A. Proofs

A.1. Proof of Proposition 1

We aim to solve the optimization problem:

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

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 $\|A\|_F^2 = \mathrm{Tr}(A^\top A)$. Therefore, we have:

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

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:

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

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

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

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

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

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

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

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 $\lambda_1 \geq \cdots \geq \lambda_d$ and orthonormal eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_n$. Following the orthonormality of the eigenvectors, we have the following for every $j \in \{1, \dots, d\}$:

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

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 $H(\mathrm{Mode}) = \sum_{i=1}^{d} \lambda_i \log(1/\lambda_i)$. On the other hand, if an adversary has the side knowledge of the text it can correctly predict $\mathrm{Mode} = j$ 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 $Y_{\mathrm{adv}} = e$ as the error, then the conditional entropy will be:

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

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

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

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

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

Note that the above provides our definition of the Schur-Complement-Entropy for the image part $\mathrm{Scendi}_\mathrm{I}$ and the text part $\mathrm{Scendi}_\mathrm{T}$ as follows:

$$\mathrm{Scendi}_I(x_1, \dots, x_n) \;= \sum_{j=1}^{d} \lambda_j^{(\Lambda_I)} \log \frac{\mathrm{Tr}(\Lambda_I)}{\lambda_j^{(\Lambda_I)}} \tag{9}$$

$$\mathrm{Scendi}_T(x_1, \dots, x_n) \;:= \sum_{j=1}^{d} \lambda_j^{(\Lambda_T)} \log \frac{\mathrm{Tr}(\Lambda_T)}{\lambda_j^{(\Lambda_T)}} \tag{10}$$

where $\lambda_j^{(\Lambda_I)}$ denotes the jth eigenvalue of matrix Λ_I and $\mathrm{Tr}(\Lambda_I) = \sum_{j=1}^{d} \lambda_j^{(\Lambda_I)}$ 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 7, we constructed a dataset of animals with traffic signs using FLUX.1-schnell [21]. 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 [19] 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 5) 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 5, we conducted similar experiment with animals and objects dataset in Figure 17. Our resuts 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 overlayed 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 [28]. Because Scendi is compatible with any cross-modal embedding, we evaluated diversity using three additional variants: OpenCLIP [13], FairCLIP, and BLIP2 [25] 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 [39], DALL-E 3 [32], Kandinsky 3 [40], and FLUX.1-schnell [21], tested on 5,000 MSCOCO [27] 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.

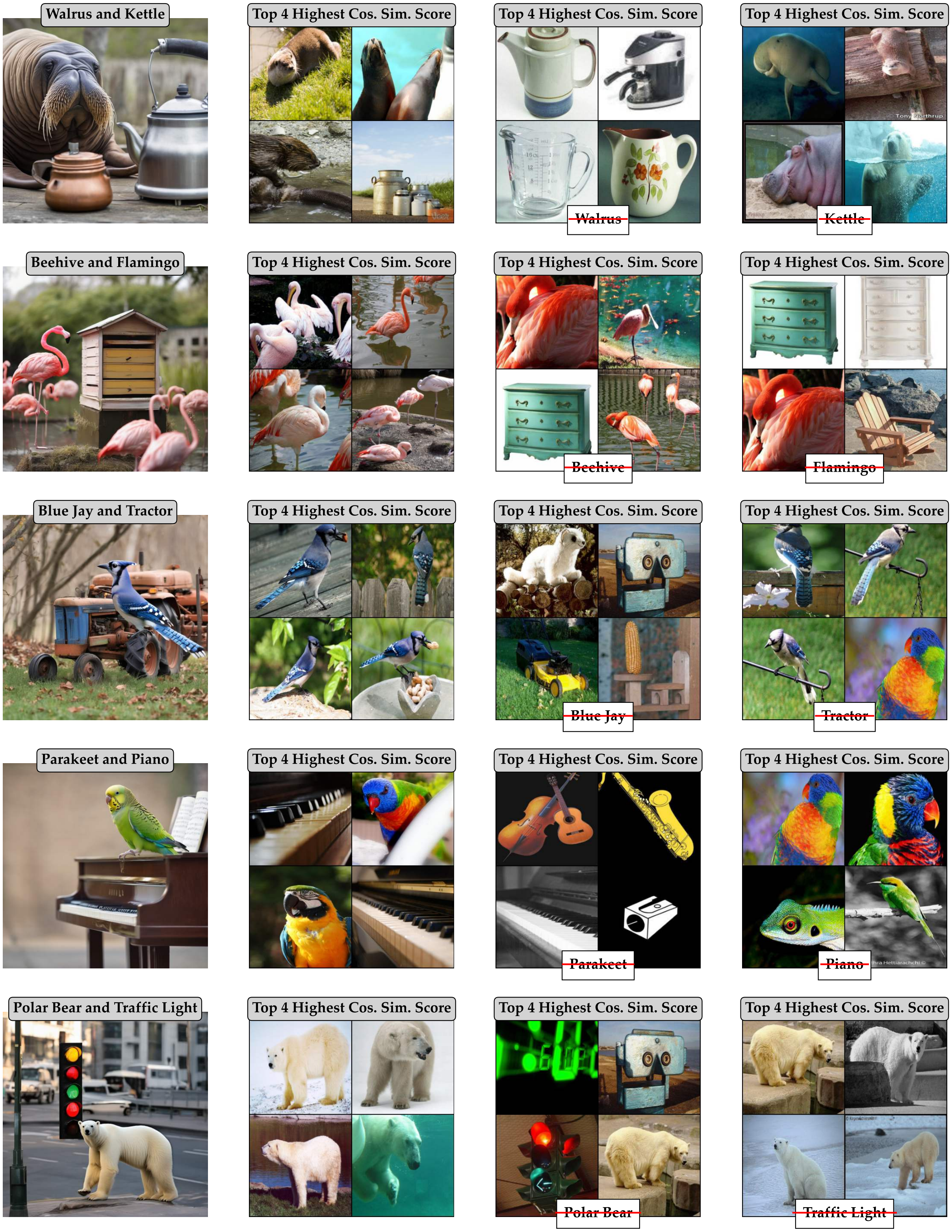

Walrus and Kettle
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
~~Walrus~~
~~Kettle~~
Beehive and Flamingo
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
~~Beehive~~
~~Flamingo~~
Blue Jay and Tractor
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
~~Blue Jay~~
~~Tractor~~
Parakeet and Piano
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
~~Parakeet~~
~~Piano~~
Polar Bear and Traffic Light
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
~~Polar Bear~~
~~Traffic Light~~

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.

