C.2. Proof of Corollary 1

As we showed in Proposition 1, the Hadarmard product $\frac{1}{n}K_X \odot K_T$ shares the same non-zero eigenvalues with $\widehat{C}_{X,T}$. Also, as noted by [17], $\frac{1}{n}K_X$ and $\frac{1}{n}K_T$ have the same non-zero eigenvalues as of $\widehat{C}_X$ and $\widehat{C}_T$, respectively. Since the order-α matrix-based entropy is only a function of the input matrix's non-zero eigenvalues (zero eigenvalues have no impact on the entropy value), we can conclude that

$$\begin{aligned}
H_\alpha(X|T) &:= H_\alpha\big(\frac{1}{n}K_X \odot K_T\big) - H_\alpha\big(\frac{1}{n}K_T\big) \\
&= H_\alpha(\widehat{C}_{X,T}) - H_\alpha(\widehat{C}_T),
\end{aligned}$$

and also

$$\begin{aligned}
I_\alpha(X;T) &:= H_\alpha\big(\frac{1}{n}K_X\big) + H_\alpha\big(\frac{1}{n}K_T\big) - H_\alpha\big(\frac{1}{n}K_X \odot K_T\big) \\
&= H_\alpha(\widehat{C}_X) + H_\alpha(\widehat{C}_T) - H_\alpha(\widehat{C}_{X,T}).
\end{aligned}$$

C.3. Proof of Theorem 1

To prove Theorem 1, we begin by showing the following lemma.

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

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

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

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

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 $1 - \exp(-z) \leq z$ 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

$$u_i \;:= \phi(\mu_i) - \sum_{j=1}^{i-1} \langle \phi(\mu_i), u_j \rangle u_j.$$

As a result, the following holds

$$\begin{aligned}
&\Big\| \sum_{i=1}^{m} \omega_i C_i \otimes \phi(\mu_i)\phi(\mu_i)^\top - \sum_{i=1}^{m} \omega_i C_i \otimes u_i u_i^\top \Big\|_F^2 \\
&= \Big\| \sum_{i=1}^{m} \omega_i C_i \otimes \Big(\phi(\mu_i)\phi(\mu_i)^\top - u_i u_i^\top \Big) \Big\|_F^2 \\
&\overset{(h)}{\le} \sum_{i=1}^{m} \omega_i \Big\| C_i \otimes \Big(\phi(\mu_i)\phi(\mu_i)^\top - u_i u_i^\top \Big) \Big\|_F^2 \\
&= \sum_{i=1}^{m} \omega_i \Big\| C_i \Big\|_F^2 \Big\| \phi(\mu_i)\phi(\mu_i)^\top - u_i u_i^\top \Big\|_F^2 \\
&\overset{(i)}{\le} \sum_{i=1}^{m} \omega_i \Big\| \phi(\mu_i)\phi(\mu_i)^\top - u_i u_i^\top \Big\|_F^2 \\
&\overset{(j)}{=} \sum_{i=1}^{m} \omega_i \Big(1 + \|u_i\|^4 - 2\big(u_i^\top \phi_T(\mu_i)\big)^2 \Big) \\
&\le \sum_{i=1}^{m} \omega_i \Big(2 - 2\big(u_i^\top \phi(\mu_i)\big)^2 \Big) \\
&= 2 \sum_{i=1}^{m} \omega_i \Big(1 + u_i^\top \phi(\mu_i) \Big) \Big(1 - u_i^\top \phi(\mu_i) \Big) \\
&\overset{(k)}{\le} 4 \sum_{i=1}^{m} \sum_{j=1}^{i-1} \omega_i \exp\Big(\frac{-\|\mu_i - \mu_j\|_2^2}{\sigma^2} \Big).
\end{aligned}$$

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 $\|uu^\top - vv^\top\|_F^2 = \|u\|_2^4 + \|v\|_2^4 - 2\langle u, v \rangle^2$. (k) holds because $u_i^\top \phi_T(\mu_i) \le 1$ and

$$u_i^\top \phi_T(\mu_i) = 1 - \sum_{j=1}^{i-1} \langle \phi_T(\mu_i), u_j \rangle^2 \ge 1 - \sum_{j=1}^{i-1} \exp\big(\frac{-\|\mu_i - \mu_j\|_2^2}{\sigma^2} \big).$$

Since we know that for every matrices $A, B \in \mathbb{R}^{d \times d}$, $\|A + B\|_F^2 \le 2\|A\|_F^2 + 2\|B\|_F^2$, the above results show that

$$\Big\| C_{X \otimes T} - \sum_{i=1}^{m} \omega_i C_i \otimes u_i u_i^\top \Big\|_F^2 \le \sum_{i=1}^{m} 4\omega_i \frac{\sigma_i^2}{\sigma^2} + \sum_{i=2}^{m} \sum_{j=1}^{i-1} 8\omega_i \exp\Big(\frac{-\|\mu_i - \mu_j\|_2^2}{\sigma^2} \Big).$$

As a result, the Hoffman-Wielandt inequality shows that for the sorted eigenvalues vector $\boldsymbol{\lambda}$ of $C_{X \otimes T}$ and sorted eigenvalues vector $\widetilde{\boldsymbol{\lambda}}$ of $\sum_{i=1}^{m} \omega_i C_i \otimes u_i u_i^\top$ the following holds:

$$\begin{aligned}
\big\| \boldsymbol{\lambda} - \widetilde{\boldsymbol{\lambda}} \big\|_2^2 \;&\le\; \Big\| C_{X \otimes T} - \sum_{i=1}^{m} \omega_i C_i \otimes u_i u_i^\top \Big\|_F^2 \\
&\le \sum_{i=1}^{m} 4\omega_i \frac{\sigma_i^2}{\sigma^2} + \sum_{i=2}^{m} \sum_{j=1}^{i-1} 8\omega_i \exp\Big(\frac{-\|\mu_i - \mu_j\|_2^2}{\sigma^2} \Big).
\end{aligned}$$

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 $\omega_i \|u_i\|_2^2$ times the eigenvalues of C_i. Also, we know that $1 \geq \|u_i\|_2^2 \geq 1 - 2\sum_{j=1}^{i-1} \exp(-\frac{\|\mu_i - \mu_j\|_2^2}{\sigma^2})$. 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 1

$$\begin{aligned}
\Big|\widetilde{H}_\alpha(X,T) - \frac{1}{1-\alpha}\log\big(\|\widehat{\lambda}_{x\otimes t}\|_\alpha^\alpha\big)\Big| &\leq g\big(\|\widetilde{\lambda}_{x\otimes t}\|_\alpha - \|\widehat{\lambda}_{x\otimes t}\|_\alpha\big) \\
&\leq g\big(\|\text{sort}\big(\widetilde{\lambda}_{x\otimes t}\big) - \text{sort}\big(\widehat{\lambda}_{x\otimes t}\big)\|_\alpha\big) \\
&\leq g\big(\|\text{sort}\big(\widetilde{\lambda}_{x\otimes t}\big) - \text{sort}\big(\widehat{\lambda}_{x\otimes t}\big)\|_2\big) \\
&\leq g\Big(\sum_{i=1}^m 4\omega_i \frac{\sigma_i^2}{\sigma^2} + \sum_{i=2}^m \sum_{j=1}^{i-1} 16\omega_i \exp\Big(\frac{-\|\mu_i - \mu_j\|_2^2}{\sigma^2}\Big)\Big).
\end{aligned}$$

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:

$$\Big|\widetilde{H}_\alpha(T) - \frac{1}{1-\alpha}\log\big(\|[\omega_1, \ldots, \omega_m]\|_\alpha^\alpha\big)\Big| \leq g\Big(\sum_{i=1}^m 4\omega_i \frac{\sigma_i^2}{\sigma^2} + \sum_{i=2}^m \sum_{j=1}^{i-1} 16\omega_i \exp\Big(\frac{-\|\mu_i - \mu_j\|_2^2}{\sigma^2}\Big)\Big).$$

