

Scendi Score: Prompt-Aware Diversity Evaluation via Schur Complement of CLIP Embeddings

Azim Ospanov*

aospanov9@cse.cuhk.edu.hk

Mohammad Jalali*

mjalali24@cse.cuhk.edu.hk

Farzan Farnia*

farnia@cse.cuhk.edu.hk

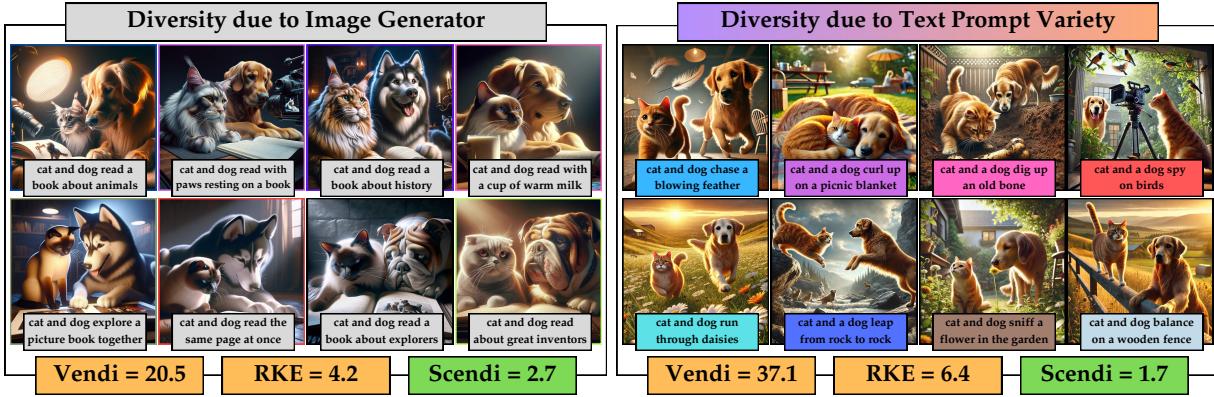


Figure 1. Comparison of model-driven diversity (left) and prompt-driven diversity (right) in 200 images generated for similar "cat and dog" prompts (left) vs. 200 images generated for diverse "cat and dog" prompts (right). The diversity metrics for unconditional generation, Vendi and RKE, favor the right-side case, but the Scendi score captures intrinsic model diversity, suggesting the left side has higher diversity.

Abstract

The use of CLIP embeddings to assess the fidelity of samples produced by text-to-image generative models has been extensively explored in the literature. While the widely adopted CLIPScore, derived from the cosine similarity of text and image embeddings, effectively measures the alignment of a generated image, it does not quantify the diversity of images generated by a text-to-image model. In this work, we extend the application of CLIP embeddings to quantify and interpret the intrinsic diversity of text-to-image models, which are responsible for generating diverse images from similar text prompts, which we refer to as prompt-aware diversity. To achieve this, we propose a decomposition of the CLIP-based kernel covariance matrix of image data into text-based and non-text-based components. Using the Schur complement of the joint image-text kernel covariance matrix, we perform this

decomposition and define the matrix-based entropy of the decomposed component as the Schur Complement ENtropy DIversity (Scendi) score, as a measure of the prompt-aware diversity for prompt-guided generative models. Additionally, we discuss the application of the Schur complement-based decomposition to nullify the influence of a given prompt on the CLIP embedding of an image, enabling focus or defocus of the embedded vectors on specific objects. We present several numerical results that apply our proposed Scendi score to evaluate text-to-image and LLM (text-to-text) models. Our numerical results indicate the success of the Scendi score in capturing the intrinsic diversity of prompt-guided generative models. The codebase is available at <https://github.com/aziksh-ospanov/scendi-score>.

1. Introduction

Prompt-guided generative models, which generate data guided by an input text prompt, have gained significant

*The Chinese University of Hong Kong, Department of Computer Science & Engineering

attention in the computer vision community. In particular, text-to-image and text-to-video models, which create visual data based on input text, have found many applications and are widely used across various content creation tasks. Given the important role of prompt-guided generative AI models in numerous machine learning applications, their training and evaluation have been extensively studied in recent years. A comprehensive evaluation of these models, addressing fidelity and diversity, is essential to ensure their effectiveness and adaptability across different use cases.

As the CLIP model [45] provides a joint representation for text and image data, the CLIP embeddings have been widely used to evaluate and interpret the performance of text-to-image models. By calculating the cosine similarity between the CLIP embeddings of text and image data, the CLIPScore [14] serves as a fidelity metric for measuring the alignment between the text and the generated image. While the CLIPScore and similar uses of CLIP embeddings focus on evaluating the quality of generated samples, these embeddings have not yet been applied to assess the diversity of data produced by text-to-image models. Existing evaluation frameworks typically address diversity scores for unconditional sample generation (considering no input prompt), such as Recall [28, 50], Coverage [39], RKE [20], and Vendi [9], which assess image samples independently of the input text.

However, the diversity of images generated by a text-to-image model depends both on the variety of input text data and on the model’s intrinsic diversity, driving the model to produce varied images in response to similar prompts. Therefore, existing diversity metrics for unconditional sample generation cannot differentiate between diversity arising from varied prompts, i.e., *prompt-driven diversity*, and diversity contributed by the model itself, i.e., *model-driven diversity*. In this work, we leverage CLIP embeddings to propose a framework for quantifying and interpreting the intrinsic diversity of both text-to-image and image captioning (image-to-text) models. The primary objective of our approach is to decompose the CLIP image embedding into a text-based component, influenced by the input text, and a non-text-based component, arising from the model’s inherent randomness.

To achieve this goal, we extend the kernel matrix entropy measures, i.e., Vendi and RKE scores, to enable a prompt aware diversity measurement. To do this, we consider a kernel similarity function and focus on the kernel covariance matrix for the generated (text, image) pairs. Considering the matrix-based entropy of the Image-based kernel covariance component C_{II} , one obtains the existing prompt-unaware Vendi [9] and RKE [20] scores. In this work, we propose applying the Gaussian elimination approach, and then decompose the image sub-covariance

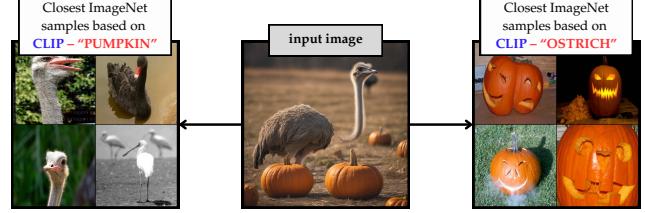


Figure 2. Example of the CLIP embedding decomposition for an input image of "Pumpkin next to Ostrich" generated by SDXL [44]. ImageNet [7] samples with the highest CLIP similarity to the input image after removing the "pumpkin" direction from its CLIP embedding via the Schur complement (left) and after removing the "ostrich" direction (right).

matrix, C_{II} as follows:

$$C_{II} = \underbrace{C_{II} - C_{IT}C_{TT}^{-1}C_{IT}^\top}_{\text{Model-driven Covariance Component}} + \underbrace{C_{IT}C_{TT}^{-1}C_{IT}^\top}_{\text{Prompt-driven Covariance Component}}$$

Note that the above is the sum of the *Schur Complement* component, $C_{II} - C_{IT}C_{TT}^{-1}C_{IT}^\top$, which represents the model-driven diversity component, and the remainder term $C_{IT}C_{TT}^{-1}C_{IT}^\top$, which captures the prompt-driven variety in the generated images. Extending the Vendi and RKE score to prompt-aware diversity evaluation, we define the matrix-based entropy of the Schur complement component, as *Schur Complement ENtropy DIversity (Scendi)* score. We highlight that Scendi is a measure of model-driven diversity in the prompt-guided generated data.

To illustrate the effectiveness of the Scendi score in distinguishing model-driven diversity from prompt-driven effects, we present a comparative example. Figure 1 compares the Scendi diversity scores for two sets of samples generated by DALL-E 3 [40]. The left-side subfigure is designed to represent model-driven diversity, showing images of diverse-breed cats and dogs generated using similar prompts like "cat and dog read a text." In contrast, the right-side subfigure highlights the effects of prompt-driven diversity, featuring same-breed cats and dogs across diverse-occasion prompts. The unconditional diversity metrics, Vendi and RKE, report higher diversity for the right-side case. In contrast, our proposed Scendi score, designed to isolate intrinsic model diversity from prompt-driven effects, indicates greater diversity in the left-side case. This example highlights the advantage of the prompt-aware diversity measurement by Scendi score over existing unconditional diversity metrics in capturing the model-driven diversity.

We present several numerical applications of the proposed framework to evaluate and interpret standard text-to-image and image-captioning models. Our results on several simulated scenarios with known ground-truth diversity indicate that the proposed entropy metric correlates with the non-text-based diversity in images,

capturing variation not attributable to the text prompt. Additionally, we show that the decomposed feature component can neutralize the influence of specific objects in the text prompt within the image embedding. Specifically, we use this decomposed feature to diminish the impact of visible text in images, reducing its effect on the embedding. We also demonstrate how the decomposed image embedding can enhance or reduce the focus on particular objects or styles within an image, which can have implications for downstream applications of CLIP embeddings when emphasizing specific elements. For instance, Figure 2 shows a generated image of an ostrich next to a pumpkin. We showcase the ImageNet samples that achieve the highest CLIPScore when compared to the modified image embeddings obtained by canceling the “pumpkin” and “ostrich” directions. This demonstrates that the modified embeddings effectively eliminate the influence of those unwanted objects. The main contributions of this work are summarized as follows:

- Proposing a Schur Complement-based approach to text-to-image diversity evaluation that decomposes the diversity metric into prompt-induced and model-induced components
- Providing a decomposition method that allows to remove directions from the CLIP embedding based on the Schur Complement modified image embedding
- Presenting numerical results on the Schur complement-based decomposition of CLIP embeddings and Scendi score performance under various diversity scenarios and data modalities.

2. Related Work

CLIP interpretability and decomposition. Contrastive vision-and-language models, such as CLIP [45] are a class of models that were trained on paired text prompts and images. The notable feature of CLIP is a shared embedding space between image and text data. A common interpretability method involves heatmaps to highlight relevant image areas [5, 11, 52, 53]. However, heatmaps are used to identify objects and lack spatially dependent information, such as object size and embedding output. Other approaches require decomposing the model architecture and analyzing attention heads. [10] introduced *TextSpan*, which finds a vector direction for each attention head and assigns it an appropriate text label. Another prominent approach, proposed in [2], decomposes dense CLIP embeddings into sparse, interpretable semantic concepts to enhance embedding description. [37] suggests disentangling written words from underlying concepts via orthogonal projection of learned CLIP representations. [4] utilize semantic guidance to move between concepts and introduce edits during diffusion process. Recently, [32] studied the geometry of the CLIP embeddings and identified

numerous properties such as modality mass of image and text components in the embedding space.

Evaluation of Diversity in Generative Models. Diversity evaluation in generative models has been extensively explored in the literature, with existing metrics broadly classified as either reference-based or reference-free. Reference-based metrics, such as FID [15], KID [3], and IS [51], assess diversity by quantifying the distance between true and generated data. Metrics like Density and Coverage [39], and Precision and Recall [28] evaluate quality and diversity by analyzing the data manifold. Reference-free metrics, including Vendi [9, 43], RKE [20], and their improved and convergent variants FKEA-Vendi [42] and Truncated Vendi [41], measure the entropy of a similarity metric (e.g., kernel matrix), capturing the number of distinct modes in the data. Since these metrics do not require a reference distribution, they are particularly suitable for text-to-image model evaluation, where selecting an appropriate reference dataset is challenging. Another concurrent work on prompt-aware diversity is the Conditional Vendi score that measures the entropy of the Hadamard kernel product subtracted by the text component. [21]. Also, the kernel-based evaluation scores have been further utilized in the context of online model selection [16–18, 49], novelty evaluation and detection [55, 56], distributed evaluation [54], and embedding comparison and alignment [12, 23].

Evaluation of Text-to-Image Generative Models. Conditional generation models have been extensively studied, with CLIPScore [14] and its variations, such as Heterogeneous CLIPScore [26], being the standard metrics for assessing prompt-image alignment using cosine similarity. Another key approach is the FID framework adapted for conditional models, which measures the joint distribution distance between prompts and images, known as FJD [8]. Holistic evaluation methods, such as the HEIM [30] and HELM [34] benchmarks, unify different aspects of generated data to provide comprehensive assessments. Diversity evaluation typically involves generating multiple images per prompt and measuring their inter-diversity [1, 25]. Existing metrics require redundant image generation for diversity measurement, whereas our proposed Schur Complement-based decomposition bypasses this need, enabling evaluation on pre-generated datasets without requiring multiple images per prompt.

3. Preliminaries

Kernel Functions. We call $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ a kernel function if for samples $x_1, \dots, x_n \in \mathcal{X}$, the resultant kernel matrix $K = [k(x_i, x_j)]_{1 \leq i, j \leq n}$ is a PSD (positive semi-definite) matrix. Moreover, K can be decomposed into a dot product of the kernel feature maps $\phi : \mathcal{X} \rightarrow \mathbb{R}^d$ as follows:

$$k(x, x') = \langle \phi(x), \phi(x') \rangle \quad (1)$$

In this work, we provide several numerical results for the cosine-similarity Kernel function defined as:

$$k_{\text{Cosine-Similarity}}(x, x') := \frac{\langle x, x' \rangle}{\|x\|_2 \|x'\|_2} \quad (2)$$

Also, we consider the Gaussian Kernel defined as:

$$k_{\text{Gaussian}(\sigma)}(x, x') := \exp\left(-\frac{\|x - x'\|_2^2}{2\sigma^2}\right) \quad (3)$$

Note that both of the presented kernels are normalized kernels, since $k(x, x) = 1$ holds for every $x \in \mathcal{X}$. To apply entropy measures, we consider the normalized kernel matrix given by $\frac{1}{n}K$. Therefore, we observe $\text{Tr}(\frac{1}{n}K) = 1$ for every normalized kernel, implying that the eigenvalues of $\frac{1}{n}K$ form a probability model, since they are non-negative and sum up to 1. Note that $\text{Tr}(\cdot)$ denotes the matrix trace.

Matrix-based Entropy and Kernel Covariance Matrix.