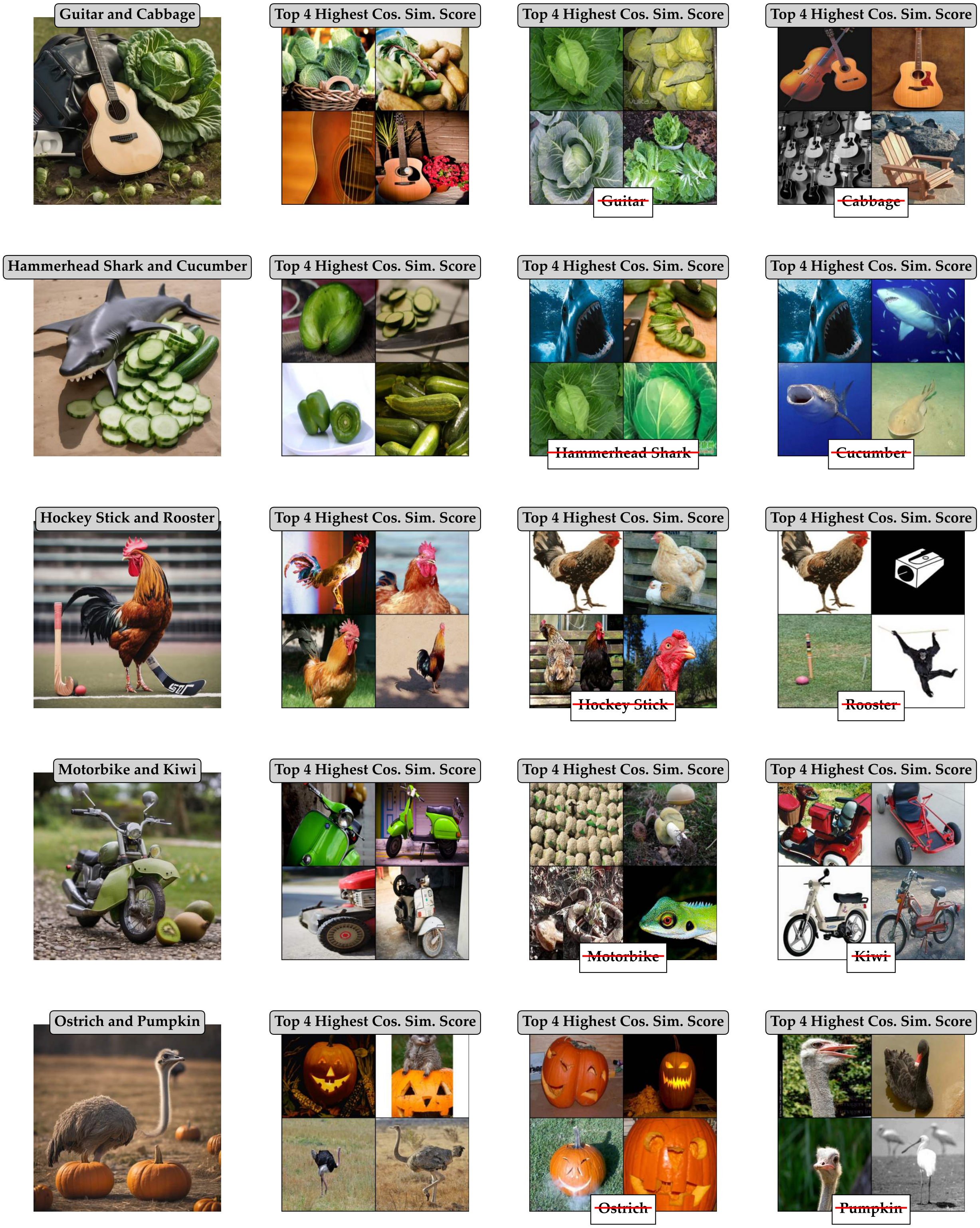

Guitar and Cabbage
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
~~Guitar~~
~~Cabbage~~
Hammerhead Shark and Cucumber
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
~~Hammerhead Shark~~
~~Cucumber~~
Hockey Stick and Rooster
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
~~Hockey Stick~~
~~Rooster~~
Motorbike and Kiwi
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
~~Motorbike~~
~~Kiwi~~
Ostrich and Pumpkin
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
~~Ostrich~~
~~Pumpkin~~

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.