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

$$\begin{aligned}
&\Big|\widetilde{H}_\alpha(X|T) - \Big(\frac{1}{1-\alpha}\log\big(\|\widehat{\lambda}_{x\otimes t}\|_\alpha^\alpha\big) - \frac{1}{1-\alpha}\log\big(\|[\omega_1, \ldots, \omega_m]\|_\alpha^\alpha\big)\Big)\Big| \\
&\leq 2g\Big(\sum_{i=1}^m 4\omega_i \frac{\sigma_i^2}{\sigma^2} + \sum_{i=2}^m \sum_{j=1}^{i-1} 16\omega_i \exp\Big(\frac{-\|\mu_i - \mu_j\|_2^2}{\sigma^2}\Big)\Big).
\end{aligned}$$

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 $H(X|G=i) = f_\alpha^{-1}(\|\lambda_{C_i}\|_\alpha^\alpha)$. As a result, we can combine the previous two equations and complete the proof as:

$$\begin{aligned}
&\Big|\widetilde{H}_\alpha(X|T) - f_\alpha^{-1}\Big(\sum_{i=1}^m \frac{\omega_i^\alpha}{\sum_{j=1}^m \omega_j^\alpha} f_\alpha\big(\widetilde{H}_\alpha(X|G=i)\big)\Big)\Big| \\
&\leq 2g\Big(\sum_{i=1}^m 4\omega_i \frac{\sigma_i^2}{\sigma^2} + \sum_{i=2}^m \sum_{j=1}^{i-1} 16\omega_i \exp\Big(\frac{-\|\mu_i - \mu_j\|_2^2}{\sigma^2}\Big)\Big).
\end{aligned}$$

D. Additional Numerical Results

D.1. Measuring Conditional-Vendi across prompt types

In this section, we conducted additional experiments similar to those in Figure 4. We created 5,000 prompts across different categories using GPT4o and generated corresponding images with text-to-image models. We reported Conditional-Vendi for the top 3 groups in the text data on PixArt-α, Stable Diffusion XL text-to-image generative models.

As shown in Figure 7, Figure 8 and Figure 9, 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.

Figure 7. Quantifying image diversity for different clusters of text prompts. Images are generated using the Stable Diffusion XL model.

D.2. Quantifying model-induced diversity via Conditional-Vendi.

In this section, we provided a more detailed version of Figure 2. As shown in Figure 10, we found that Conditional-Vendi increased at a more rapid rate when the prompts did not specify the type of animal in the picture. In contrast, when the animal types were specified in the prompts, there was only a slight increase in the Conditional-Vendi score.

D.3. Additional Numerical Evaluation of the Conditional-Vendi Score

To further experiment the correlation between the intrinsic model diversity and the defined Conditional-Vendi score, we have performed experiments of quantifying the diversity scores for unspecified and type-specified prompts when generating data from standard text-to-image models. To conduct an extensive evaluation of the Conditional-Vendi score, we performed the experiments on the nine combinations of three category types, 1) animals, 2) fruits, 3) objects, and three SOTA text-to-image generation models SDXL, Kandinsky, and PixArt-Σ Chen et al. [9]. In each of the nine experiments, we generated prompts on 10 different types related to each category and created image samples by inputting the prompts to the text-to-image model. In each experiment, we simulated 10 prompt-based generative models by considering image samples from j types for $j \in \{1, \ldots, 10\}$.

To validate the correlation between the Conditional-Vendi score and the groundtruth non-prompt-induced diversity, we evaluated the conditional-Vendi scores of the generated samples when the type is specified in the prompt (i.e. the original prompt) and when the type is unspecified by the prompt. Figures 13,15,14 show the Conditional-Vendi scores evaluated under type-specified and unspecified. In these figures, we validate that by specifying the category type in the prompt, the conditional Vendi score will have a lower value and grows slowly with the number of types expressed in the prompts. In addition, given the original prompts specifying the type of category in the image, we compared the Vendi and Conditional-Vendi scores among the three text-to-image models. Figures ??? show the comparison between the scores of the three models, which suggest the higher intrinsic diversity measure by Conditional-Vendi for the SD-XL and PixArt-Σ models.

Figure 8. Quantifying image diversity for different clusters of text prompts. Images are generated using the PixArt-α model.

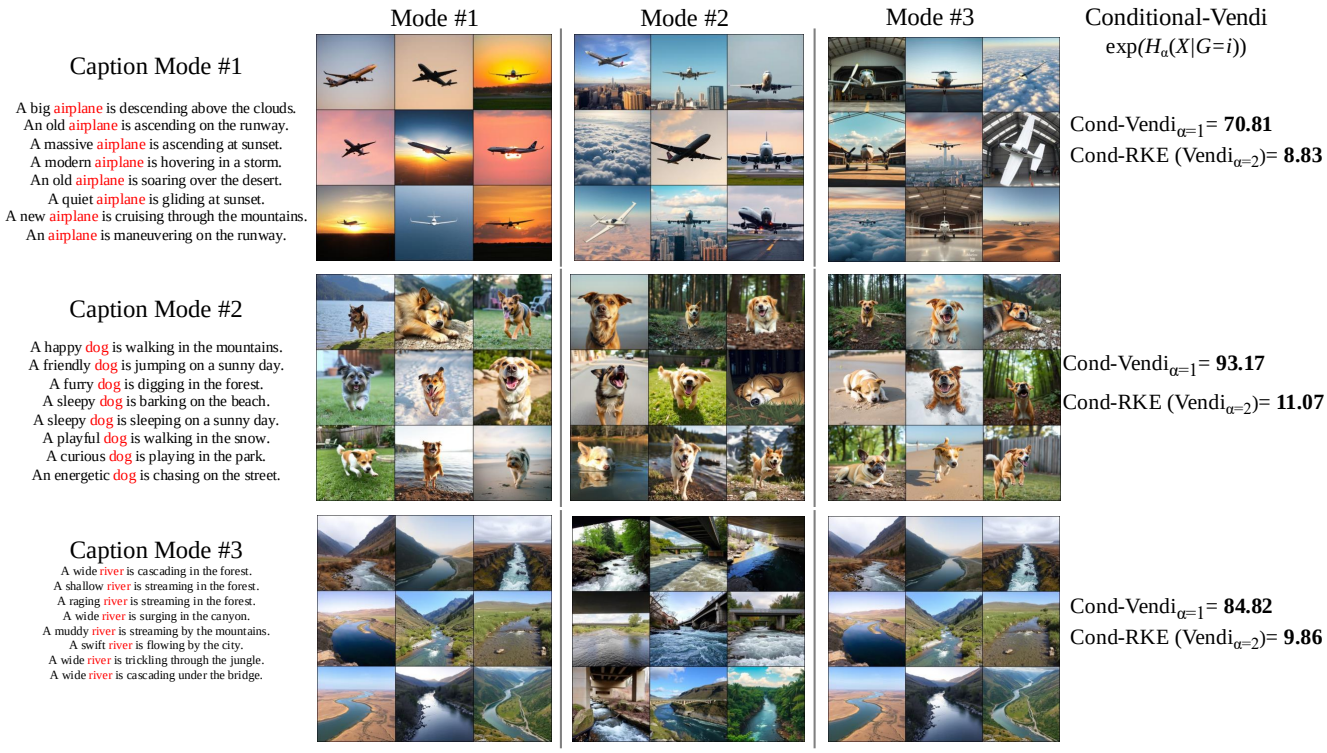
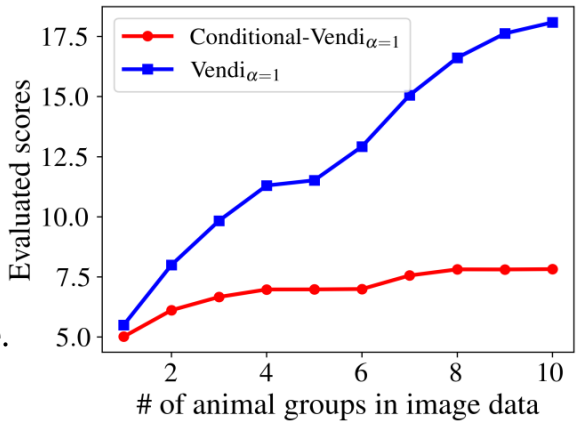


Figure 9. Quantifying image diversity for different clusters of text prompts. Images are generated using the Flux model.