Given a PSD matrix A with unit trace and eigenvalues $\lambda_1, \dots, \lambda_n \geq 0$, the Von-Neumann entropy is defined as:

$$H(A) := \sum_{i=1}^n \lambda_i \log \frac{1}{\lambda_i}. \quad (4)$$

[9] discusses that the Von-Neumann entropy of the normalized kernel matrix $\frac{1}{n}K$ can effectively capture the entropy of the cluster variable in the collected data, and proposes the Von Neumann entropy diversity (Vendi) score for measuring the diversity of a sampleset. Observe that $\frac{1}{n}K$ shares the same non-zero eigenvalues with the kernel covariance matrix C_X defined as:

$$C_X := \frac{1}{n} \sum_{i=1}^n \phi(x_i) \phi(x_i)^\top = \frac{1}{n} \Phi^\top \Phi \quad (5)$$

where $\Phi \in \mathbb{R}^{n \times d}$ is an $n \times d$ matrix whose rows are the feature presentations of samples $\phi(x_1), \dots, \phi(x_n)$. Therefore, given the eigenvalues $\lambda_1, \dots, \lambda_d$ of the kernel covariance matrix C_X , the Vendi and RKE [20] scores are

$$\begin{aligned} \text{Vendi}(x_1, \dots, x_n) &= \exp\left(\sum_{i=1}^d \lambda_i \log \frac{1}{\lambda_i}\right), \\ \text{RKE}(x_1, \dots, x_n) &= \frac{1}{\sum_{i=1}^d \lambda_i^2} = \frac{1}{\|C_X\|_F^2}. \end{aligned}$$

Feature representation varies between kernel methods. In Cosine Similarity Kernel, $\phi(x) = \text{CLIP}(x)/\|\text{CLIP}(x)\|_2$, i.e. normalized CLIP embedding of sample x , whereas in shift-invariant kernels, ϕ is a proxy feature map of the Gaussian kernel following the random Fourier features [46].

Schur Complement. Consider a block matrix Λ with a symmetric matrix partition as follows:

$$\Lambda = \begin{bmatrix} B & C \\ C^\top & D \end{bmatrix}$$

where B and D are square symmetric submatrices. If B is invertible, the Schur Complement of B in Λ is given by $S = D - C^\top B^{-1}C$. In general, even if B is not invertible, B^{-1} can be replaced with the Moore-Penrose pseudoinverse B^\dagger in the Schur complement definition. Note that the Schur complement $S \succeq 0$ will be PSD for a PSD matrix $\Lambda \succeq 0$.

4. Diversity Evaluation for Text-to-Image Generative Models via CLIP Embedding

As discussed earlier, the CLIP model offers a joint embedding of text and image data, which enables defining joint (text,image) kernel covaraince matrices for the collected data. Suppose that we have collected n paired text,image samples (T_j, I_j) for $j = 1, \dots, n$. Here I_j represents the j th image and T_j represents the corresponding text. The application of CLIP embedding transfers the pair to the share embedding space \mathcal{X} and results in embedded samples (x_{T_j}, x_{I_j}) in the CLIP space. As a consequence of the joint embedding, not only can we compute the kernel function between (text,text) and (image,image) pairs, but also we can compute the kernel function for a (text,image) input.

To analyze the embedded sample, consider a kernel function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ with feature map $\phi : \mathcal{X} \rightarrow \mathbb{R}^d$ where \mathcal{X} is the CLIP space and d is the dimension of the kernel feature map. Then, for the collected samples, we define the embedded image feature matrix $\Phi_I \in \mathbb{R}^{n \times d}$ whose j th row is $\phi(x_{I_j})$ for the j th image. Similarly, we define the embedded text feature matrix $\Phi_T \in \mathbb{R}^{n \times d}$ for the text samples. Note that the resulting CLIP-based kernel covariance matrix for the joint (text,image) map $[\phi(\mathbf{x}_T), \phi(\mathbf{x}_I)]$ is

$$C_{\text{joint (I,T)}} := \begin{bmatrix} C_{II} & C_{IT} \\ C_{IT}^\top & C_{TT} \end{bmatrix}$$

In the above, we define the sub-covariances as follows:

$$C_{II} = \frac{1}{n} \Phi_I^\top \Phi_I, \quad C_{IT} = \frac{1}{n} \Phi_I^\top \Phi_T, \quad C_{TT} = \frac{1}{n} \Phi_T^\top \Phi_T$$

Note that the above matrix is PSD, which implies that we can leverage the Schur complement to decompose the image-based block C_{II} as follows:

$$C_{II} = \underbrace{C_{II} - C_{IT} C_{TT}^{-1} C_{IT}^\top}_{\text{model-driven component } \Lambda_I} + \underbrace{C_{IT} C_{TT}^{-1} C_{IT}^\top}_{\text{text component } \Lambda_T} \quad (6)$$

Proposition 1 Define text-to-image conversion matrix $\Gamma^* = C_{IT} C_{TT}^{-1}$. Then, Γ^* is an optimal solution to the following ($\|\cdot\|_F$ denotes the Frobenius norm):

$$\underset{\Gamma \in \mathbb{R}^{d \times d}}{\operatorname{argmin}} \frac{1}{n} \left\| \Phi_I^\top - \Gamma \Phi_T^\top \right\|_F^2$$

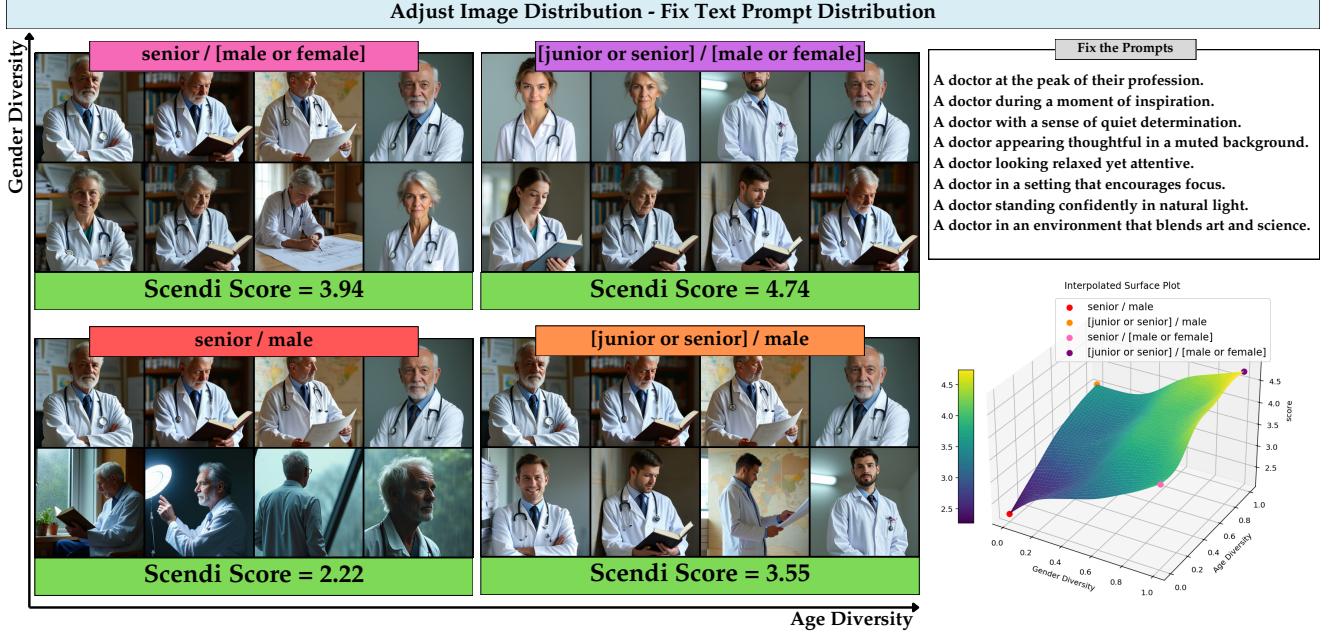


Figure 3. Evaluated Scendi scores with fixed text prompt distribution about doctors in various settings. Interpolated surface plot in the bottom right visualises the change in diversity (z-axis) when doctors have a varying diversity of age and gender features. (x and y axis)

Then, in (6), Λ_T is the covariance matrix of $\Gamma^* \phi(x_T)$, and Λ_I is the kernel covariance matrix of $\phi(x_I) - \Gamma^* \phi(x_T)$.

Remark 1 Proposition 1 shows that given the optimal text-to-image conversion matrix $\Gamma^* = C_{IT}C_{TT}^{-1}$, the effect of a text T on the embedding of an image I can be canceled by considering the remainder term $\phi(x_I) - \Gamma^* \phi(x_T)$ to decorrelate the image and text embedding.

If we consider the cosine-similarity kernel feature map $\phi(x) = x/\|x\|$, the above discussion reveals an approach to cancel the effect of the input text on the CLIP embedding of the output images. Given a paired dataset of prompts and images, we first compute the modification matrix $\Gamma^* = C_{IT}C_{TT}^{-1}$ and modify the CLIP embedding for the input image $x_I = \text{CLIP}(I)$ and prompt $x_T = \text{CLIP}(T)$ as:

$$\text{CLIP}_{\text{modified}}(I|T) := \text{CLIP}(I) - \Gamma^* \text{CLIP}(T) \quad (7)$$

Note that using the cosine similarity kernel, the dimension of Γ^* matches with the CLIP dimension, i.e., 512.

Furthermore, in the decomposition in (6), both the model-driven covariance component Λ_I and the text prompt-driven covariance component Λ_T are PSD matrices, which have non-unit trace values. To apply the matrix-based entropy definition which requires the unit trace, we rewrite the identity as

$$C_{II} = \text{Tr}(\Lambda_I) \cdot \frac{1}{\text{Tr}(\Lambda_I)} \Lambda_I + (1 - \text{Tr}(\Lambda_I)) \cdot \frac{1}{\text{Tr}(\Lambda_T)} \Lambda_T$$

Definition 1 We define the Schur-Complement-ENtropy Diversity (Scendi) score as follows:

$$\text{Scendi}(x_1, \dots, x_n; t_1, \dots, t_n) := \exp \left(\sum_{j=1}^d \lambda_j^{(\Lambda_I)} \log \frac{\text{Tr}(\Lambda_I)}{\lambda_j^{(\Lambda_I)}} \right)$$

where $\lambda_j^{(\Lambda_I)}$ denotes the j th eigenvalue of matrix Λ_I and $\text{Tr}(\Lambda_I) = \sum_{j=1}^d \lambda_j^{(\Lambda_I)}$ is the sum of the eigenvalues.

It can be seen that the defined Schur complement entropy is the conditional entropy of the hidden image cluster variable $\text{Mode} \in \{1, 2, \dots, d\}$, which follows the spectral decomposition of the image covariance matrix $C_{II} = \sum_{j=1}^d \lambda_j v_j v_j^\top$, where the mode $\text{Mode} = i$ has probability λ_i . Here we define a text-based guess of the image mode Y_T , which as we discuss in the Appendix can correctly predict the cluster with probability $\text{Tr}(\Lambda_T)$ and outputs erasure e with probability $\text{Tr}(\Lambda_I)$. Then, we show in the Appendix that following standard Shannon entropy definition:

$$\text{Scendi}(x_1, \dots, x_n; t_1, \dots, t_n) = \exp(H(\text{Mode}|Y_T))$$

Therefore, this analysis allows us to decompose the diversity of generated images into model-driven and text-driven components.

5. Numerical Results

We evaluated our proposed Schur Complement-based approach for CLIP decomposition across various scenarios

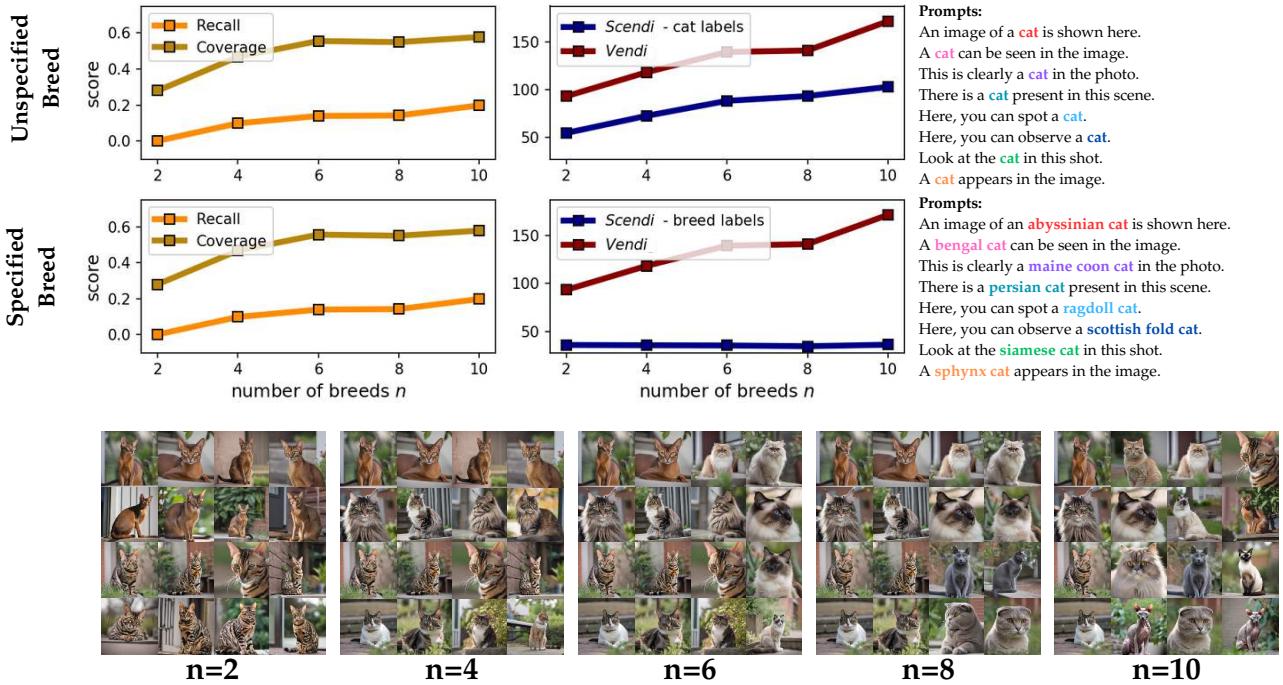


Figure 4. Evaluated Scendi, Vendi, Recall and Coverage scores with Gaussian Kernel on different cat breeds dataset.