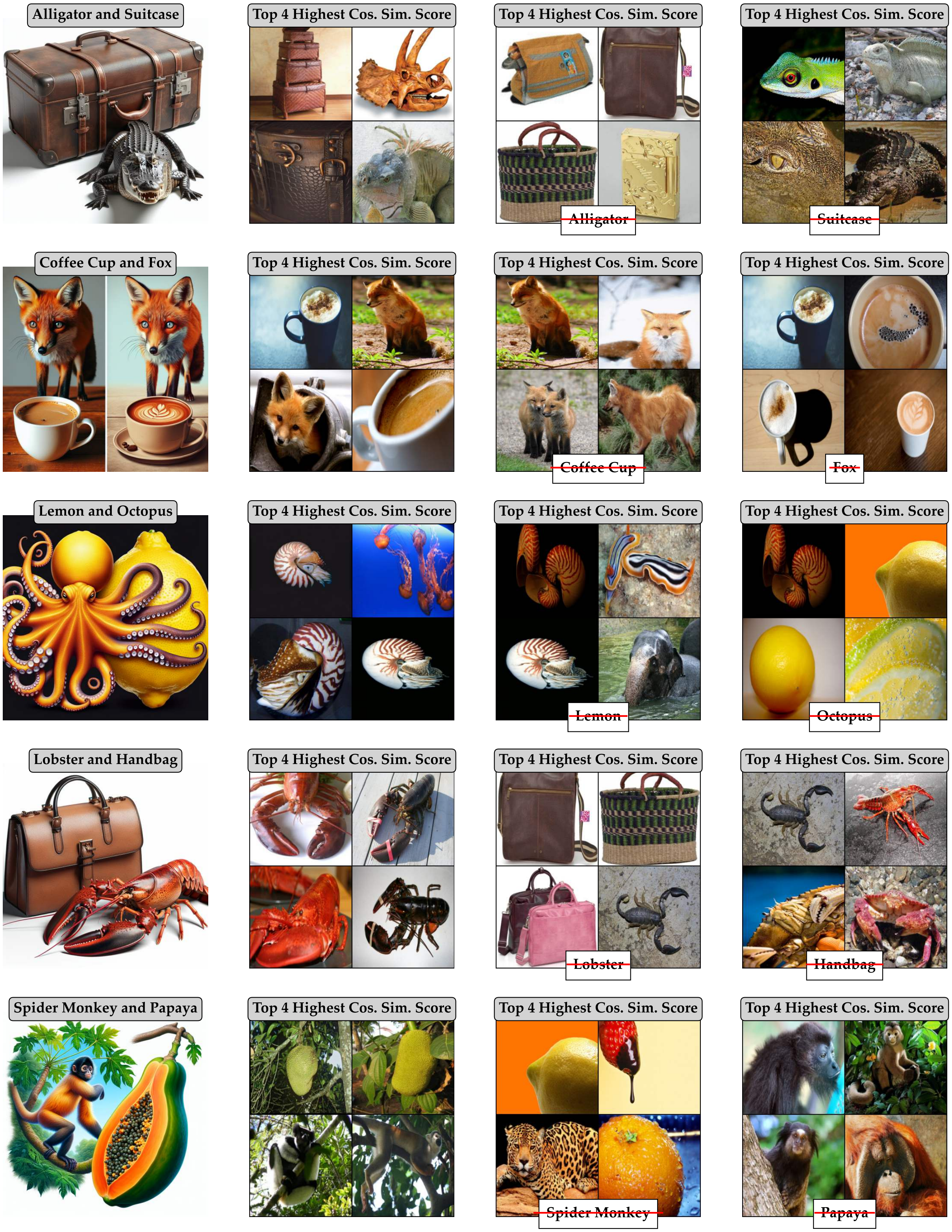

Alligator and Suitcase
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
~~Alligator~~
~~Suitcase~~
Coffee Cup and Fox
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
~~Coffee Cup~~
~~Fox~~
Lemon and Octopus
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
~~Lemon~~
~~Octopus~~
Lobster and Handbag
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
~~Lobster~~
~~Handbag~~
Spider Monkey and Papaya
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
~~Spider Monkey~~
~~Papaya~~

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.