Type-specified animal prompts

Prompts:

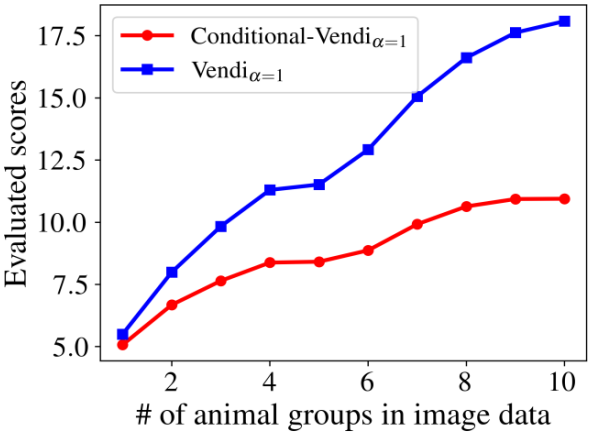
A **fox** is rolling in the grass.
A **camel** is resting near a sand dune.
A **wolf** is drinking from a river.
A **cow** is resting under a tree.
A **sheep** is resting in a grassy pasture.
An **elephant** is walking through thick jungle.
A **giraffe** is reaching up for leaves.



Unspecified animal prompts

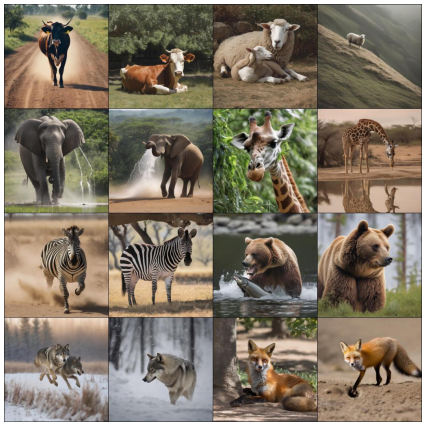
Prompts:

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.



Model 2:
Samples from 2 animal groups

Model 4:
Samples from 4 animal groups



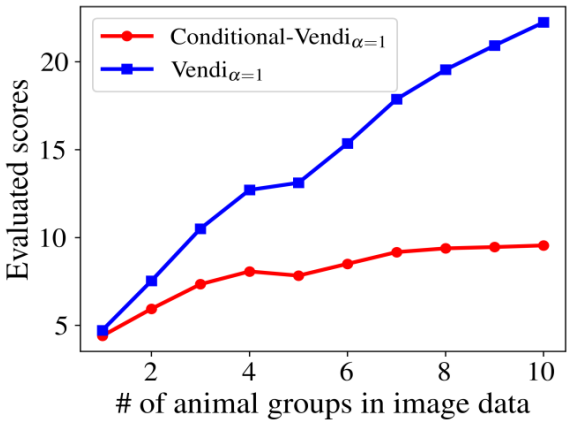
Model 8:
Samples from 8 animal groups

Figure 10. Comparing Conditional-Vendi with Vendi on different animal groups generated by Stable Diffusion-XL.

Type-specified animal prompts

Prompts:

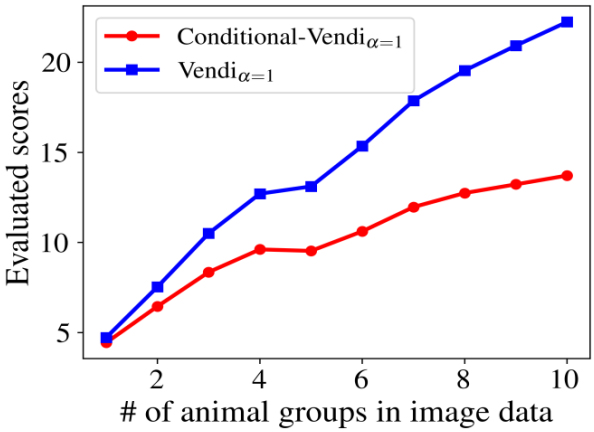
A **fox** is rolling in the grass.
A **camel** is resting near a sand dune.
A **wolf** is drinking from a river.
A **cow** is resting under a tree.
A **sheep** is resting in a grassy pasture.
An **elephant** is walking through thick jungle
A **giraffe** is reaching up for leaves.



Unspecified animal prompts

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.



Model 2:
Samples from 2 animal groups

Model 4:
Samples from 4 animal groups

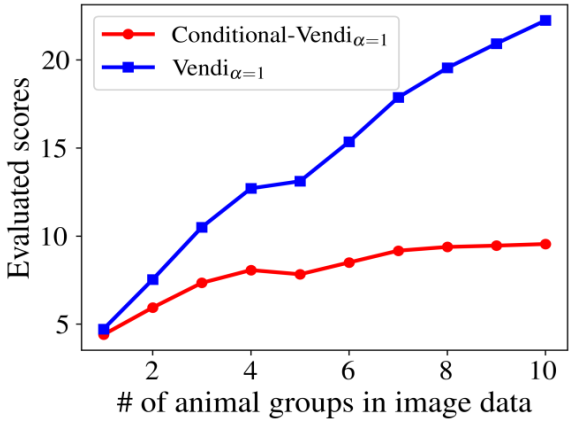
Model 8:
Samples from 8 animal groups

Figure 11. Comparing Conditional-Vendi with Vendi on different animal groups generated by PixArtΣ.

Type-specified animal prompts

Prompts:

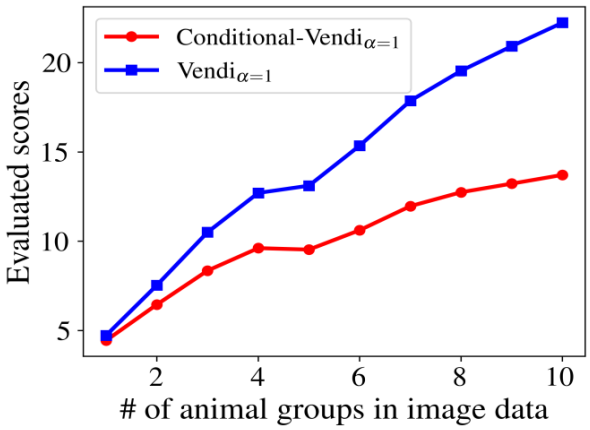
A **fox** is rolling in the grass.
A **camel** is resting near a sand dune.
A **wolf** is drinking from a river.
A **cow** is resting under a tree.
A **sheep** is resting in a grassy pasture.
An **elephant** is walking through thick jungle
A **giraffe** is reaching up for leaves.



Unspecified animal prompts

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.