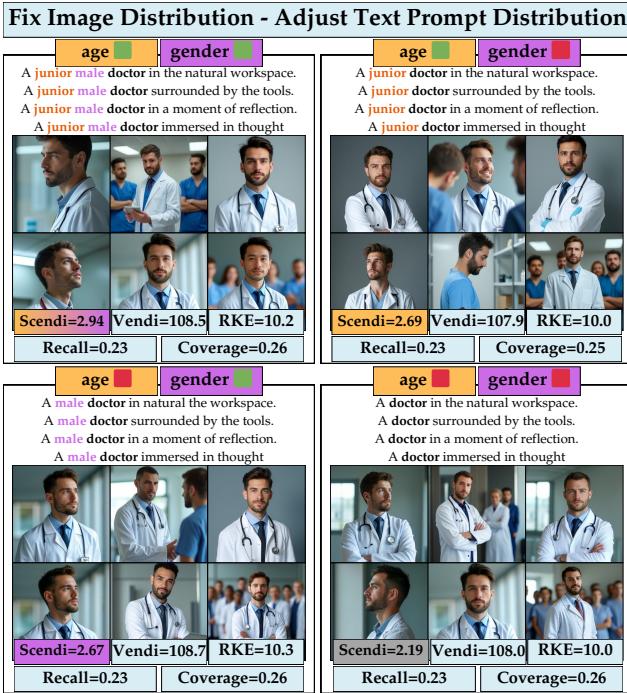


Figure 5. Evaluated Scendi scores with fixed text prompt distribution. Scendi changes to reflect the diversity contributed by the varying level of information in text prompts.

using both real and synthetic datasets. Our experimental results are reported for the standard Cosine Similarity

and Gaussian kernel functions. For the Gaussian kernel function, we use random Fourier features [46] with a random Fourier dimension $r = 2000$ to embed the kernel using a finite feature dimension. In the diversity evaluation, we report our defined Scendi scores as well as the Vendi score [9], RKE[20], Recall[28] and Coverage[39] for unconditional diversity assessment without taking the text data into account. Note that Recall and Coverage may only be performed in the presence of a valid representative reference dataset. The clustering experiments were performed using Kernel PCA with the Gaussian kernel, highlighting the top clusters that align with the top eigenvector directions in the data.

Measuring Text-to-Image Model Diversity. To highlight the strengths of Scendi-based diversity evaluation, we conducted numerical experiments comparing unconditional and conditional diversity assessments. Figure 3 presents an experiment with the FLUX.1-dev [29] model, where individuals are generated based on an ambiguous prompt. We fixed the text prompt distribution and conditioned the model to generate doctors of varying gender and age. The results show that when the model is restricted to producing only "senior male doctors," the Scendi score is the lowest. Conversely, when the model generates a broader range of age groups and genders, the diversity score increases. We also provide an interpolated surface plot illustrating the intermediate mixing results. A diversity level of 0 indicates a strong bias toward

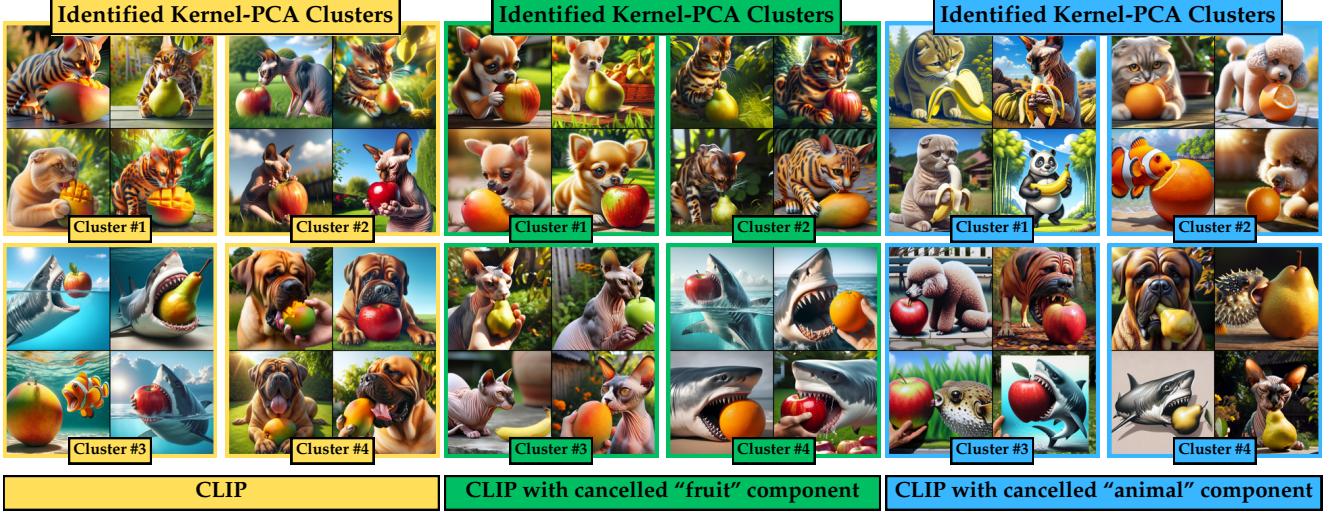


Figure 6. Clusters of DALL-E 3 [40] generated images of animals with fruits. The yellow column shows Kernel-PCA (KPCA) clusters using CLIP; the green column shows KPCA clusters with the application the proposed Schur-Complement-based method to remove "fruit" direction from CLIP embedding; and the blue column shows clusters with the "animal" direction removed from CLIP embedding.

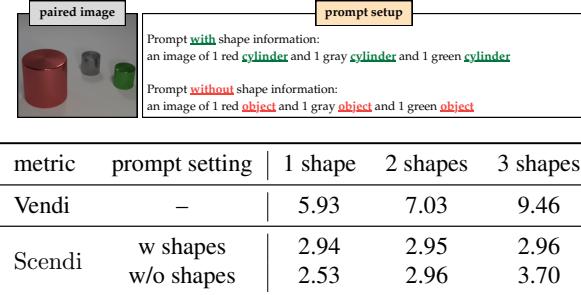


Figure 7. Diversity measurement on the CLEVR dataset using Vendi and Scendi metrics. Vendi stays constant for all prompts, while Scendi only increases when shapes aren't specified, highlighting true image-driven diversity.

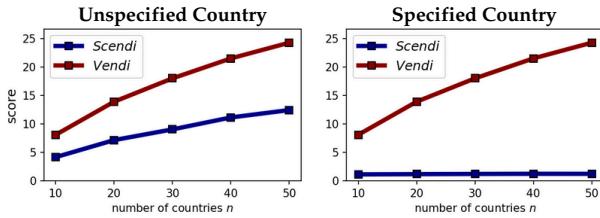


Figure 8. LLM countries with Qwen3-0.6B [57]

a single characteristic (e.g., all images depict males), whereas a score of 1 signifies an equal representation of all characteristics (e.g., an equal number of male and female images).

Figure 5 illustrates the inverse scenario: the image distribution is fixed while the accompanying text prompts vary in their specificity. We generated four sets of "junior male doctor" images with different seeds, while prompts

were assigned four levels of detail—fully specified (age and gender), partially specified (age only or gender only), or unspecified (no details). We measured Scendi, Vendi, RKE, Coverage and Recall scores across four scenarios. Reference set for R/C is "doctor" class in IdenProf dataset [38]. Vendi, RKE, Recall and Coverage scores remain relatively unchanged, whereas Scendi distinguishes between these cases. When the description is incomplete, Scendi is lowest due to the strong presence of "junior male doctors" across all images, indicating bias. However, when prompts explicitly state that all images should depict "junior male doctors," diversity increases and is evaluated based on auxiliary features such as background details rather than age or gender. When partial information is given, Scendi falls between the fully specified and unspecified cases. These results highlight the need for differentiation between model-driven and prompt-driven diversity.

To further quantify our findings, in Figure 4 we conducted an experiment on SDXL generated cat images of different breeds. We generated 1,000 images per class and compared Scendi with other diversity metrics. When prompts did not specify cat breeds, the diversity captured by Scendi aligns with unconditional Vendi score. Conversely, when breed information was included in the prompts, diversity mainly stemmed from the text prompts rather than the image generator, leading to a non-increasing Scendi score. In both scenarios, Vendi, Coverage, and Recall scores remained unchanged. The reference set for Recall and Coverage consisted of ImageNet [7] 'cat' samples. Additional experiments on other cases are discussed in the Appendix. We evaluated our metric on a subset of CLEVR [24] images containing only one shape to



Figure 9. Kernel PCA clusters before and after CLIP correction on captioned ImageNet dataset.

benchmark Scendi in a fully controlled environment. Each image was paired with two prompts: one specifying the shape and one omitting it, while keeping object count and colors constant. In a diversity evaluation analogous to Figure 4, similar trend remains.

This work primarily focuses on text-to-image evaluation using CLIP; however, the Scendi score naturally extends to any uni-modal evaluation setting, such as text-to-text, which has become extremely popular with LLMs. Figure 8 mirrors our two previous experimental setups, but here the dataset consists of country prompts paired with short, fact-only responses generated by Deepseek-V3 [6]. Notably, this experiment demonstrates that in a uni-modal context the Scendi score applies directly and does not require CLIP or other cross-modal embeddings, any suitable embedding will suffice.

CLIP Decomposition. We use the eigenspace of the Schur complement component to interpret the diversity in the generated images that is independent of the text prompt, revealing the unique elements introduced by the generation model in the images. The eigenspace-based interpretation follows the application of the Kernel PCA (KPCA) clustering method and leads to a visualization of the sample clusters due to the input text clusters and the clusters with a feature added by the model to the generated image. Figure 6 shows the resulting KPCA-detected clusters for DALL-E 3 [40] generated images in response to prompts of type “[animal] eating a [fruit]” with 5 different animal and fruit names. The detected clusters of CLIP image embedding (leftmost case) shows clusters with mixed animals and fruits. On the other hand, by considering the prompt of only specifying the fruit, the Schur complement component’s KPCA clusters highlight the animals in each cluster (middle case), and similarly by considering the prompt of only specifying the fruit, the animals are captured by each KPCA cluster of the Schur complement matrix. The results show how decomposing and removing the influence of concepts present in the prompts leads to changes in the major clusters. Further decomposition results and analysis are provided in the Appendix.

Typographic Attacks. CLIP’s sensitivity to text

within images, as showed in [37], makes it vulnerable to typographic attacks, where overlaying misleading text on an image influences its classification [31]. To investigate this, we constructed a dataset of 10 ImageNet classes, each image engraved with a random class label. Figure 9 demonstrates CLIP’s susceptibility, showing that top eigenvector directions capture the engraved text rather than the actual image content. By conditioning on prompts like “text reading ‘cassette player’ on top of image” and removing this direction, we re-clustered images based on the corrected CLIP embedding. This adjustment allowed us to identify clusters based on the actual image content rather than the engraved text. Additional results are provided in the Appendix.

6. Conclusion

In this work, we proposed a Schur Complement-based approach to decompose the kernel covariance matrix of CLIP image embeddings into the sum of text-induced and model-induced kernel covariance components. We demonstrated the application of this Schur Complement-based approach to define the Scendi score that evaluates the diversity of prompt-guided generated data in a prompt-aware fashion. This method extends the application of CLIP embeddings from relevance evaluation to assessing diversity in text-to-image generation models. Additionally, we applied our approach to modify CLIP image embeddings by modifying or canceling the influence of an input text. Our numerical results showcase the potential for canceling text effects and its applications to computer vision tasks. Future research directions include extending the Schur Complement-based approach to other generative AI models, such as text-to-video and language models. Additionally, leveraging this decomposition to fine-tune the CLIP model for enhanced understanding of diverse concepts represents another future direction. Finally, applying Scendi score for diversity guidance in diffusion models, similar to RKE guidance in [22] and Vendi guidance in [13], will be relevant for future studies.

Acknowledgements

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

References

- [1] Pietro Astolfi, Marlene Careil, Melissa Hall, Oscar Mañas, Matthew Muckley, Jakob Verbeek, Adriana Romero Soriano, and Michal Drozdzal. Consistency-diversity-realism pareto fronts of conditional image generative models, 2024. 3
- [2] Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio P. Calmon, and Himabindu Lakkaraju. Interpreting clip with sparse linear concept embeddings (splice). In *Advances in Neural Information Processing Systems*, 2024. 3
- [3] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 3
- [4] Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. SEGA: Instructing text-to-image models using semantic guidance. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3
- [5] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791, 2021. 3
- [6] DeepSeek-AI. Deepseek-v3 technical report, 2024. 8
- [7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 7
- [8] Terrance DeVries, Adriana Romero, Luis Pineda, Graham W. Taylor, and Michal Drozdzal. On the evaluation of conditional gans, 2019. 3
- [9] Dan Friedman and Adji Bouso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on machine learning research*, 2023. 2, 3, 4, 6
- [10] Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting CLIP’s image representation via text-based decomposition. In *The Twelfth International Conference on Learning Representations*, 2024. 3
- [11] Shizhan Gong, LEI Haoyu, Qi Dou, and Farzan Farnia. Boosting the visual interpretability of clip via adversarial fine-tuning. In *The Thirteenth International Conference on Learning Representations*, 2025. 3
- [12] Shizhan Gong, Yankai Jiang, Qi Dou, and Farzan Farnia. Kernel-based unsupervised embedding alignment for enhanced visual representation in vision-language models. *arXiv preprint arXiv:2506.02557*, 2025. 3
- [13] Reyhane Askari Hemmat, Melissa Hall, Alicia Sun, Candace Ross, Michal Drozdzal, and Adriana Romero-Soriano. Improving geo-diversity of generated images with contextualized vendi score guidance. In *European Conference on Computer Vision*, 2024. 8
- [14] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 2, 3
- [15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 3
- [16] Xiaoyan Hu, Ho fung Leung, and Farzan Farnia. A multi-armed bandit approach to online selection and evaluation of generative models, 2025. 3
- [17] Xiaoyan Hu, Ho-fung Leung, and Farzan Farnia. An online learning approach to prompt-based selection of generative models and llms. In *Forty-second International Conference on Machine Learning*, 2025.
- [18] Xiaoyan Hu, Lauren Pick, Ho fung Leung, and Farzan Farnia. Promptwise: Online learning for cost-aware prompt assignment in generative models. 2025. 3
- [19] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 4
- [20] Mohammad Jalali, Cheuk Ting Li, and Farzan Farnia. An information-theoretic evaluation of generative models in learning multi-modal distributions. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2, 3, 4, 6
- [21] Mohammad Jalali, Azim Ospanov, Amin Gohari, and Farzan Farnia. Conditional vendi score: An information-theoretic approach to diversity evaluation of prompt-based generative models, 2024. 3
- [22] Mohammad Jalali, Haoyu Lei, Amin Gohari, and Farzan Farnia. Sparke: Scalable prompt-aware diversity guidance in diffusion models via rke score. *arXiv preprint arXiv:2506.10173*, 2025. 8
- [23] Mohammad Jalali, Bahar Dibaei Nia, and Farzan Farnia. Towards an explainable comparison and alignment of feature embeddings. *arXiv preprint arXiv:2506.06231*, 2025. 3
- [24] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Conference on Computer Vision and Pattern Recognition*, 2017. 7
- [25] Nithish Kannen, Arif Ahmad, Marco Andreetto, Vinodkumar Prabhakaran, Utsav Prabhu, Adji Bouso Dieng, Pushpak Bhattacharyya, and Shachi Dave. Beyond aesthetics: Cultural competence in text-to-image models, 2024. 3
- [26] Dongkyun Kim, Mingi Kwon, and Youngjung Uh. Attribute based interpretable evaluation metrics for generative