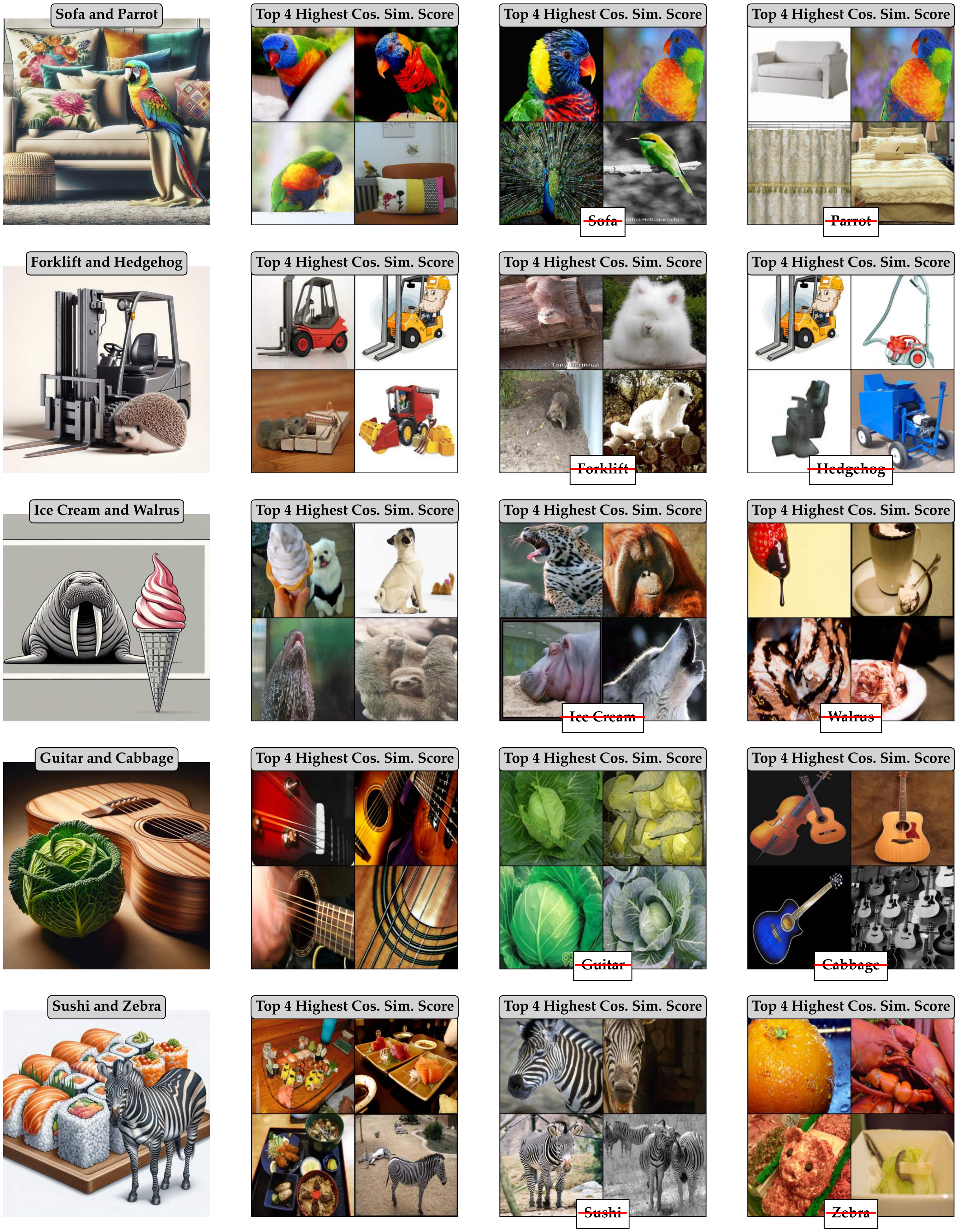

Sofa and Parrot
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
~~Sofa~~
Top 4 Highest Cos. Sim. Score
~~Parrot~~
Forklift and Hedgehog
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
~~Forklift~~
Top 4 Highest Cos. Sim. Score
~~Hedgehog~~
Ice Cream and Walrus
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
~~Ice Cream~~
Top 4 Highest Cos. Sim. Score
~~Walrus~~
Guitar and Cabbage
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
~~Guitar~~
Top 4 Highest Cos. Sim. Score
~~Cabbage~~
Sushi and Zebra
Top 4 Highest Cos. Sim. Score
Top 4 Highest Cos. Sim. Score
~~Sushi~~
Top 4 Highest Cos. Sim. Score
~~Zebra~~

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.

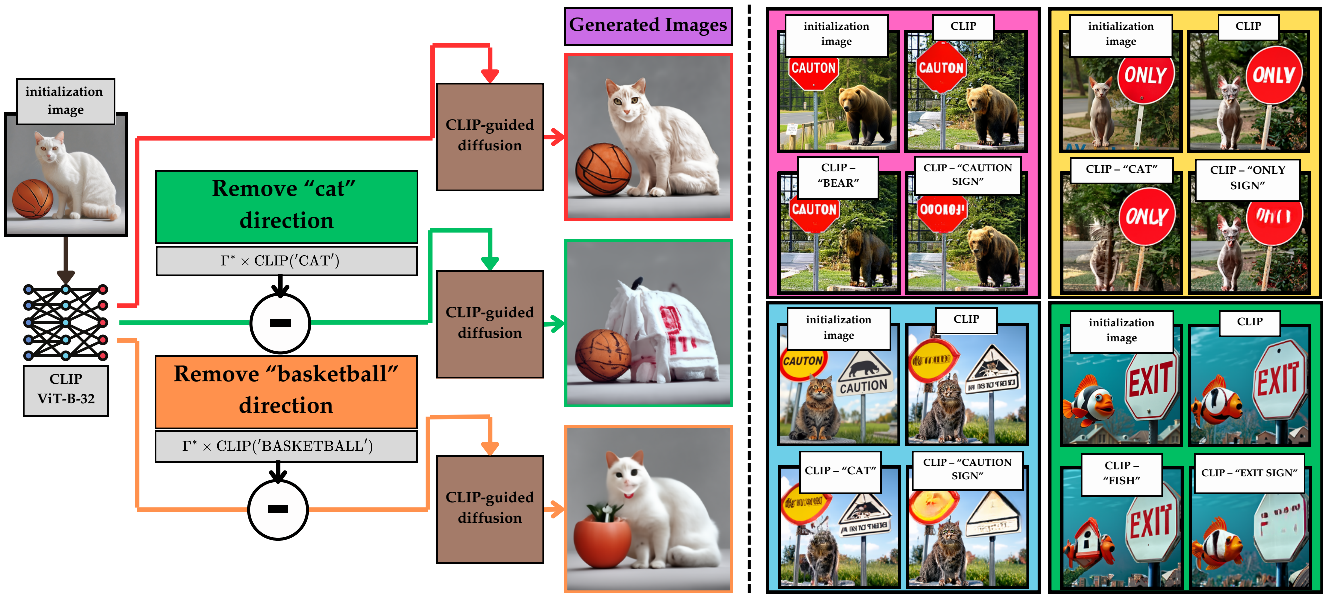
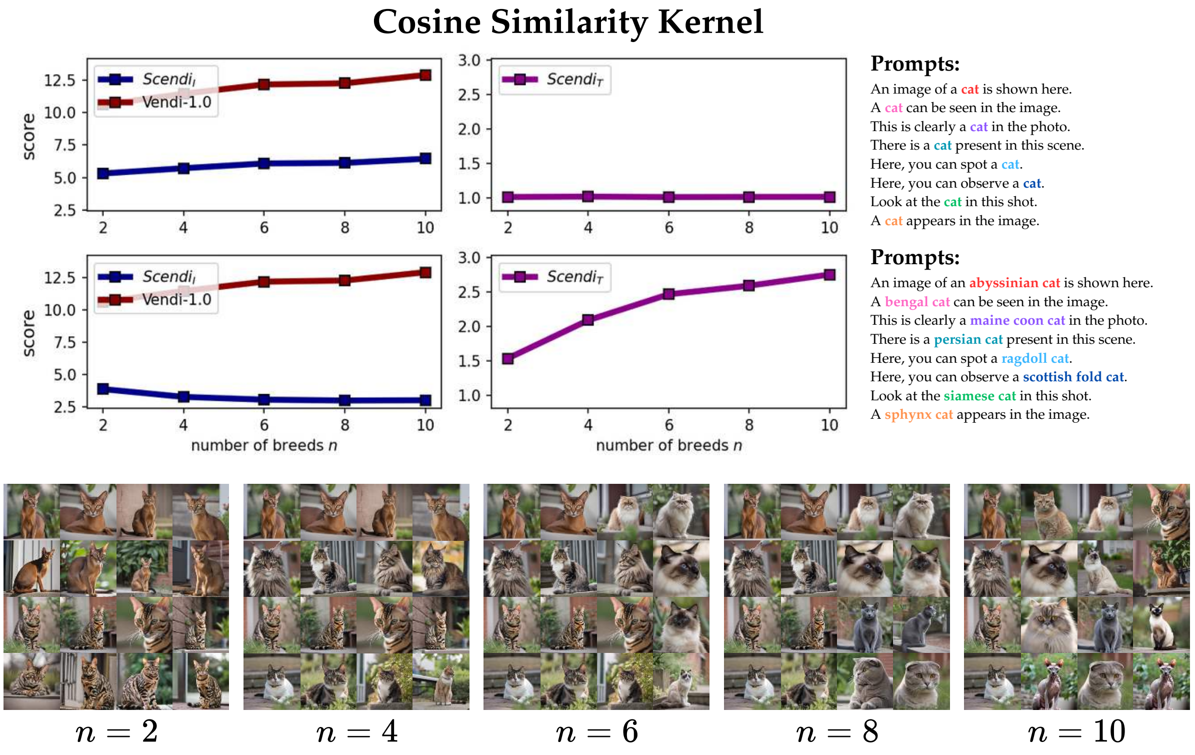