Model 2:
Samples from 2 animal groups

Model 4:
Samples from 4 animal groups

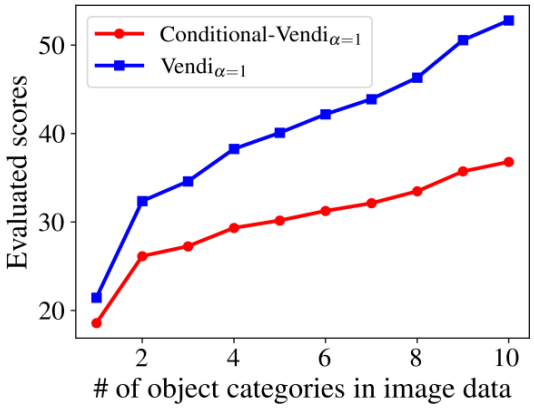
Model 8:
Samples from 8 animal groups

Figure 12. Comparing Conditional-Vendi with Vendi on different animal groups generated by Kandinsky.

Type-specified object prompts

Prompts:

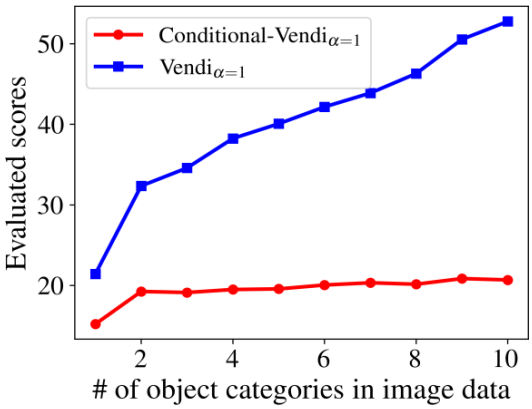
A **Chair** is placed under a sprawling tree.
A **Sofa** is glowing in the light of a nearby fire.
A **Book** is balancing on the edge of a table.
A **Clock** is positioned in an office setup.
A **Lamp** is sitting under a hanging light bulb.
A **laptop** is half-hidden behind a stack of boxes.
A **car** is in the corner of a large warehouse.
The **cup** is precariously balanced on rocks.

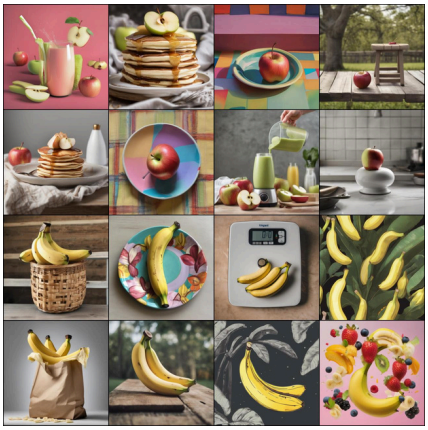


Unspecified object prompts

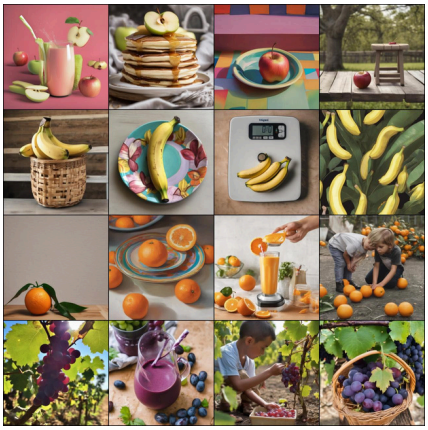
Prompts:

An **object** is placed under a sprawling tree.
An **object** is glowing in the light of a nearby fire.
An **object** is balancing on the edge of a table.
An **object** is positioned in an office setup.
An **object** is sitting under a hanging light bulb.
An **object** is half-hidden behind a stack of boxes.
An **object** is in the corner of a large warehouse.
The **object** is precariously balanced on rocks.





Model 2:
Samples from 2 fruit types



Model 4:
Samples from 4 fruit types

Model 8:
Samples from 8 fruit types

Figure 13. Comparing Conditional-Vendi with Vendi on different fruit types generated by Stable Diffusion-XL.

D.4. 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:

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

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 1, is expected to correlate with our defined Conditional-Vendi score.

In the numerical settings of the previous section, where we know the ground-truth clusters based on the type of animal, fruit, or object in the texts, we computed the value of the GroundTruth-Cluster-Vendi score and compared it with the evaluated Conditional-Vendi score. As demonstrated in Figures **??**, the two diversity scores, Conditional-Vendi and GroundTruth-Cluster-Vendi, highly correlate for the ten 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. Our numerical results with the K-Means clustering algorithm in Figure 19 also demonstrated these clustering challenges and, in several cases, failed to find the ground-truth clusters with high accuracy.

D.5. Algorithm for Computing Conditional-Vendi and Information-Vendi

In this section, we present the algorithm to compute the Conditional-Vendi and Information-Vendi scores. Using the definition provided in Section 4, combined with the entropy definition in (3), we calculate the Conditional-Vendi score. The steps are outlined in Algorithm 1.

D.6. Qualitative results for generative models trained on MS-COCO dataset

In this section, we provide images and prompts corresponding to Figure 3. Figure 21 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.

D.7. Effect of Bandwidth on Conditional-Vendi and Information-Vendi

To further investigate the effect of bandwidth on Conditional-Vendi and Information-Vendi, we began by selecting the image bandwidth similar to prior works Friedman and Dieng [11], Ospanov et al. [31]. We then measured and plotted the scores using varying text kernel bandwidths. Figure 22 demonstrates consistent rankings of the four models across different bandwidth parameters. The results indicate that as the kernel bandwidth increases, the number of text clusters increases, leading to a decrease in the Information-Vendi value.

Algorithm 1 Conditional-Vendi and Information-Vendi

1: **Input:** Sample sets $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and $\{\mathbf{t}_1, \ldots, \mathbf{t}_n\}$, Gaussian kernel bandwidths σ_i^2, σ_t^2, order α.

2: **Compute kernel matrices**: $K_{\mathbf{X}} = \frac{1}{n}[k(\mathbf{x}_i, \mathbf{x}_j)]_{n\times n}$, $K_{\mathbf{T}} = \frac{1}{n}[k(\mathbf{t}_i, \mathbf{t}_j)]_{n\times n}$

3: **Perform eigendecomposition** on the $K_{\mathbf{X}}$, $K_{\mathbf{T}}$ and $\frac{1}{n}K_{\mathbf{X}} \odot K_{\mathbf{T}}$ matrices:

$$\{\lambda_1^{\mathbf{X}}, \ldots, \lambda_n^{\mathbf{X}}\} \leftarrow \text{Eigendecomposition}(K_{\mathbf{X}})$$

$$\{\lambda_1^{\mathbf{T}}, \ldots, \lambda_n^{\mathbf{T}}\} \leftarrow \text{Eigendecomposition}(K_{\mathbf{T}})$$

$$\{\lambda_1^{\mathbf{X},\mathbf{T}}, \ldots, \lambda_n^{\mathbf{X},\mathbf{T}}\} \leftarrow \text{Eigendecomposition}(\frac{1}{n}K_{\mathbf{X}} \odot K_{\mathbf{T}})$$

4: **Compute** $H_\alpha(\frac{1}{n}K_{\mathbf{X}})$**,** $H_\alpha(\frac{1}{n}K_{\mathbf{T}})$ **and** $H_\alpha(\frac{1}{n}K_{\mathbf{X}} \odot K_{\mathbf{T}})$ using their eigenvalues.

$$H_\alpha\big(\frac{1}{n}K_{\mathbf{X}}\big) \leftarrow \frac{1}{1-\alpha}\log\Big(\sum_{i=1}^{n}(\lambda_i^{\mathbf{x}})^\alpha\Big)$$

$$H_\alpha\big(\frac{1}{n}K_{\mathbf{T}}\big) \leftarrow \frac{1}{1-\alpha}\log\Big(\sum_{i=1}^{n}(\lambda_i^{\mathbf{T}})^\alpha\Big)$$

$$H_\alpha\big(\frac{1}{n}K_{\mathbf{X}} \odot K_{\mathbf{T}}\big) \leftarrow \frac{1}{1-\alpha}\log\Big(\sum_{i=1}^{n}(\lambda_i^{\mathbf{X},\mathbf{T}})^\alpha\Big)$$