- models. In *Forty-first International Conference on Machine Learning*, 2024. 3
- [27] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2426–2435, 2022. 3
- [28] Tuomas Kynkänniemi, Tero Karras, Samuli Laine, Jaakk Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 3, 6
- [29] Black Forest Lab. Flux: A diffusion-based text-to-image (t2i) model. <https://github.com/blackforestlab/flux>, 2024. Accessed: 2024-09. 6, 3, 4
- [30] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Benita Teufel, Marco Bellagente, Minguk Kang, Taesung Park, Jure Leskovec, Jun-Yan Zhu, Li Fei-Fei, Jiajun Wu, Stefano Ermon, and Percy Liang. Holistic evaluation of text-to-image models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 3
- [31] Yoann Lemesle, Masataka Sawayama, Guillermo Valle-Perez, Maxime Adolphe, Hélène Sauzéon, and Pierre-Yves Oudeyer. Language-biased image classification: evaluation based on semantic representations. In *International Conference on Learning Representations*, 2022. 8
- [32] Meir Yossef Levi and Guy Gilboa. The double-ellipsoid geometry of CLIP. In *Forty-second International Conference on Machine Learning*, 2025. 3
- [33] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 4
- [34] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. Featured Certification, Expert Certification. 3
- [35] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 4
- [36] Yan Luo, Min Shi, Muhammad Osama Khan, Muhammad Muneeb Afzal, Hao Huang, Shuaihang Yuan, Yu Tian, Luo Song, Ava Kouhana, Tobias Elze, Yi Fang, and Mengyu Wang. Fairclip: Harnessing fairness in vision-language learning, 2024. 4
- [37] Joanna Materzynska, Antonio Torralba, and David Bau. Disentangling visual and written concepts in clip. In *CVPR*, 2022. 3, 8
- [38] Olafenwa Moses. Idenprof: A pre-trained deep learning model for identifying professionals in images, 2018. Accessed: 2025-03-08. 7
- [39] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, pages 7176–7185. PMLR, 2020. 2, 3, 6
- [40] OpenAI. Dall-e 3. <https://openai.com/index/dall-e-3/>, 2023. 2, 7, 8, 4
- [41] Azim Ospanov and Farzan Farnia. Do vendi scores converge with finite samples? truncated vendi score for finite-sample convergence guarantees. In *The 41st Conference on Uncertainty in Artificial Intelligence*, 2025. 3
- [42] Azim Ospanov, Jingwei Zhang, Mohammad Jalali, Xuenan Cao, Andrej Bogdanov, and Farzan Farnia. Towards a scalable reference-free evaluation of generative models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3
- [43] Amey Pasarkar and Adji Bousso Dieng. Cousins of the vendi score: A family of similarity-based diversity metrics for science and machine learning. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2024. 3
- [44] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 11
- [45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 3
- [46] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007. 4, 6
- [47] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 4
- [48] Anton Razhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov, Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion, 2023. 4
- [49] Parham Rezaei, Farzan Farnia, and Cheuk Ting Li. Be more diverse than the most diverse: Optimal mixtures of generative models via mixture-ucb bandit algorithms, 2025. 3

- [50] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 2018. [2](#)
- [51] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2016. [3](#)
- [52] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626. IEEE Computer Society, 2017. [3](#)
- [53] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *CoRR*, abs/1703.01365, 2017. [3](#)
- [54] Zixiao Wang, Farzan Farnia, Zhenghao Lin, Yunheng Shen, and Bei Yu. On the distributed evaluation of generative models. *arXiv preprint arXiv:2310.11714*, 2023. [3](#)
- [55] Jingwei Zhang, Cheuk Ting Li, and Farzan Farnia. An interpretable evaluation of entropy-based novelty of generative models. *arXiv preprint arXiv:2402.17287*, 2024. [3](#)
- [56] Jingwei Zhang, Mohammad Jalali, Cheuk Ting Li, and Farzan Farnia. Unveiling differences in generative models: A scalable differential clustering approach. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8269–8278, 2025. [3](#)
- [57] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025. [7](#)

A. Proofs

A.1. Proof of Proposition 1

We aim to solve the optimization problem:

$$\Gamma^* = \underset{\Gamma \in \mathbb{R}^{d \times d}}{\operatorname{argmin}}; \frac{1}{n} \|\Phi_I^\top - \Gamma \Phi_T^\top\|_F^2,$$

where $\Phi_I, \Phi_T \in \mathbb{R}^{d \times n}$ are given matrices, and $\|\cdot\|_F$ denotes the Frobenius norm.

To find the optimal Γ^* , we begin by expanding the objective function. Recall that the squared Frobenius norm of a matrix A is given by $\|A\|_F^2 = \operatorname{Tr}(A^\top A)$. Therefore, we have:

$$\begin{aligned} f(\Gamma) &= \frac{1}{n} \|\Phi_I^\top - \Gamma \Phi_T^\top\|_F^2 \\ &= \frac{1}{n} \operatorname{Tr} \left[(\Phi_I^\top - \Gamma \Phi_T^\top)^\top (\Phi_I^\top - \Gamma \Phi_T^\top) \right] \\ &= \frac{1}{n} \operatorname{Tr} [\Phi_I \Phi_I^\top - \Gamma \Phi_I \Phi_T^\top - \Phi_T \Phi_I^\top \Gamma^\top + \Gamma \Phi_T \Phi_T^\top \Gamma^\top]. \end{aligned}$$

Let us define the covariance matrices:

$$\begin{aligned} C_{II} &= \Phi_I \Phi_I^\top \in \mathbb{R}^{d \times d}, \\ C_{IT} &= \Phi_I \Phi_T^\top \in \mathbb{R}^{d \times d}, \\ C_{TI} &= \Phi_T \Phi_I^\top = C_{IT}^\top \in \mathbb{R}^{d \times d}, \\ C_{TT} &= \Phi_T \Phi_T^\top \in \mathbb{R}^{d \times d}. \end{aligned}$$

Substituting these definitions into $f(\Gamma)$, we obtain:

$$f(\Gamma) = \frac{1}{n} \operatorname{Tr} [C_{II} - \Gamma C_{IT} - C_{TI} \Gamma^\top + \Gamma C_{TT} \Gamma^\top].$$

To find the minimizer, we compute the Jacobian of $f(\Gamma)$ with respect to Γ . Using standard matrix derivative identities, we have:

$$\begin{aligned} J_\Gamma f(\Gamma) &= \frac{1}{n} (-C_{IT}^\top - C_{TI} + 2\Gamma C_{TT}) \\ &= \frac{1}{n} (-C_{IT}^\top - C_{IT}^\top + 2\Gamma C_{TT}) \quad (\text{since } C_{TI} = C_{IT}^\top) \\ &= \frac{1}{n} (-2C_{IT}^\top + 2\Gamma C_{TT}). \end{aligned}$$

We observe that by choosing $\Gamma^* = C_{TI} C_{TT}^{-1}$, we will have

$$J_\Gamma f(\Gamma^*) = -\frac{2}{n} C_{IT}^\top + \frac{2}{n} \Gamma^* C_{TT} = \mathbf{0}.$$

Therefore, Γ^* is a stationary point in the optimization problem with a convex objective function and hence is an optimal solution to the minimization task.

A.2. Conditional Entropy Interpretation of Scendi Score

As discussed in the main text, in Equation (6), both image component Λ_I and text component Λ_T are PSD matrices with unit trace. Furthermore, we have

$$C_{II} = \operatorname{Tr}(\Lambda_I) \cdot \frac{1}{\operatorname{Tr}(\Lambda_I)} \Lambda_I + (1 - \operatorname{Tr}(\Lambda_I)) \cdot \frac{1}{\operatorname{Tr}(\Lambda_T)} \Lambda_T$$

Next, we consider the spectral decomposition of matrix $C_{II} = \sum_{i=1}^d \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$ given its non-negative eigenvalues $\lambda_1 \geq \dots \geq \lambda_d$ and orthonormal eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_n$. Following the orthonormality of the eigenvectors, we have the following for every $j \in \{1, \dots, d\}$:

$$\lambda_j = \text{Tr}(\Lambda_I) \cdot \frac{1}{\text{Tr}(\Lambda_I)} \mathbf{v}_j^\top \Lambda_I \mathbf{v}_j + (1 - \text{Tr}(\Lambda_I)) \cdot \frac{1}{\text{Tr}(\Lambda_T)} \mathbf{v}_j^\top \Lambda_T \mathbf{v}_j$$

Therefore, if we define the Mode random variable over $\{1, \dots, d\}$ with probabilities $\lambda_1, \dots, \lambda_d$, its unconditional Shannon entropy will be $H(\text{Mode}) = \sum_{i=1}^d \lambda_i \log(1/\lambda_i)$. On the other hand, if an adversary has the side knowledge of the text it can correctly predict Mode = j with probability $\mathbf{v}_j^\top \Lambda_T \mathbf{v}_j$. If we define Y_{adv} as the correct prediction of this adversary when the text can be correctly mapped to the mode variable and else we define $Y_{\text{adv}} = e$ as the error, then the conditional entropy will be:

$$\begin{aligned} H(\text{Mode}|Y_{\text{adv}}) &= P(Y_{\text{adv}} = e)H(\text{Mode}|Y_{\text{adv}} = e) + P(Y_{\text{adv}} \neq e)H(\text{Mode}|Y_{\text{adv}} \neq e) \\ &= P(Y_{\text{adv}} = e)H(\text{Mode}|Y_{\text{adv}} = e) + P(Y_{\text{adv}} \neq e) \times 0 \\ &= P(Y_{\text{adv}} = e)H(\text{Mode}|Y_{\text{adv}} = e) \\ &= \left(\sum_{j=1}^d v_j^\top \Lambda_I v_j \right) \sum_{j=1}^d \frac{v_j^\top \Lambda_I v_j}{\sum_{t=1}^d v_t^\top \Lambda_I v_t} \log \frac{\sum_{t=1}^d v_t^\top \Lambda_I v_t}{v_j^\top \Lambda_I v_j} \\ &= \sum_{j=1}^d (v_j^\top \Lambda_I v_j) \log \frac{\sum_{t=1}^d v_t^\top \Lambda_I v_t}{v_j^\top \Lambda_I v_j} \end{aligned}$$

Note that $\sum_{t=1}^d v_t^\top \Lambda_I v_t = \sum_{t=1}^d \text{Tr}(v_t^\top \Lambda_I v_t) = \text{Tr}(\sum_{t=1}^d v_t v_t^\top \Lambda_I) = \text{Tr}(\Lambda_I)$ which implies that

$$\begin{aligned} H(\text{Mode}|Y_{\text{adv}}) &= \sum_{j=1}^d (v_j^\top \Lambda_I v_j) \log \frac{\text{Tr}(\Lambda_I)}{v_j^\top \Lambda_I v_j} \\ &= \log(\text{Tr}(\Lambda_I)) \left(\sum_{j=1}^d (v_j^\top \Lambda_I v_j) \right) + \sum_{j=1}^d (v_j^\top \Lambda_I v_j) \log \frac{1}{v_j^\top \Lambda_I v_j} \\ &= \log(\text{Tr}(\Lambda_I)) \text{Tr}(\Lambda_I) + \sum_{j=1}^d (v_j^\top \Lambda_I v_j) \log \frac{1}{v_j^\top \Lambda_I v_j} \end{aligned}$$

which assuming that Λ_I and C_{II} share the same eigenvectors will provide

$$\begin{aligned} H(\text{Mode}|Y_{\text{adv}}) &= \log(\text{Tr}(\Lambda_I)) \text{Tr}(\Lambda_I) + \sum_{j=1}^d (\lambda_j^{(\Lambda_I)}) \log \frac{1}{\lambda_j^{(\Lambda_I)}} \\ &= \sum_{j=1}^d \lambda_j^{(\Lambda_I)} \log \frac{\text{Tr}(\Lambda_I)}{\lambda_j^{(\Lambda_I)}} \end{aligned}$$

Note that the above provides our definition of the Schur-Complement-Entropy for the image part Scendi_I and the text part Scendi_T as follows:

$$\text{Scendi}_I(x_1, \dots, x_n) = \sum_{j=1}^d \lambda_j^{(\Lambda_I)} \log \frac{\text{Tr}(\Lambda_I)}{\lambda_j^{(\Lambda_I)}} \quad (8)$$

$$\text{Scendi}_T(x_1, \dots, x_n) := \sum_{j=1}^d \lambda_j^{(\Lambda_T)} \log \frac{\text{Tr}(\Lambda_T)}{\lambda_j^{(\Lambda_T)}} \quad (9)$$

where $\lambda_j^{(\Lambda_I)}$ denotes the j th eigenvalue of matrix Λ_I and $\text{Tr}(\Lambda_I) = \sum_{j=1}^d \lambda_j^{(\Lambda_I)}$ is the sum of the eigenvalues. Note that we follow the same definition for the text part Λ_T .

B. Limitations

The Scendi framework is only compatible with cross-modal embeddings, such as those produced by CLIP. When such embeddings are unavailable for a given data modality, evaluators cannot use Scendi to measure diversity. Extending Scendi to modalities without cross-modal embeddings remains an open challenge and a promising direction for future work.

C. Additional Individual Image Decomposition Results via SC-Based Method

In this section, we present additional CLIP decomposition results for randomly selected pairs of ImageNet labels. The correction matrix was computed using the captioned MSCOCO dataset. The experimental setup follows the approach illustrated in Figure 2. We generated images containing predominantly two concepts and applied the SC-based method for decomposition. Subsequently, we measured the cosine similarity between the corrected and regular CLIP embeddings and the CLIP-embedded ImageNet samples. The top four images with the highest similarity scores are reported. These results demonstrate the effectiveness of the Schur Complement method in decomposing directions present in generated images.

Results for synthetic images generated using SDXL are shown in Figures 10 and 11. Corresponding results for DALL-E 3 are presented in Figures 12 and 13. Notably, the Schur Complement-based decomposition successfully isolates and removes image directions corresponding to a text condition that describes the concept to be excluded.

To expand on the results in Figure 6, we constructed a dataset of animals with traffic signs using FLUX.1-schnell [29]. Figure 20 illustrates that after canceling either of the subjects using the Schur complement method, Kernel-PCA clusters according to the remaining concepts in the image.