Figure 14. CLIP-guided diffusion process. Starting from an 'initialization image,' generation is guided by CLIP embeddings. The baseline (red arrow) shows unchanged denoising. Adjusted CLIP-guided results (green and orange arrows) show denoised images after removing one of the subjects. Additional clip-guided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



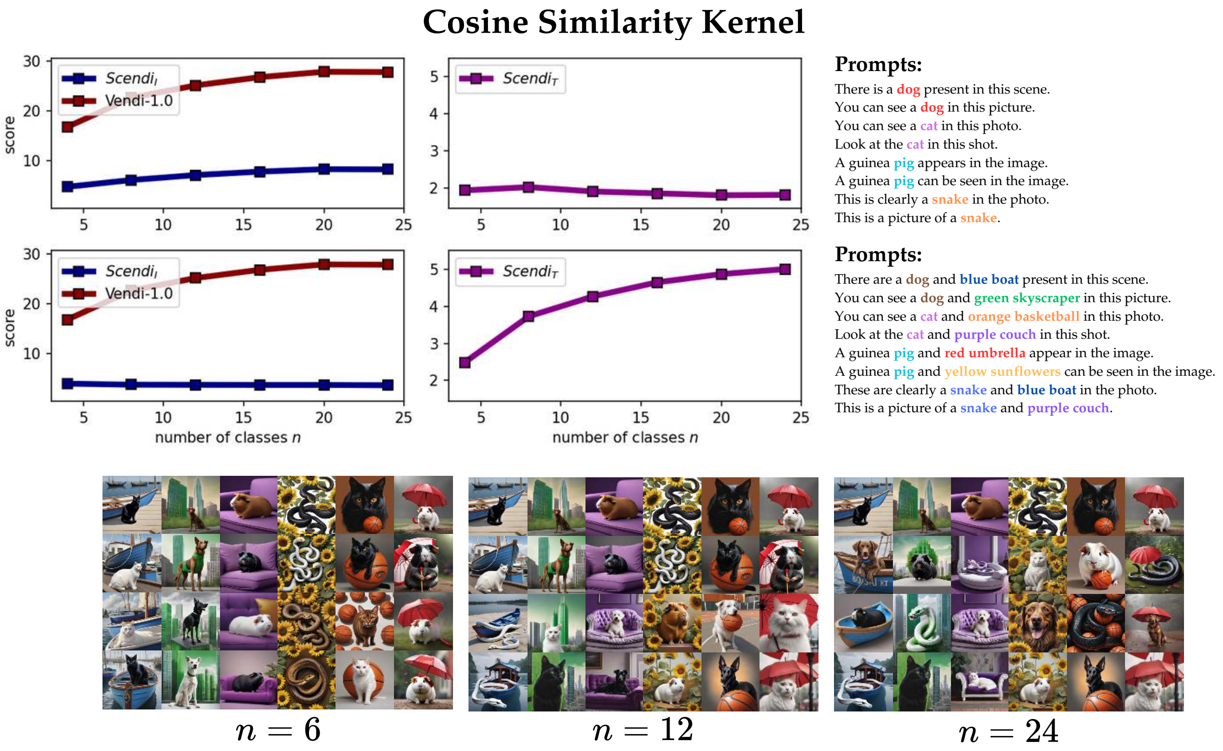
Prompts:

An image of a cat is shown here.
A cat can be seen in the image.
This is clearly a cat in the photo.
There is a cat present in this scene.
Here, you can spot a cat.
Here, you can observe a cat.
Look at the cat in this shot.
A cat appears in the image.

Prompts:

An image of an abyssinian cat is shown here.
A bengal cat can be seen in the image.
This is clearly a maine coon cat in the photo.
There is a persian cat present in this scene.
Here, you can spot a ragdoll cat.
Here, you can observe a scottish fold cat.
Look at the siamese cat in this shot.
A sphynx cat appears in the image.