5: **Compute Conditional-Vendi and Information-Vendi**

$$\text{Conditional-Vendi}_\alpha\big(x_1, \ldots, x_n \big| t_1, \ldots, t_n\big) \leftarrow \exp\Big(H_\alpha\big(\frac{1}{n}K_{\mathbf{X}} \odot K_{\mathbf{T}}\big) - H_\alpha\big(\frac{1}{n}K_{\mathbf{T}}\big)\Big)$$

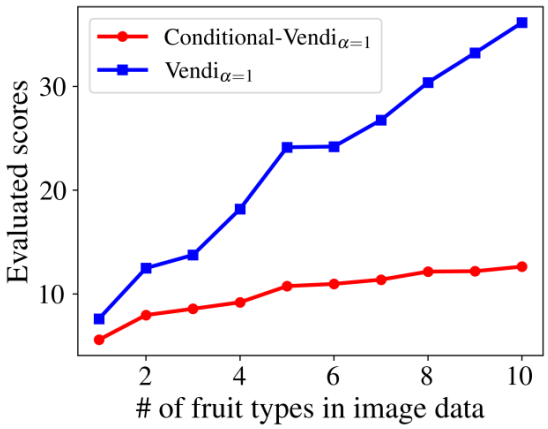
$$\text{Information-Vendi}_\alpha\big(x_1, .., x_n; t_1, .., t_n\big) \leftarrow \exp\Big(H_\alpha\big(\frac{1}{n}K_{\mathbf{X}}\big) + H_\alpha\big(\frac{1}{n}K_{\mathbf{T}}\big) - H_\alpha\big(\frac{1}{n}K_{\mathbf{X}} \odot K_{\mathbf{T}}\big)\Big)$$

6: **Output:** Conditional-Vendi and Information-Vendi

Type-specified fruit prompts

Prompts:

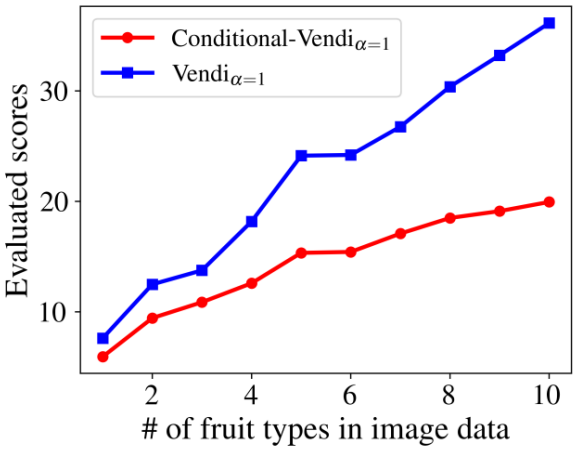
An **apple** is next to a cold glass of fresh juice.
A **banana** is being sliced with a sharp knife.
The **watermelon** is blended into a smoothie.
The **pineapple** is falling out of a grocery bag.
A **strawberry** is being washed.
The **peach** is sitting on a kitchen countertop.
A **cherry** is being sliced with a knife.
A **mango** is sitting on a colorful plate.



Unspecified fruit prompts

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.
A **fruit** is being sliced with a knife.
A **fruit** is sitting on a colorful plate.



Model 2:
Samples from 2 fruit types

Model 4:
Samples from 4 fruit types

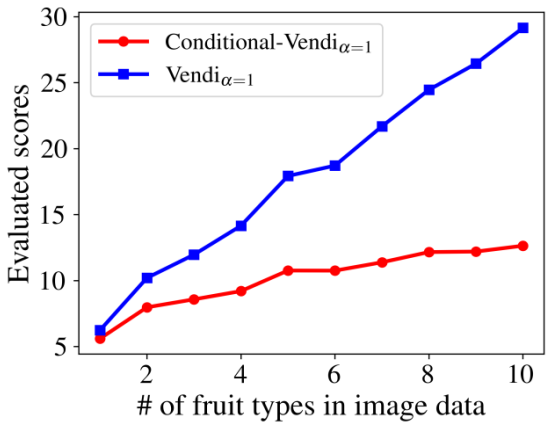
Model 8:
Samples from 8 fruit types

Figure 14. Comparing Conditional-Vendi with Vendi on different fruit types generated by Kandinsky.

Type-specified fruit prompts

Prompts:

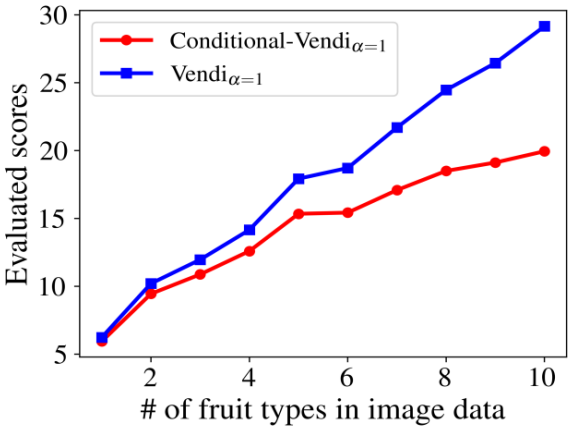
An **apple** is next to a cold glass of fresh juice.
A **banana** is being sliced with a sharp knife.
The **watermelon** is blended into a smoothie.
The **pineapple** is falling out of a grocery bag.
A **strawberry** is being washed.
The **peach** is sitting on a kitchen countertop.
A **cherry** is being sliced with a knife.
A **mango** is sitting on a colorful plate.



Unspecified fruit prompts

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.
A **fruit** is being sliced with a knife.
A **fruit** is sitting on a colorful plate.



Model 2:
Samples from 2 fruit types

Model 4:
Samples from 4 fruit types

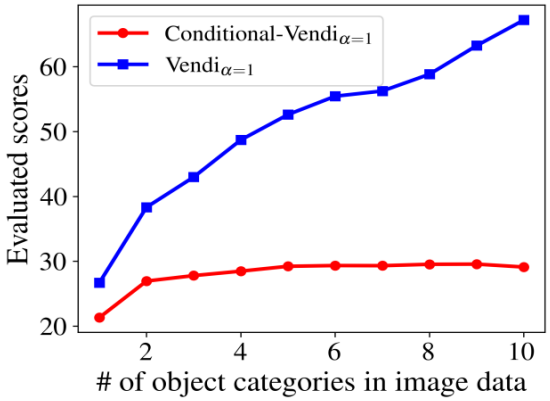
Model 8:
Samples from 8 fruit types

Figure 15. Comparing Conditional-Vendi with Vendi on different fruit types generated by PixArt-Σ.

Type-specified object prompts

Prompts:

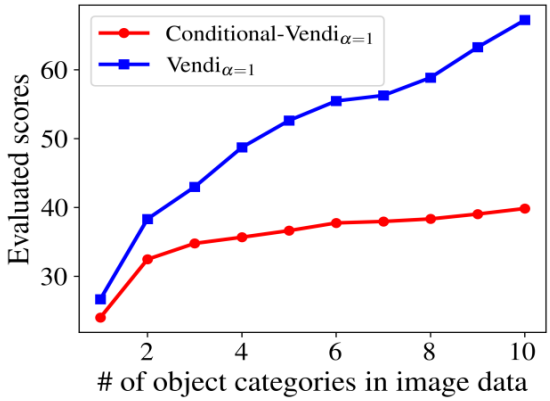
A **Chair** is placed under a sprawling tree.
A **Sofa** is glowing in the light of a nearby fire.
A **Book** is balancing on the edge of a table.
A **Clock** is positioned in an office setup.
A **Lamp** is sitting under a hanging light bulb.
A **laptop** is half-hidden behind a stack of boxes.
A **car** is in the corner of a large warehouse.
The **cup** is precariously balanced on rocks.