Moreover, to test how CLIP correction affects the underlying directions of concepts, we applied the CLIPDiffusion [27] framework to edit the image according to different CLIP embeddings. Figure 14 illustrates the setup of the problem, where we edit the ‘initialization image’ that consists of two subjects: a cat and a basketball. We then denoise and guide the generation according to three different embeddings: the unchanged CLIP embedding of the ‘initialization image’, the modified CLIP by a ‘cat’ direction, or the modified CLIP by a ‘basketball’ direction. We show that after removing a concept direction, the denoiser is no longer rewarded for generating the corresponding concept, which is reflected in the denoised images. After correction, the basketball resembles a bowl with plants, and the cat loses its features. We also note that in both cases, the other object remains intact. To further showcase these results, we performed the same diffusion on the animals with traffic signs dataset, shown on the side of Figure 14.

D. Additional Results on Diversity Evaluation

To further validate the findings presented in the main text, we conducted a similar experiment (Figure 4) using the Cosine Similarity Kernel. The results confirm that the diversity trends observed in the main text persist under a finite-dimensional kernel. Figures 15 and 16 illustrate the variation in Scendi diversity when conditioned on different text prompts.

Similar to Figure 4, we conducted similar experiment with animals and objects dataset in Figure 17. Our results mirror previous findings, strengthening the proposed diversity evaluation metric in measuring subject quantity related diversity.

Moreover, we evaluate the diversity of typographically attacked ImageNet samples in Figure 21. Specifically, we overlay the text “cassette player” onto images from 10 different ImageNet classes and measure diversity as the number of distinct classes increases. The presence of overlaid text diverts CLIP’s sensitivity away from image content, causing it to encode the direction indicated by the text instead.

To illustrate this effect, we visualize salience maps of CLIP embeddings given a prompt referring to an object behind the overlaid text. The results show that CLIP is highly sensitive to centrally placed text, even when it is unrelated to the prompted object. However, applying SC-based decomposition mitigates this bias. This correction is reflected in the diversity plots, where $Scendi_I$ increases rapidly as the number of ImageNet classes grows, whereas Vendi, Coverage, and Recall exhibit much weaker correlations with class diversity.

E. Additional Experiments on the Image Captioning Task

In the main text, we discussed SC-based decomposition for text-to-image models and demonstrated how images can be decomposed given text prompts. Here, we show that the reverse process is also possible. Specifically, we explored decomposing captions based on their corresponding images.

The experimental setup mirrors that of Figure 18, with the key difference being that, instead of generating images for text prompts, we generated captions for the corresponding images. For this task, we used *gpt4o-mini* as the captioning model. Figure 19 illustrates the experimental setup.

We selected images closely aligned with the concept we aimed to remove. For instance, to eliminate the "cat" direction in text, we used an image of a cat against a white background to better isolate the concept. After applying corrections for "animals" or "objects" in the text prompt, we observed successful decomposition, as reflected in the second column: the corrected CLIP embedding is no longer sensitive to the removed concept.

These findings highlight the versatility of the Scendi method, demonstrating its applicability across a wide range of tasks that rely on a shared embedding space.

F. Robustness of Scendi

We note that the robustness of the Scendi framework depends on the choice of underlying embedding. To address limitations in CLIP, several alternatives have been introduced, such as FairCLIP [36]. Because Scendi is compatible with any cross-modal embedding, we evaluated diversity using three additional variants: OpenCLIP [19], FairCLIP, and BLIP2 [33] on the SDXL generated cat breed dataset. The results appear in Figure 22. We observe that Scendi preserves the qualities of a diversity metric that increases as we introduce more breeds into the data pool.

G. Results with Naive Text Embedding Subtraction without considering the adjustment matrix Γ^*

In the given task setting, it may seem intuitive to assume that the difference between Φ_I and Φ_T would yield a similar outcome as the SC-based decomposition. To test this hypothesis, we compared the SC-based decomposition with a "naive" method, defined as $\Phi_I - \Phi_T$. In this naive approach, the learned correction matrix Γ^* is replaced with an identity matrix, effectively omitting its computation. Our experiments reveal that such a decomposition usually fails to achieve the desired results and often leads to a loss of coherent directionality in the embedding space.

To evaluate the performance of the naive embedding subtraction method, we used the typographic attack dataset, which consists of 10 ImageNet classes where misleading text is overlaid on the images. We measured classification accuracy before and after decomposition. Figure 24 shows the distribution of classifications for images engraved with the text "cassette player." CLIP's classification is heavily biased towards "cassette player," despite the underlying images belonging to a different class. After decomposition, the naive method removes the direction corresponding to the text but results in a skew towards "french horn," even though the image distribution is uniform across all 10 classes. In contrast, the SC-based decomposition corrects the embeddings, making them sensitive to the underlying images rather than the engraved text.

To further demonstrate the effectiveness of the SC method, we compared kernel PCA clusters in Figure 26. The clustering results for the naive decomposition closely resemble those without any correction, indicating that this method does not address the typographic attack. On the other hand, SC-based decomposition significantly improves the clustering by accurately resolving the misleading text directionality.

Additionally, we performed CLIP-guided diffusion to visualize the contents of the corrected embeddings. The setup is illustrated in figure 23 and it is similar to the one described in the main text, except we do not use Γ^* in the decomposition of CLIP. Figure 25 compares images generated using naive and SC-based decompositions. The naive method performs poorly, particularly when text overlays traffic signs, and often removes directions without preserving information about other underlying concepts in the images. In contrast, the SC-based decomposition preserves the structural and semantic information while successfully removing the undesired text directionality.

These results highlight the necessity of computing the correction matrix Γ^* to effectively remove specific directions while preserving information about other concepts within the image embeddings.

H. Additional Numerical Results

We evaluated several text-to-image models using the Scendi metric to assess their performance. Figure 27 summarizes our findings for DALL-E 2 [47], DALL-E 3 [40], Kandinsky 3 [48], and FLUX.1-schnell [29], tested on 5,000 MSCOCO [35] captions.

Our results demonstrate that the SC-Vendi metric correlates with the Vendi score, which measures the diversity of image generators. This suggests that when tested on the MSCOCO dataset, image diversity arises not only from the text prompts but also from the intrinsic properties of the generator itself.

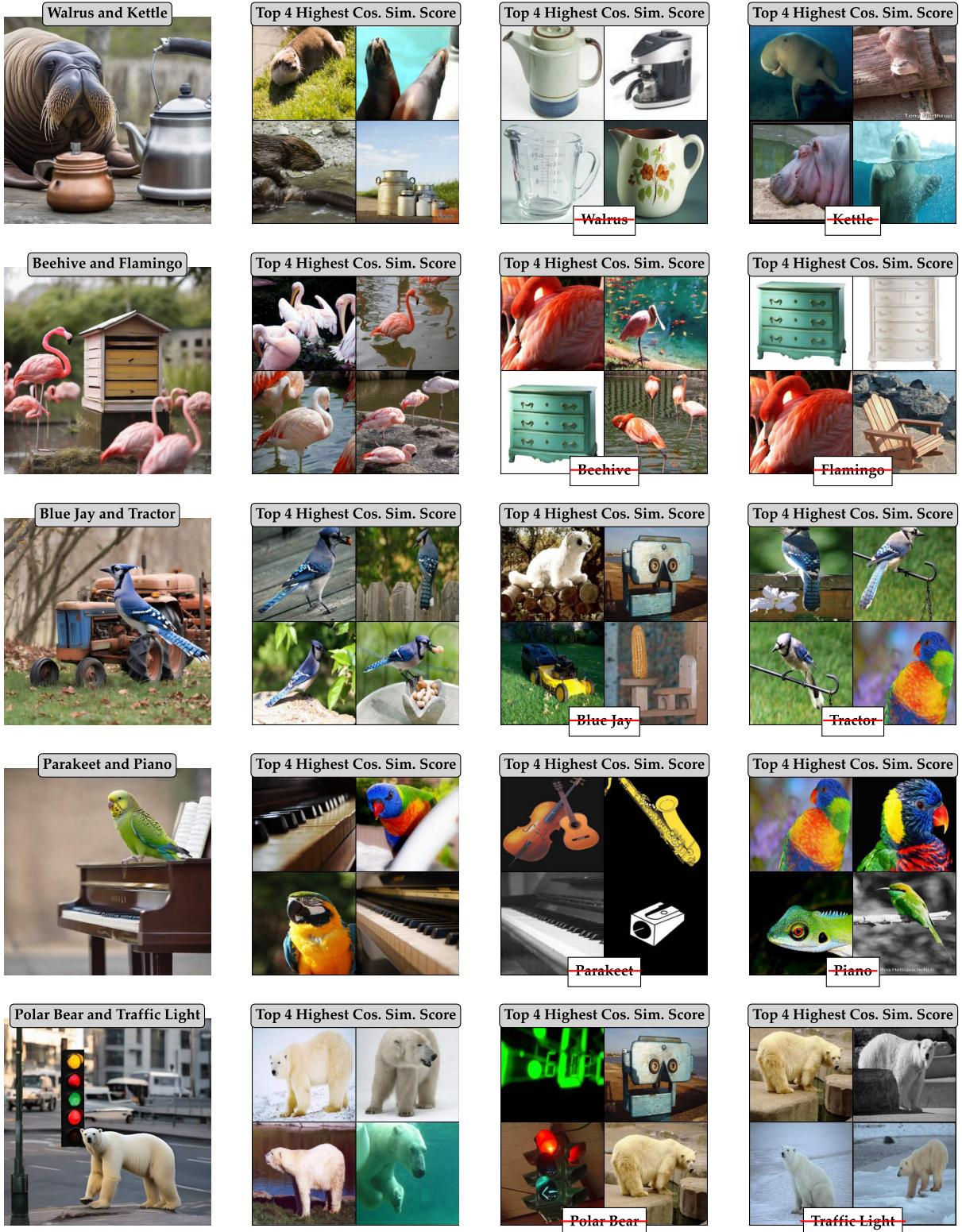


Figure 10. Diagram presenting the decomposition of SDXL generated images of two random labels from ImageNet. First column presents the generated image of a pair. Second column presents four images from ImageNet with highest Cosine Similarity Score. Third and Fourth columns showcase feature removal from the image.

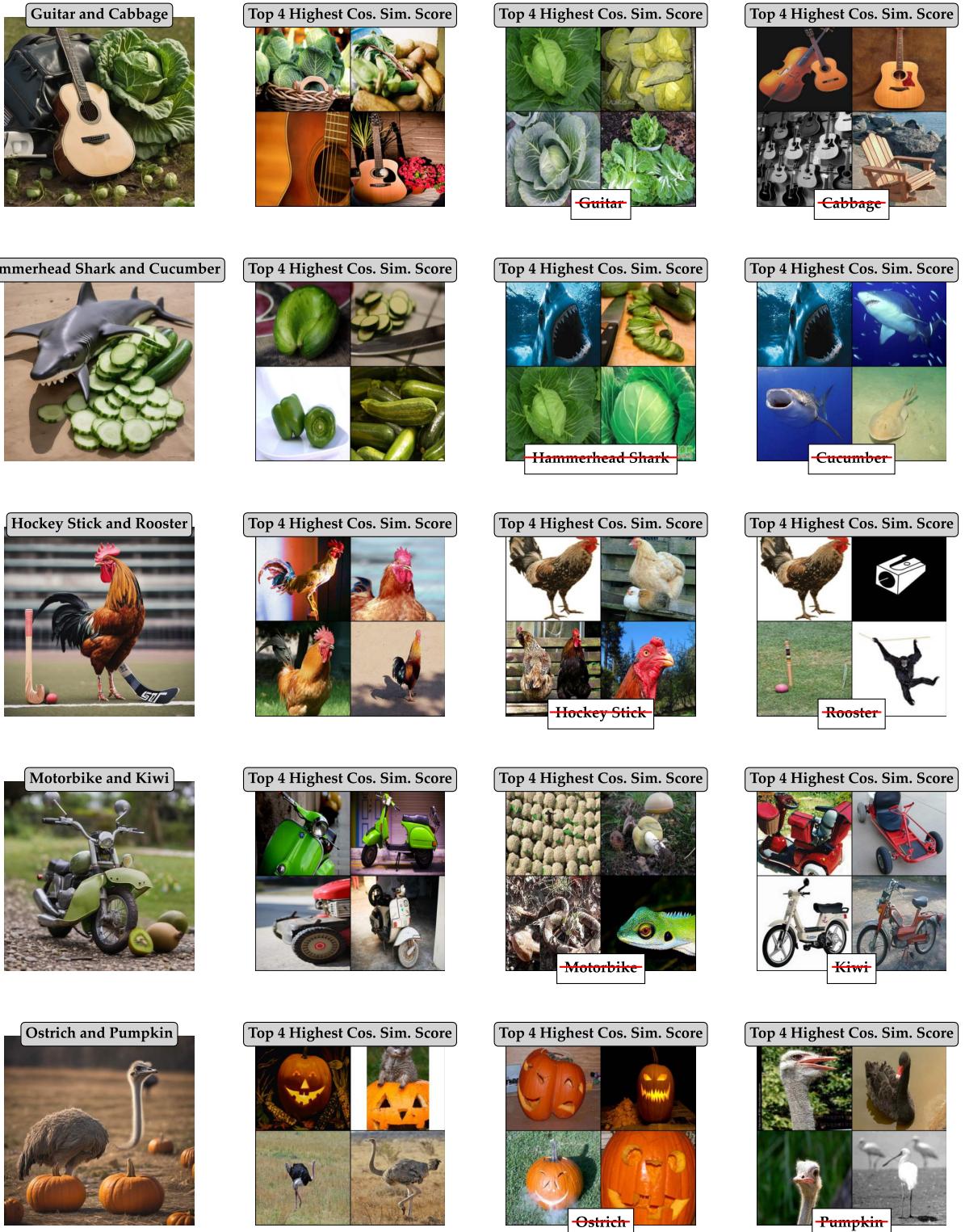


Figure 11. Diagram presenting the decomposition of SDXL generated images of two random labels from ImageNet. First column presents the generated image of a pair. Second column presents four images from ImageNet with highest Cosine Similarity Score. Third and Fourth columns showcase feature removal from the image.