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



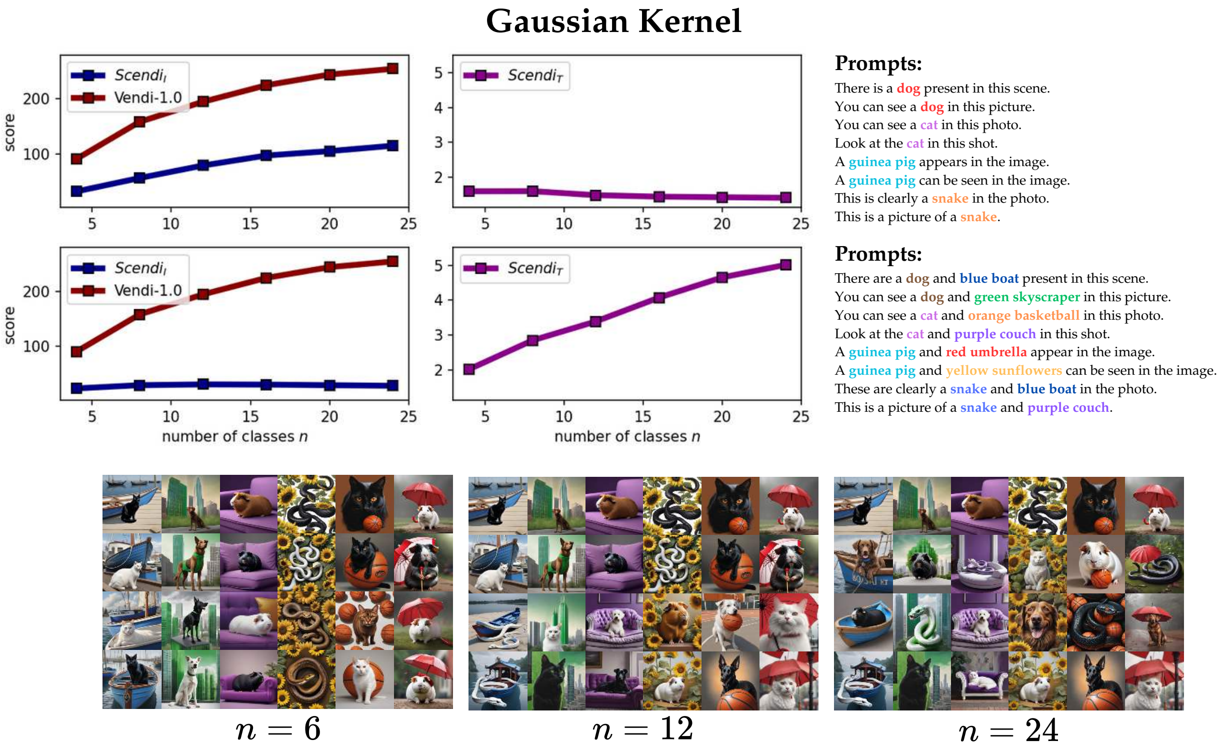
Prompts:

There is a dog present in this scene.
You can see a dog in this picture.
You can see a cat in this photo.
Look at the cat in this shot.
A guinea pig appears in the image.
A guinea pig can be seen in the image.
This is clearly a snake in the photo.
This is a picture of a snake.

Prompts:

There are a dog and blue boat present in this scene.
You can see a dog and green skyscraper in this picture.
You can see a cat and orange basketball in this photo.
Look at the cat and purple couch in this shot.
A guinea pig and red umbrella appear in the image.
A guinea pig and yellow sunflowers can be seen in the image.
These are clearly a snake and blue boat in the photo.
This is a picture of a snake and purple couch.

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



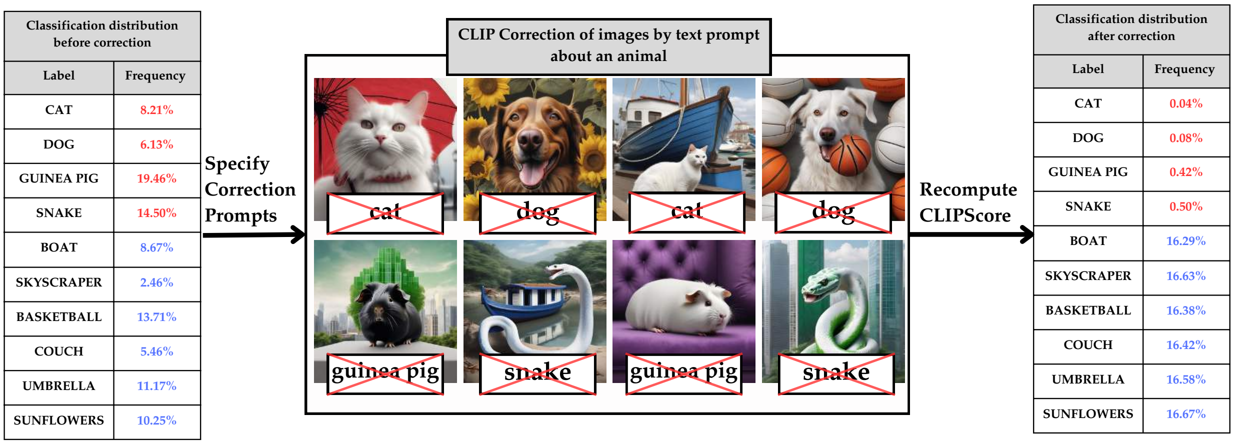
Prompts:

There is a dog present in this scene.
You can see a dog in this picture.
You can see a cat in this photo.
Look at the cat in this shot.
A guinea pig appears in the image.
A guinea pig can be seen in the image.
This is clearly a snake in the photo.
This is a picture of a snake.