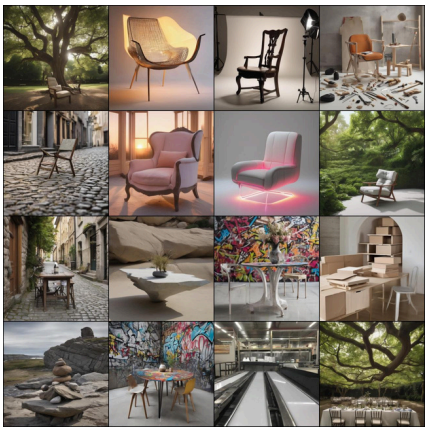


Unspecified object prompts

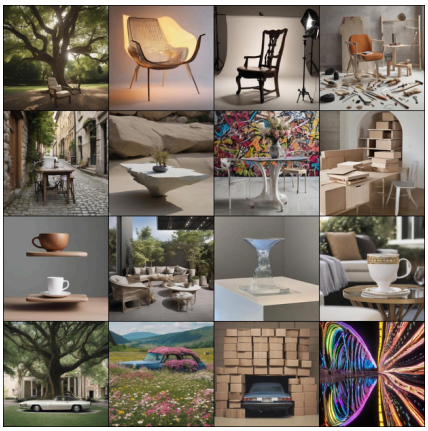
Prompts:

An **object** is placed under a sprawling tree.
An **object** is glowing in the light of a nearby fire.
An **object** is balancing on the edge of a table.
An **object** is positioned in an office setup.
An **object** is sitting under a hanging light bulb.
An **object** is half-hidden behind a stack of boxes.
An **object** is in the corner of a large warehouse.
The **object** is precariously balanced on rocks.

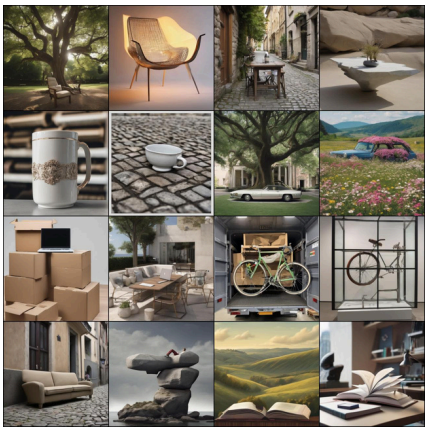




Model 2:
Samples from 2 object categories



Model 4:
Samples from 4 object categories



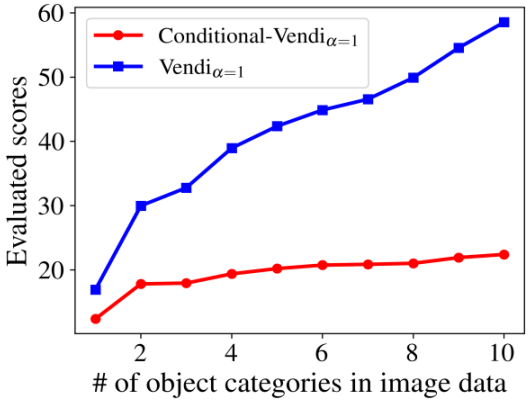
Model 8:
Samples from 8 object categories

Figure 16. Comparing Conditional-Vendi with Vendi on different fruit types generated by Stable Diffusion-XL.

Type-specified object prompts

Prompts:

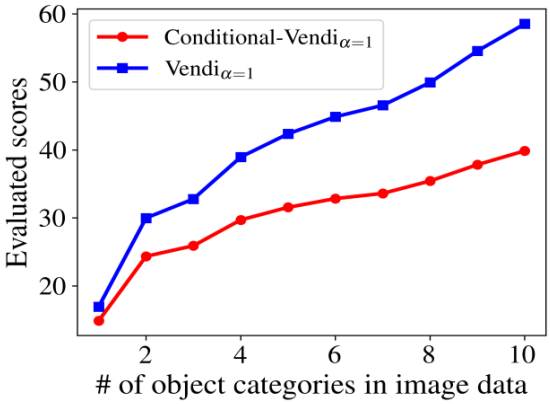
A **Chair** is placed under a sprawling tree.
A **Sofa** is glowing in the light of a nearby fire.
A **Book** is balancing on the edge of a table.
A **Clock** is positioned in an office setup.
A **Lamp** is sitting under a hanging light bulb.
A **laptop** is half-hidden behind a stack of boxes.
A **car** is in the corner of a large warehouse.
The **cup** is precariously balanced on rocks.

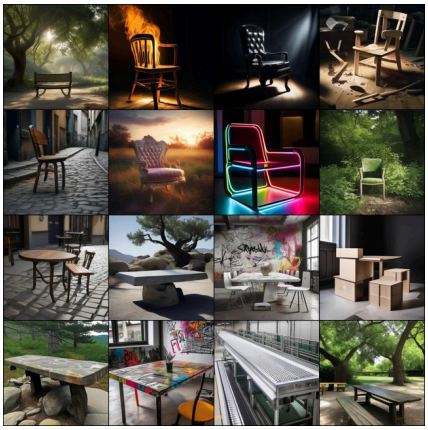


Unspecified object prompts

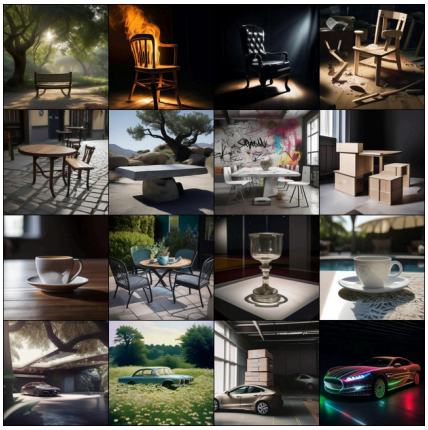
Prompts:

An **object** is placed under a sprawling tree.
An **object** is glowing in the light of a nearby fire.
An **object** is balancing on the edge of a table.
An **object** is positioned in an office setup.
An **object** is sitting under a hanging light bulb.
An **object** is half-hidden behind a stack of boxes.
An **object** is in the corner of a large warehouse.
The **object** is precariously balanced on rocks.

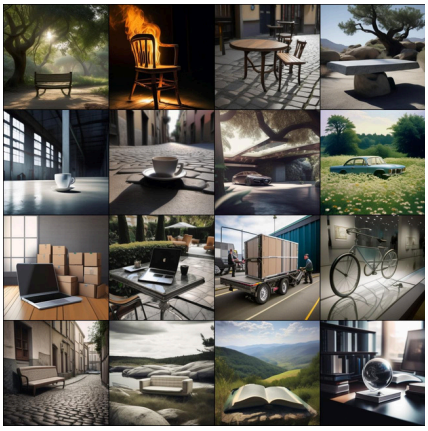




Model 2:
Samples from 2 object categories



Model 4:
Samples from 4 object categories



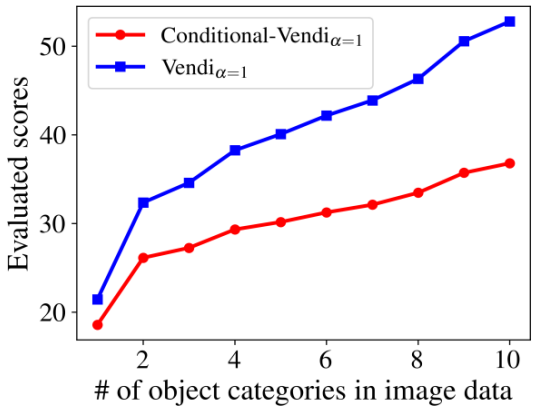
Model 8:
Samples from 8 object categories

Figure 17. Comparing Conditional-Vendi with Vendi on different object categories types generated by Kandinsky.