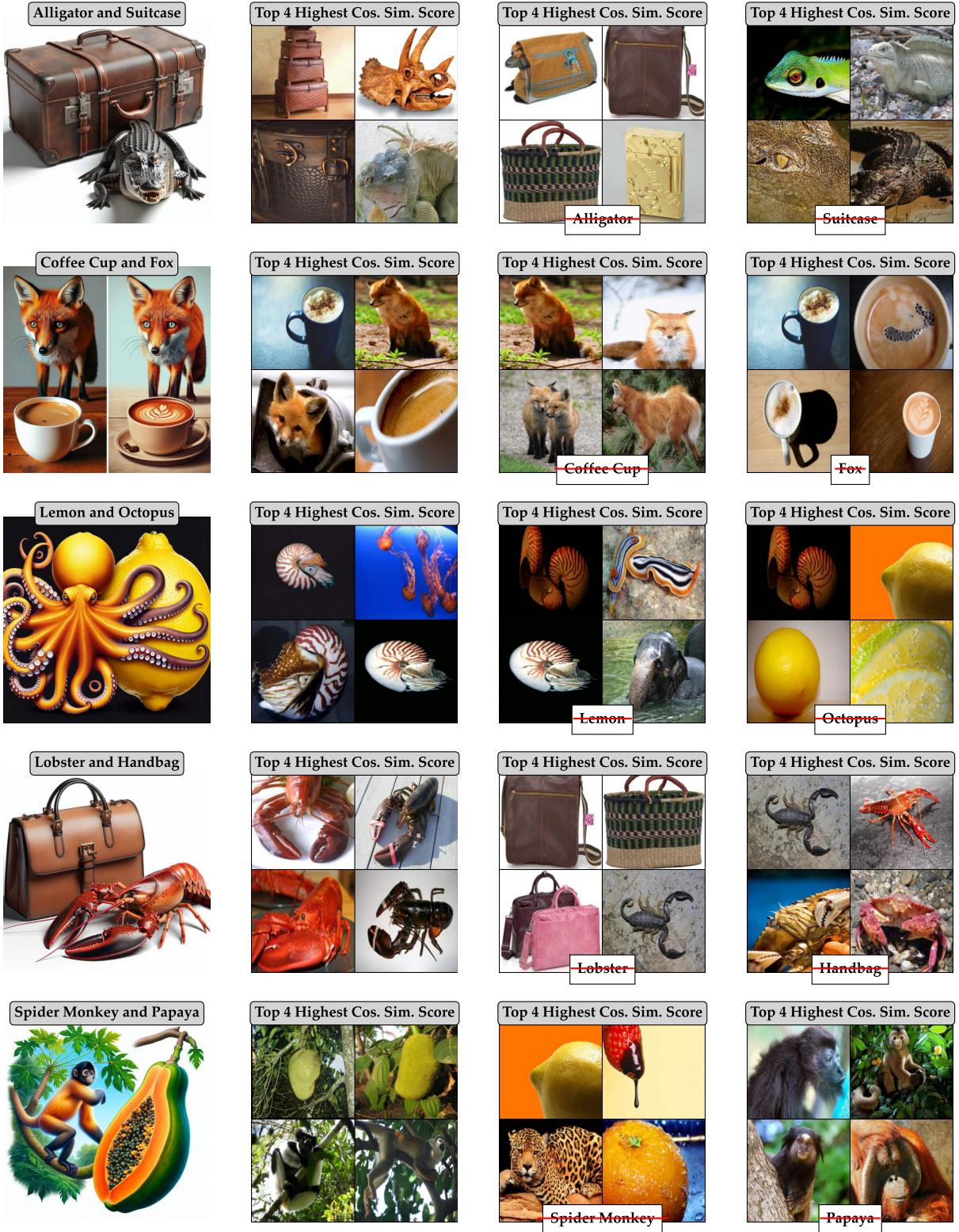


Figure 12. Diagram presenting the decomposition of DALL-E 3 generated images of two random labels from ImageNet. First column presents the generated image of a pair. Second column presents four images from ImageNet with highest Cosine Similarity Score. Third and Fourth columns showcase feature removal from the image.

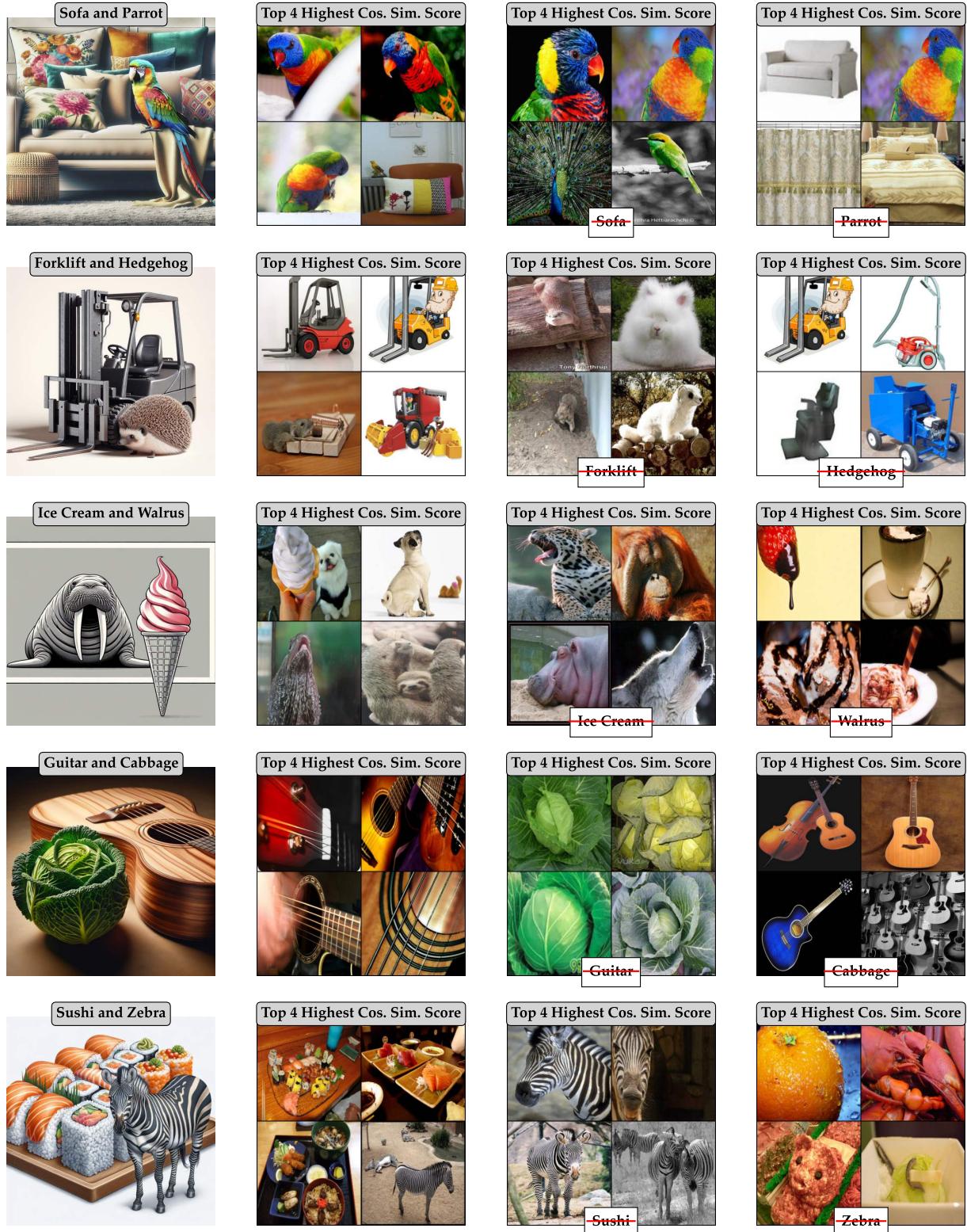


Figure 13. Diagram presenting the decomposition of DALL-E 3 generated images of two random labels from ImageNet. First column presents the generated image of a pair. Second column presents four images from ImageNet with highest Cosine Similarity Score. Third and Fourth columns showcase feature removal from the image.

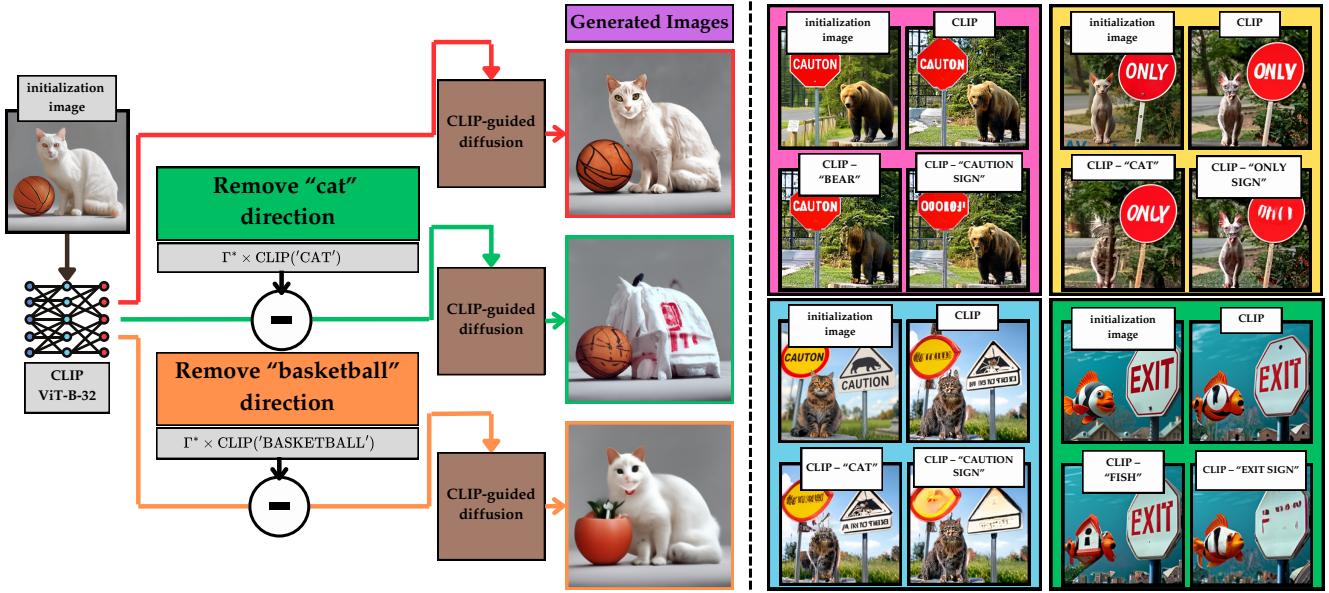


Figure 14. CLIP-guided diffusion process. Starting from an 'initialization image,' generation is guided by CLIP embeddings. The baseline (red arrow) shows unchanged denoising. Adjusted CLIP-guided results (green and orange arrows) show denoised images after removing one of the subjects. Additional clip-guided denoised samples are shown on the right.

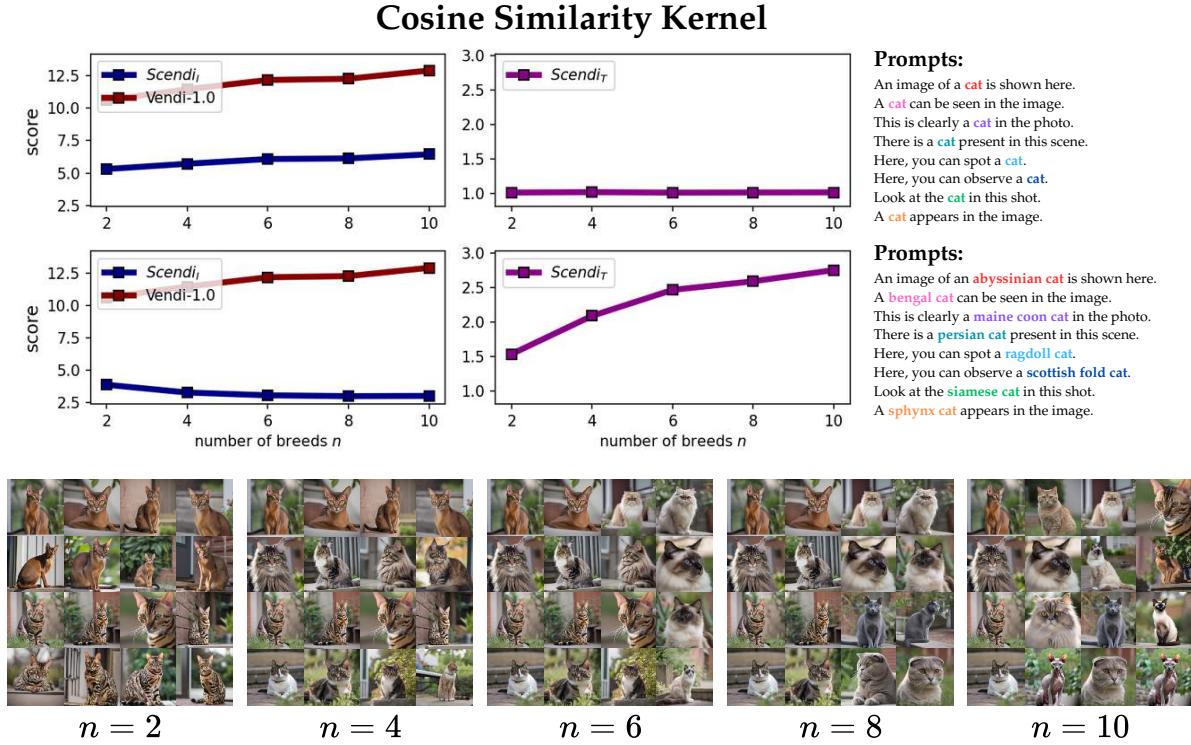


Figure 15. Plots by cancelling out 'cat' and specific cat breed prompts (Cosine Similarity Kernel)