Prompts:

There are a dog and blue boat present in this scene.
You can see a dog and green skyscraper in this picture.
You can see a cat and orange basketball in this photo.
Look at the cat and purple couch in this shot.
A guinea pig and red umbrella appear in the image.
A guinea pig and yellow sunflowers can be seen in the image.
These are clearly a snake and blue boat in the photo.
This is a picture of a snake and purple couch.

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



Classification distribution before correction	
Label	Frequency
CAT	8.21%
DOG	6.13%
GUINEA PIG	19.46%
SNAKE	14.50%
BOAT	8.67%
SKYSCRAPER	2.46%
BASKETBALL	13.71%
COUCH	5.46%
UMBRELLA	11.17%
SUNFLOWERS	10.25%

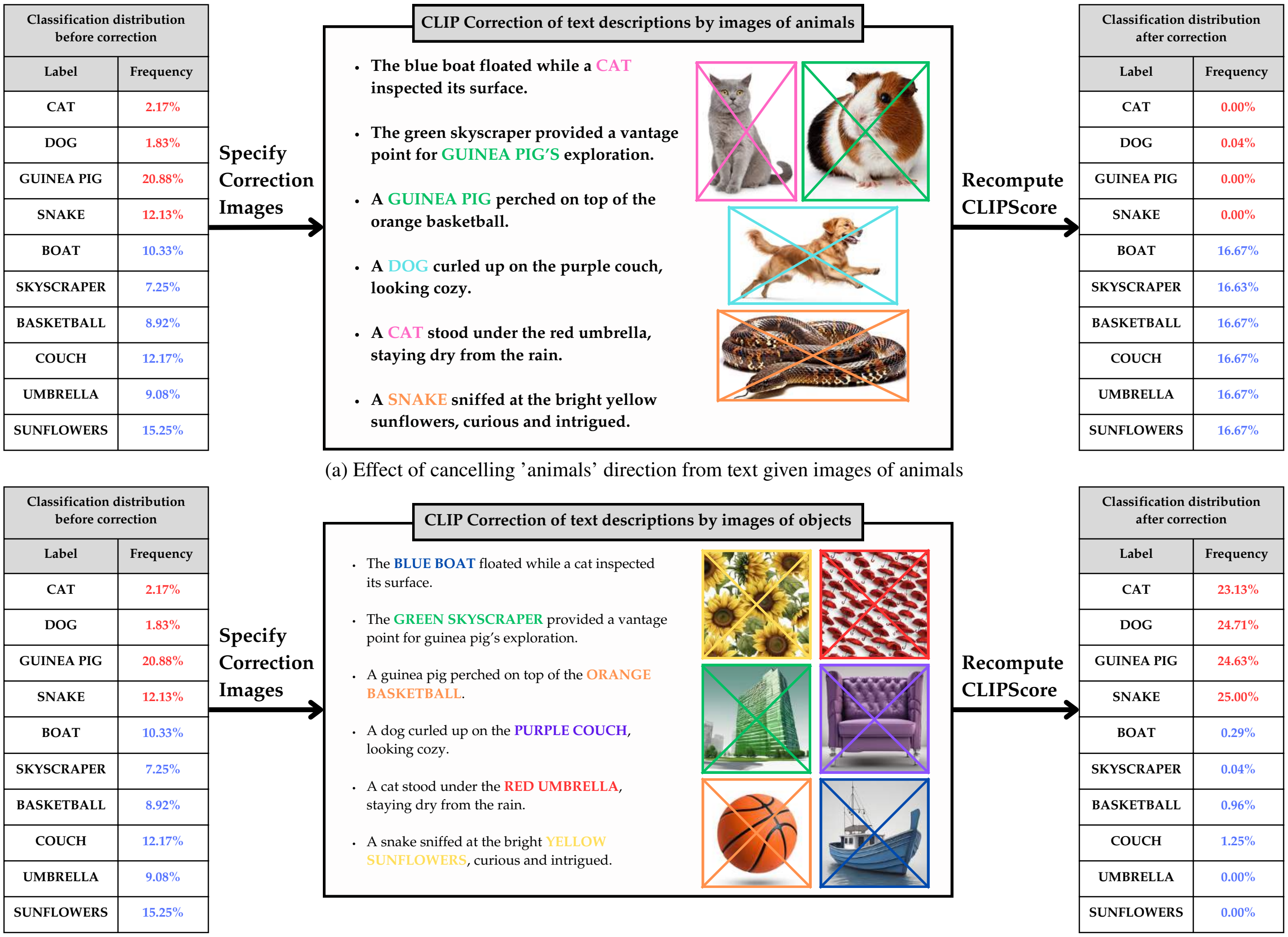
Specify Correction Prompts

CLIP Correction of images by text prompt about an animal

Recompute CLIPScore

Classification distribution after correction	
Label	Frequency
CAT	0.04%
DOG	0.08%
GUINEA PIG	0.42%
SNAKE	0.50%
BOAT	16.29%
SKYSCRAPER	16.63%
BASKETBALL	16.38%
COUCH	16.42%
UMBRELLA	16.58%
SUNFLOWERS	16.67%

Figure 18. Classification distribution before and after CLIP correction on SDXL [36] generated images of animals with objects in the background



Classification distribution before correction	
Label	Frequency
CAT	2.17%
DOG	1.83%
GUINEA PIG	20.88%
SNAKE	12.13%
BOAT	10.33%
SKYSCRAPER	7.25%
BASKETBALL	8.92%
COUCH	12.17%
UMBRELLA	9.08%
SUNFLOWERS	15.25%

Specify Correction Images

CLIP Correction of text descriptions by images of animals

- The blue boat floated while a CAT inspected its surface.
- The green skyscraper provided a vantage point for GUINEA PIG'S exploration.
- A GUINEA PIG perched on top of the orange basketball.
- A DOG curled up on the purple couch, looking cozy.
- A CAT stood under the red umbrella, staying dry from the rain.
- A SNAKE sniffed at the bright yellow sunflowers, curious and intrigued.