Type-specified object prompts

Prompts:

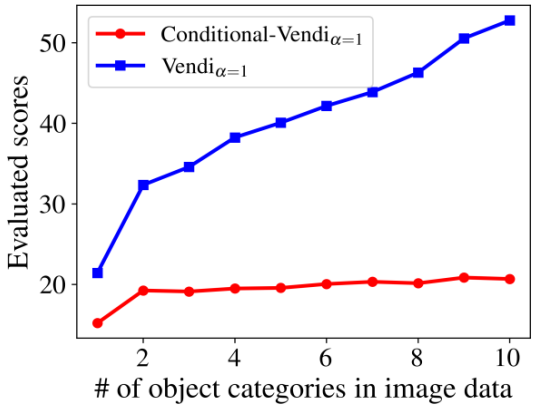
A **Chair** is placed under a sprawling tree.
A **Sofa** is glowing in the light of a nearby fire.
A **Book** is balancing on the edge of a table.
A **Clock** is positioned in an office setup.
A **Lamp** is sitting under a hanging light bulb.
A **laptop** is half-hidden behind a stack of boxes.
A **car** is in the corner of a large warehouse.
The **cup** is precariously balanced on rocks.

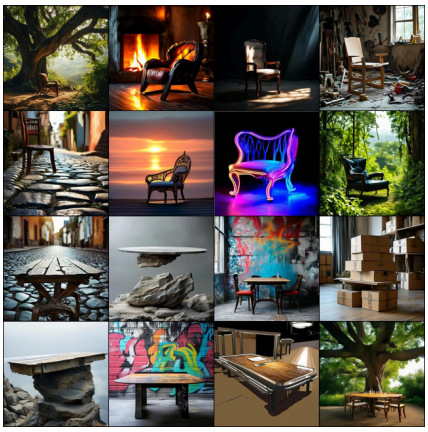


Unspecified object prompts

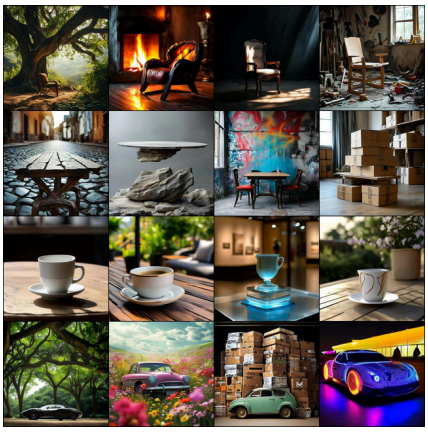
Prompts:

An **object** is placed under a sprawling tree.
An **object** is glowing in the light of a nearby fire.
An **object** is balancing on the edge of a table.
An **object** is positioned in an office setup.
An **object** is sitting under a hanging light bulb.
An **object** is half-hidden behind a stack of boxes.
An **object** is in the corner of a large warehouse.
The **object** is precariously balanced on rocks.





Model 2:
Samples from 2 object categories



Model 4:
Samples from 4 object categories

Model 8:
Samples from 8 object categories

Figure 18. Comparing Conditional-Vendi with Vendi on different object categories types generated by PixArt-Σ.

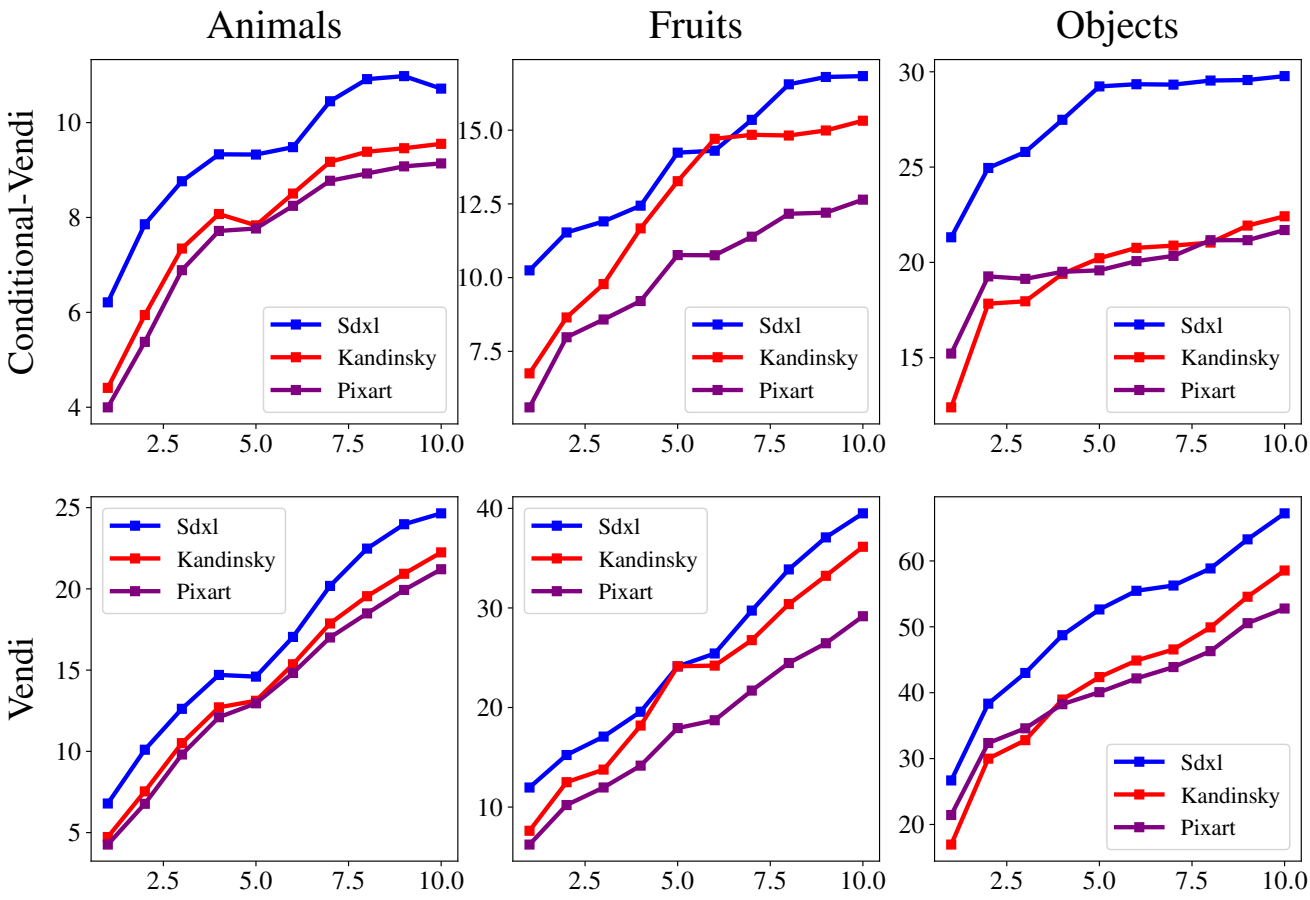


Figure 19. Evaluated (unconditional) Vendi and Conditional-Vendi scores of three text-to-image models in the category-based experiments with varying number of types within each of the categories: Animals, Fruits, Objects.