Cosine Similarity Kernel

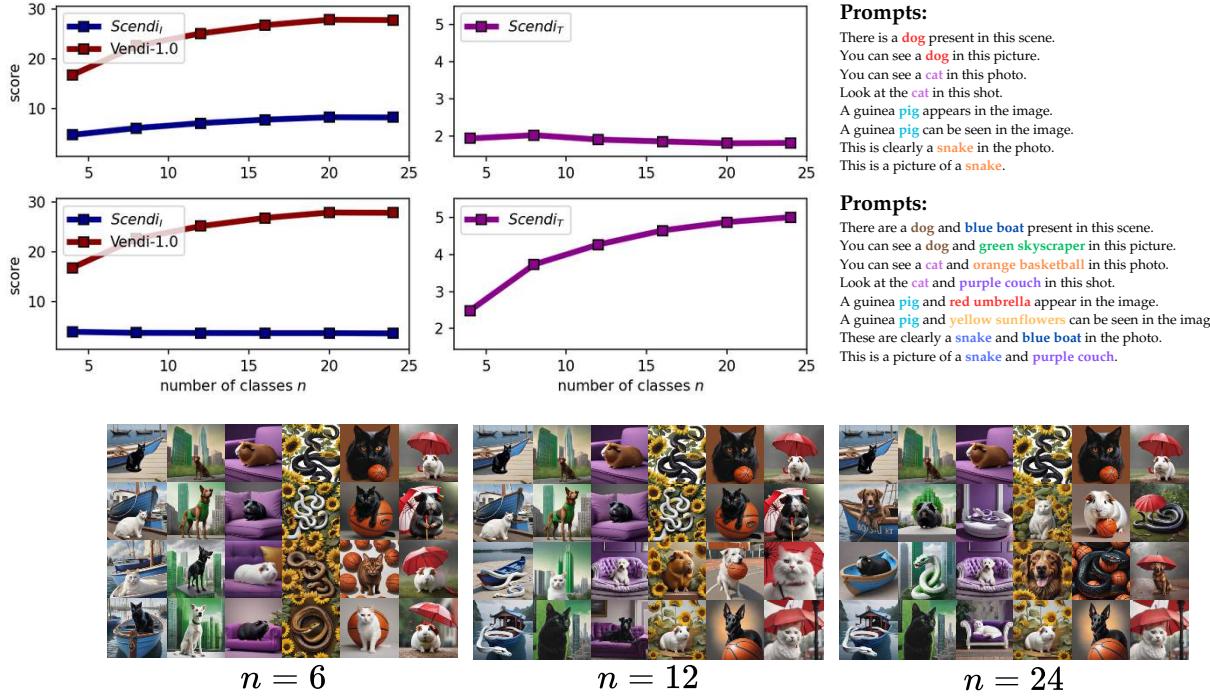


Figure 16. Plots by cancelling out animal name and specific object types prompts (Cosine Sim Kernel)

Gaussian Kernel

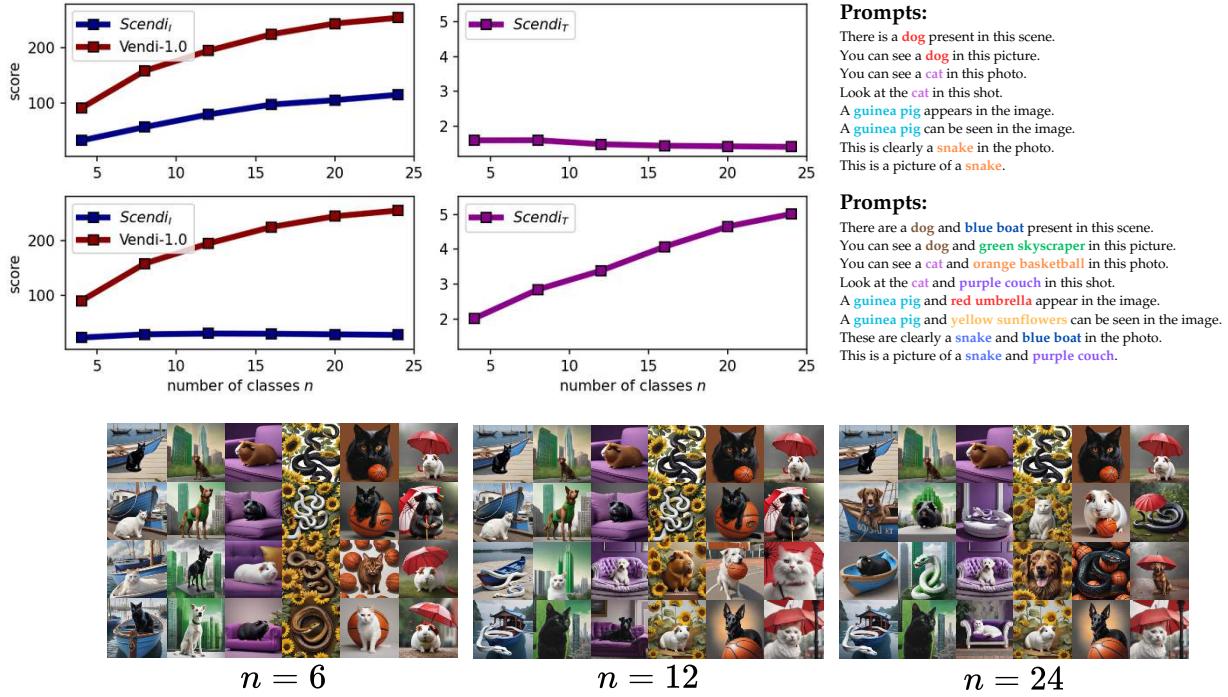


Figure 17. Evaluated Scendi and Vendi scores with Gaussian Kernel on different animals with objects.

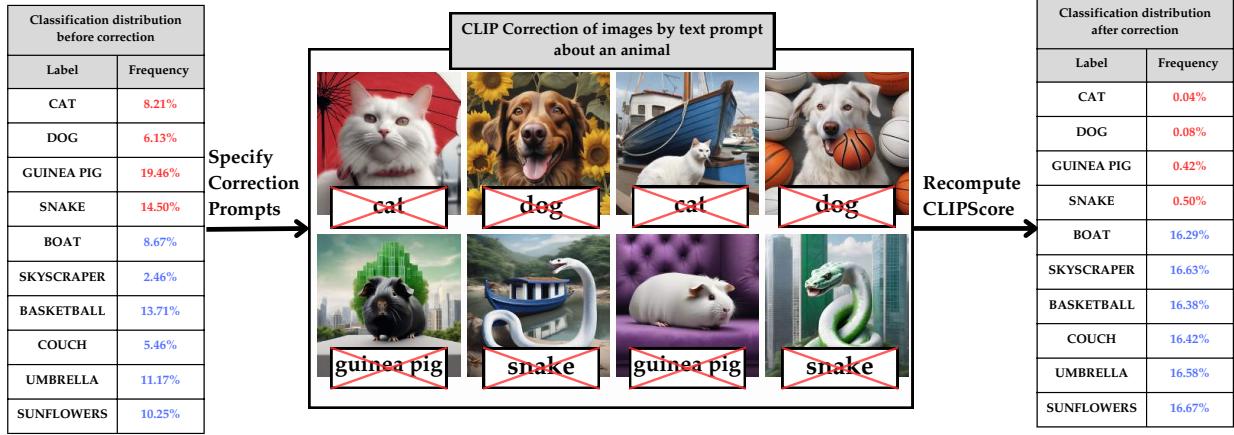


Figure 18. Classification distribution before and after CLIP correction on SDXL [44] generated images of animals with objects in the background

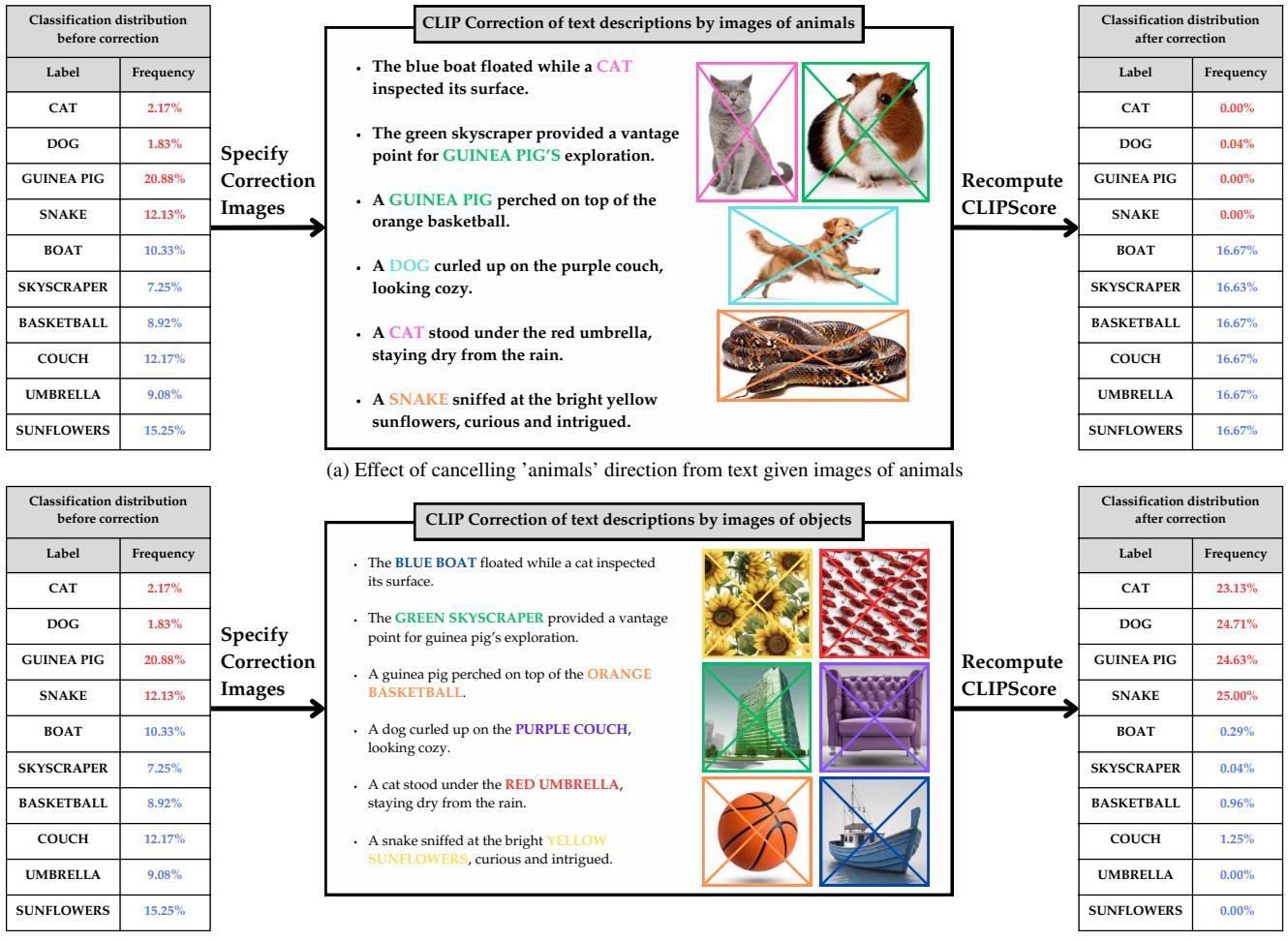


Figure 19. Evaluating the CLIPScore on GPT-4o generated captions of animals with objects

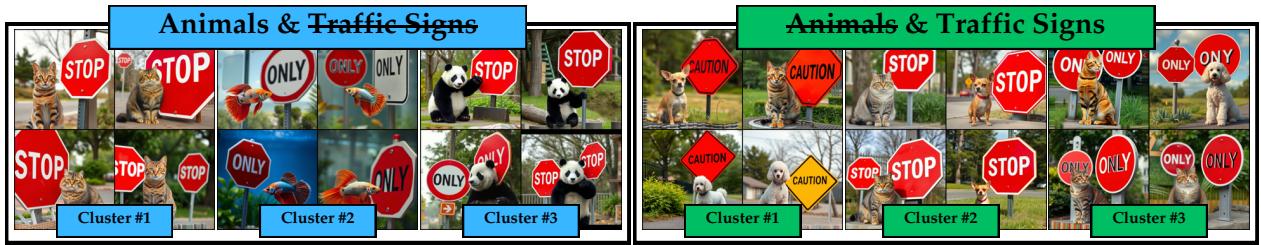


Figure 20. Identified Kernel PCA clusters on the synthetic dataset composed of random animals with traffic signs.

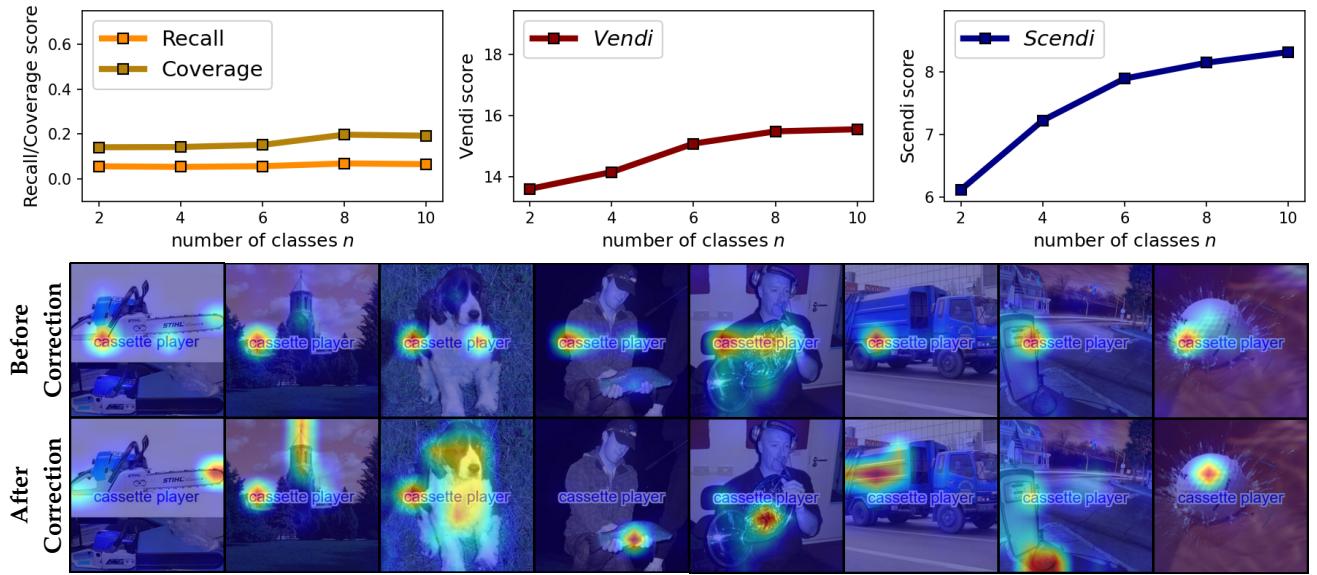


Figure 21. Evaluated Scendi, Vendi, Recall and Coverage scores with Gaussian Kernel on ImageNet with overlayed text.

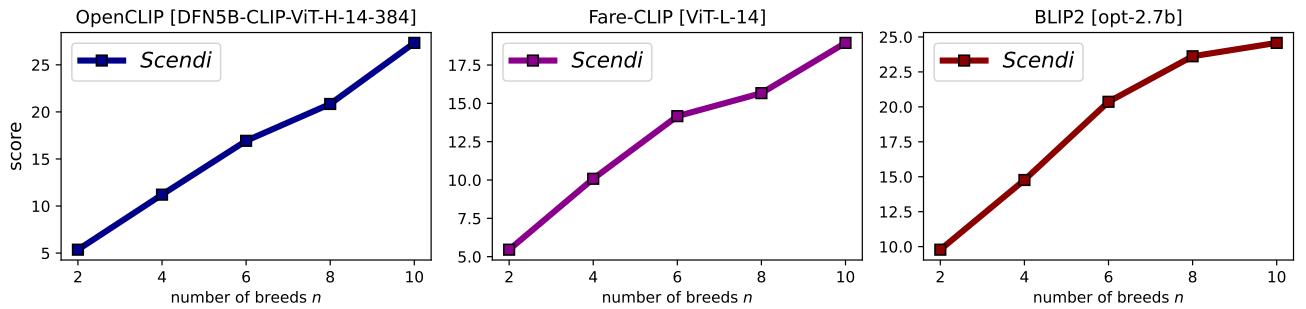


Figure 22. Figure 5's Scendi evaluation of different embeddings: OpenCLIP (left), Robust FAIR-CLIP (middle), BLIP2 (right)

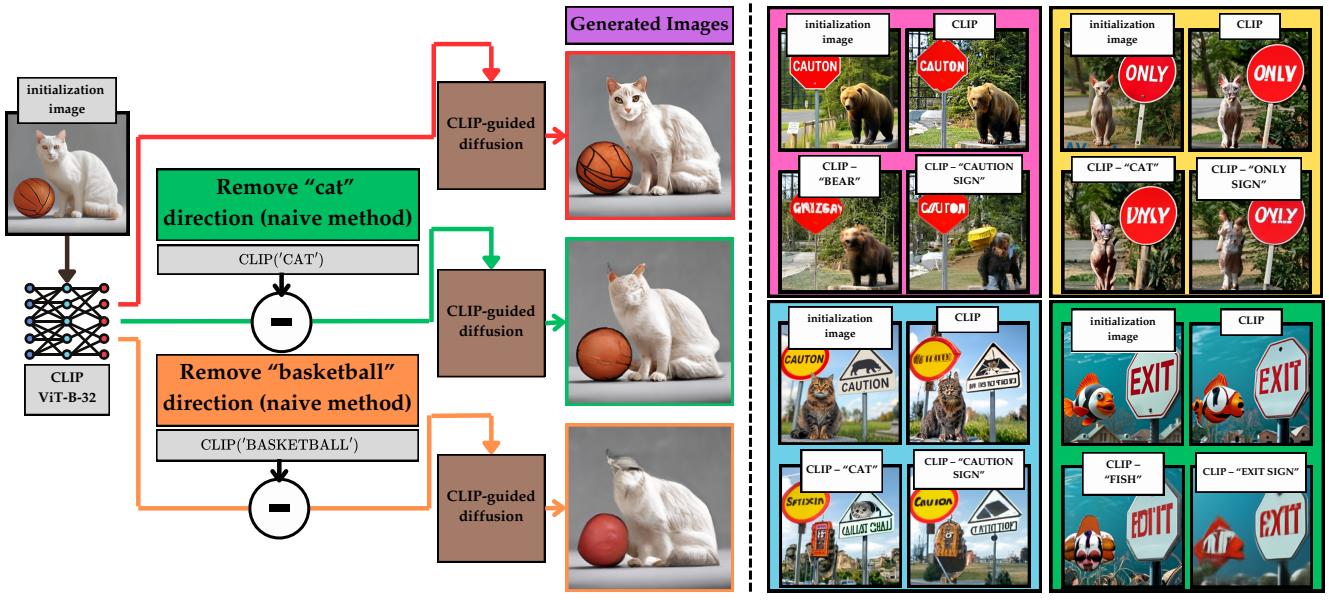


Figure 23. CLIP-guided diffusion process with Naive text embedding cancellation. Starting from an 'initialization image,' generation is guided by CLIP embeddings. The baseline (red arrow) shows unchanged denoising. Naive adjusted CLIP-guided results (green and orange arrows) show image embeddings after subtracting the text CLIP embedding.

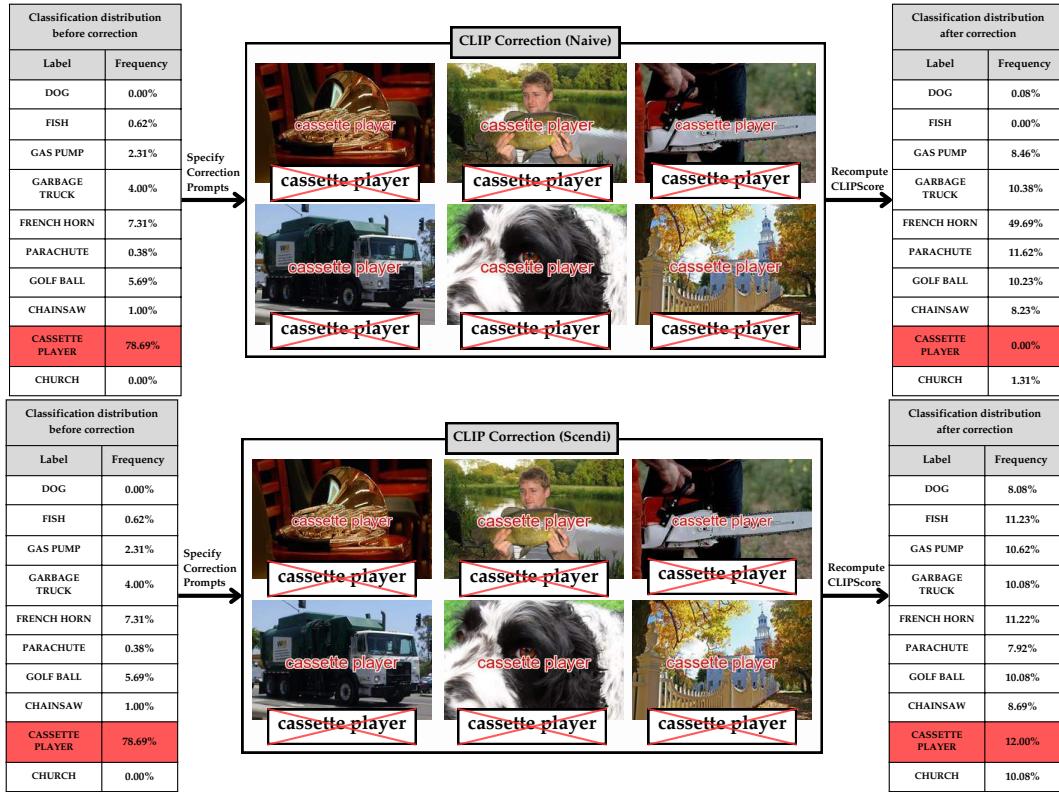


Figure 24. Effect of removing encoded "cassette player" text on top of ImageNet samples. Top figure represents naive method and bottom figure represents SC method.

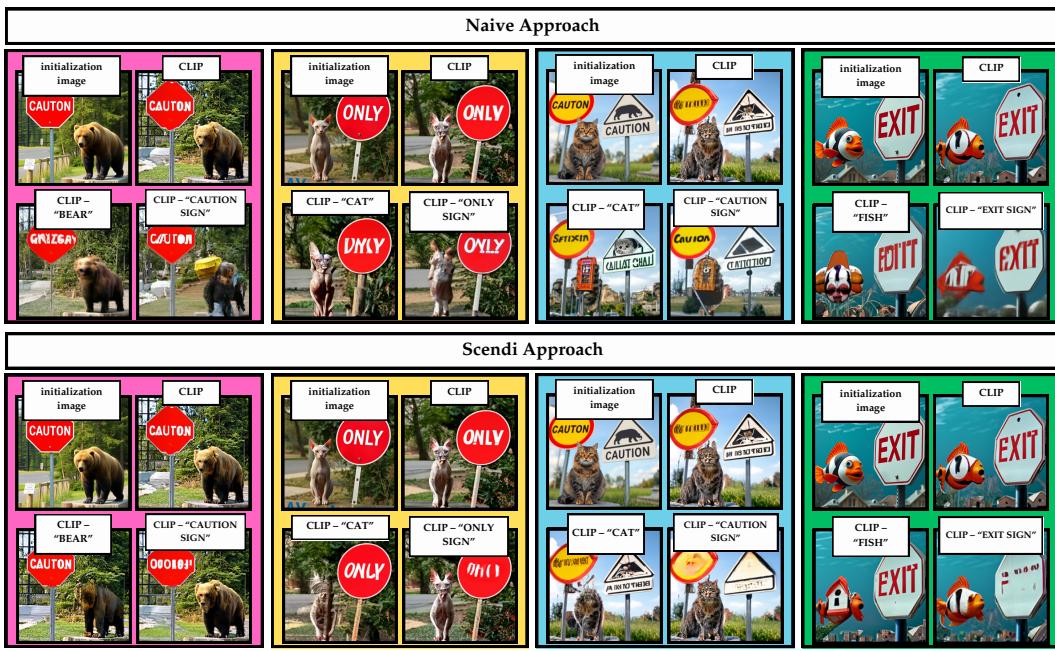


Figure 25. Comparison of samples generated by naive and SC-based decompositions of CLIP.

Kernel-PCA Clustering



Kernel-PCA Clustering after Scendi correction



Kernel-PCA Clustering after Naive correction

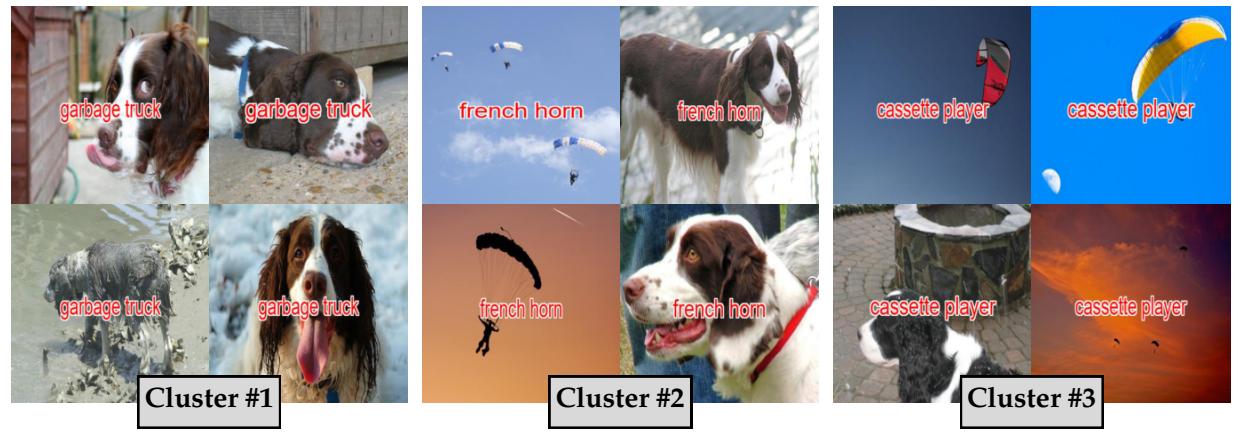


Figure 26. Kernel PCA clusters before and after CLIP correction on the captioned ImageNet dataset, comparing the Scendi decomposition approach with the naive approach.

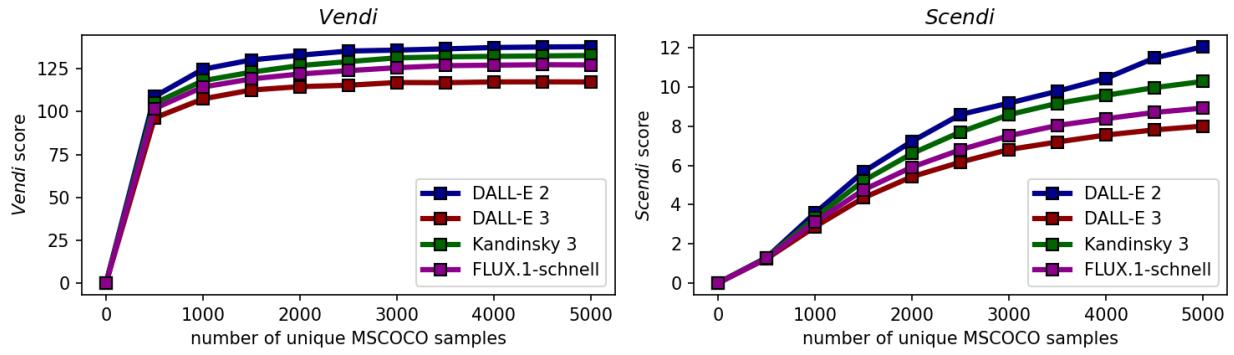


Figure 27. Comparison of different text-to-image models with *Vendi*-1.0 (generated image diversity) and *Scendi* (image generator diversity).

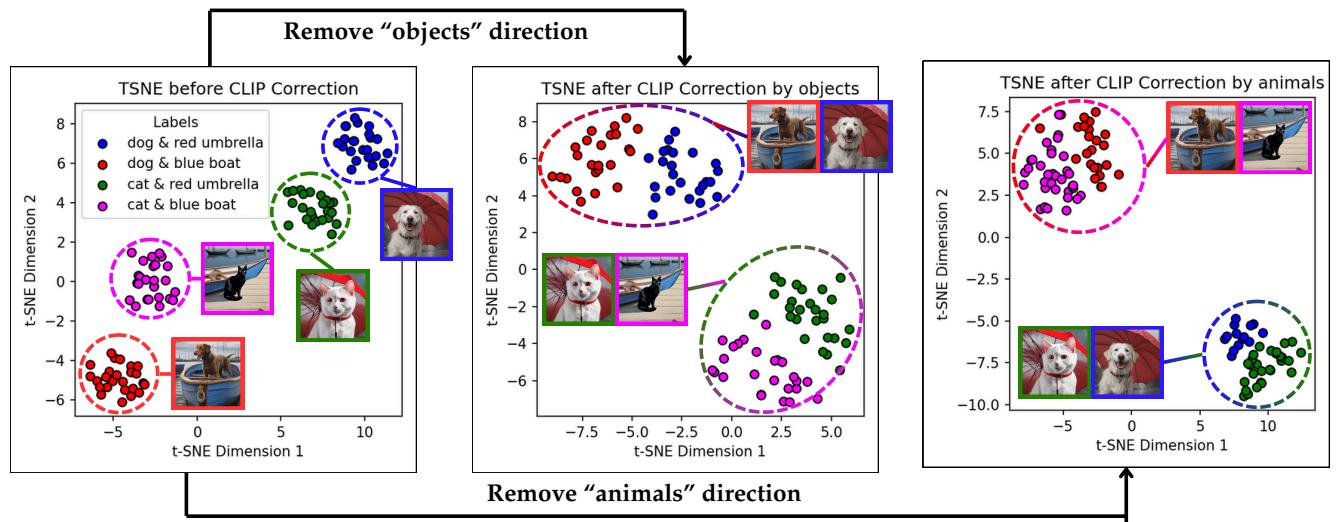


Figure 28. t-SNE plot of animals with objects dataset.