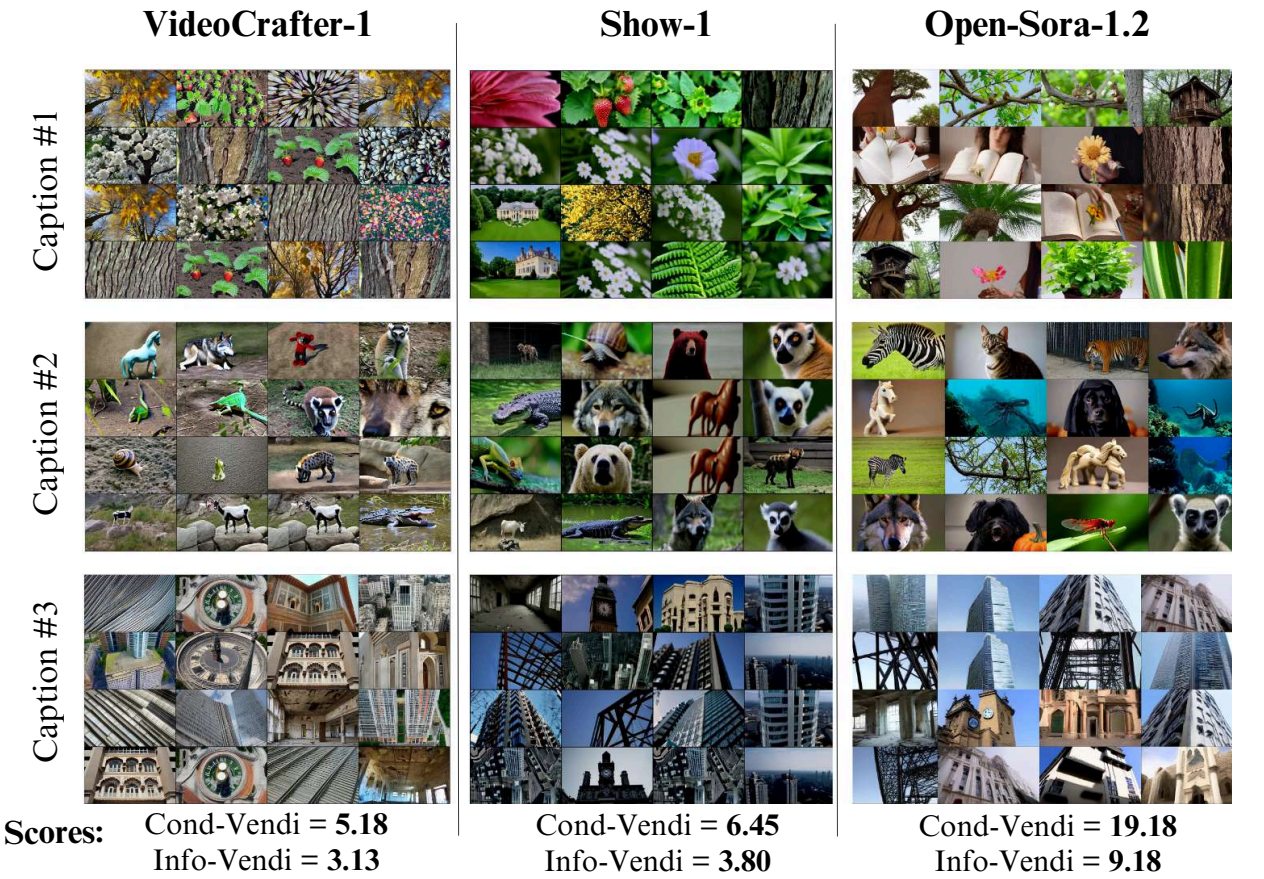
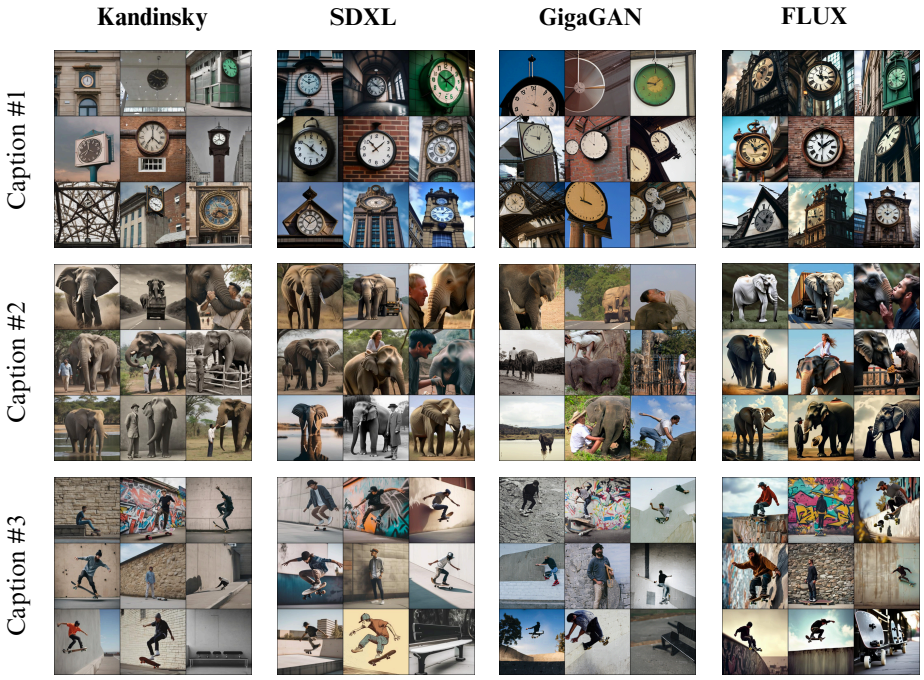


Figure 20. Measuring Conditional-Vendi and Information-Vendi for text-to-video models



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 21. Effect of text kernel bandwidth on Conditional-Vendi and Information-Vendi scores

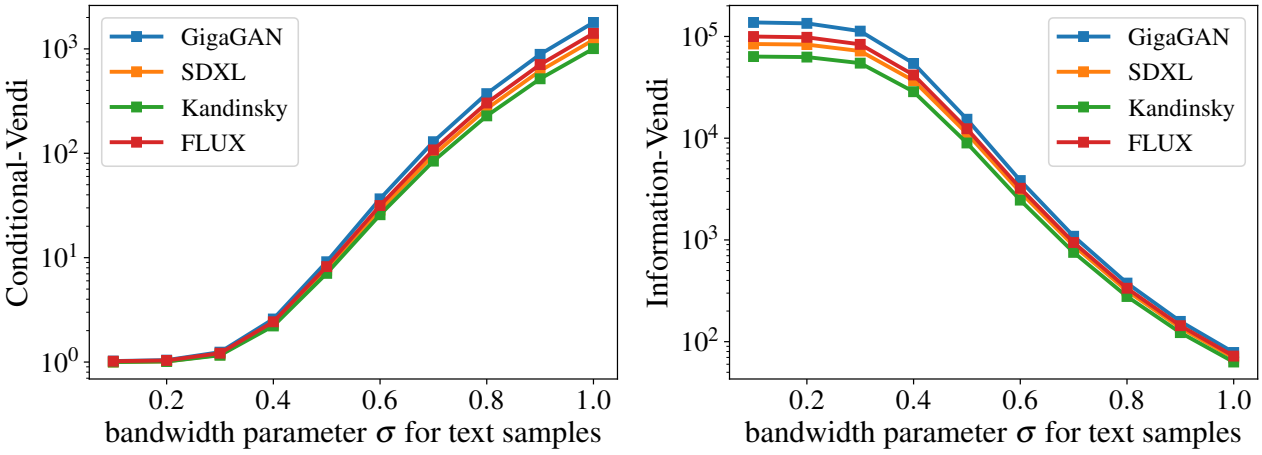


Figure 22. Effect of text kernel bandwidth on Conditional-Vendi and Information-Vendi scores