Recompute CLIPScore

Classification distribution after correction	
Label	Frequency
CAT	0.00%
DOG	0.04%
GUINEA PIG	0.00%
SNAKE	0.00%
BOAT	16.67%
SKYSCRAPER	16.63%
BASKETBALL	16.67%
COUCH	16.67%
UMBRELLA	16.67%
SUNFLOWERS	16.67%

(a) Effect of cancelling 'animals' direction from text given images of animals

Classification distribution before correction	
Label	Frequency
CAT	2.17%
DOG	1.83%
GUINEA PIG	20.88%
SNAKE	12.13%
BOAT	10.33%
SKYSCRAPER	7.25%
BASKETBALL	8.92%
COUCH	12.17%
UMBRELLA	9.08%
SUNFLOWERS	15.25%

Specify Correction Images

CLIP Correction of text descriptions by images of objects

- The BLUE BOAT floated while a cat inspected its surface.
- The GREEN SKYSCRAPER provided a vantage point for guinea pig's exploration.
- A guinea pig perched on top of the ORANGE BASKETBALL.
- A dog curled up on the PURPLE COUCH, looking cozy.
- A cat stood under the RED UMBRELLA, staying dry from the rain.
- A snake sniffed at the bright YELLOW SUNFLOWERS, curious and intrigued.

Recompute CLIPScore

Classification distribution after correction	
Label	Frequency
CAT	23.13%
DOG	24.71%
GUINEA PIG	24.63%
SNAKE	25.00%
BOAT	0.29%
SKYSCRAPER	0.04%
BASKETBALL	0.96%
COUCH	1.25%
UMBRELLA	0.00%
SUNFLOWERS	0.00%

(b) Effect of cancelling 'objects' direction from text given images of objects

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

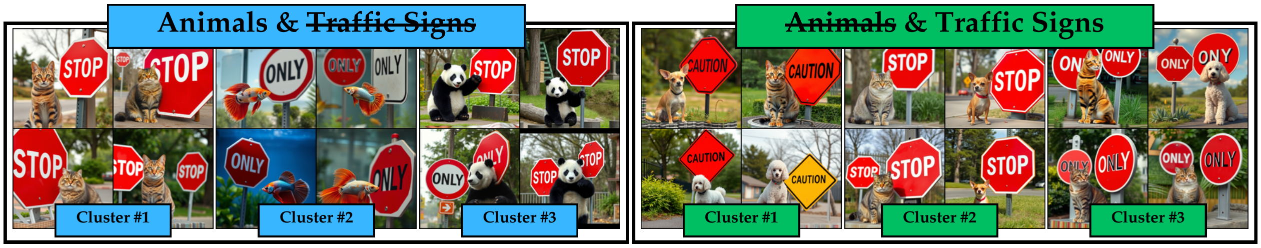


Figure 20. Identified Kernel PCA clusters on the synthetic dataset composed of random animals with traffic signs.

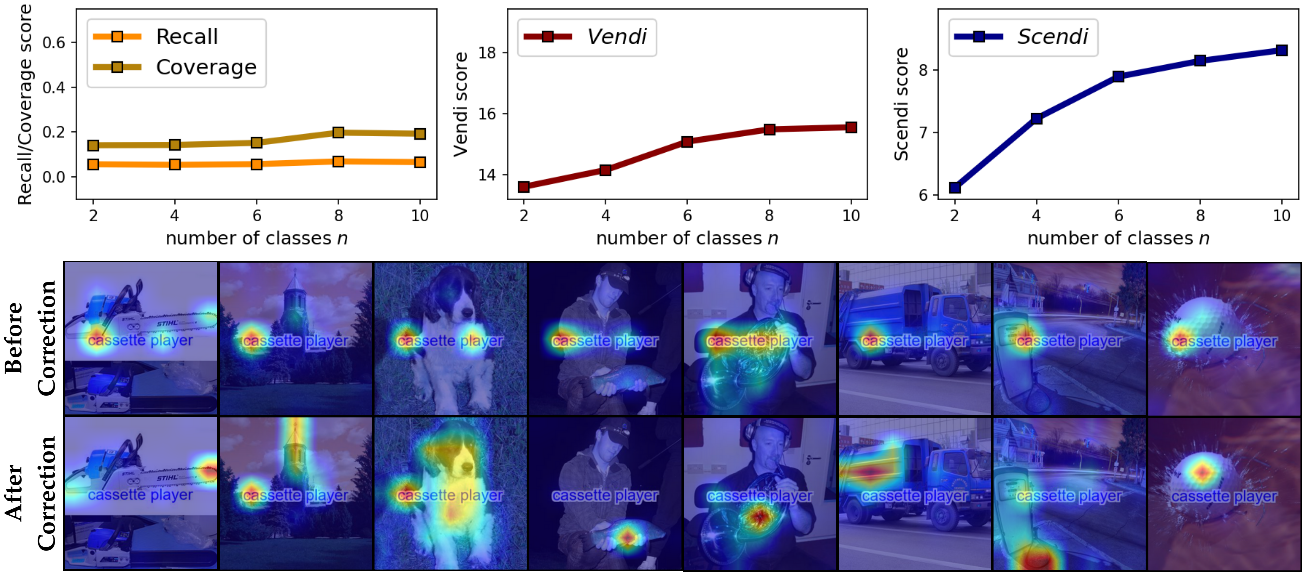


Figure 21. Evaluated Scendi, Vendi, Recall and Coverage scores with Gaussian Kernel on ImageNet with overlayed text.

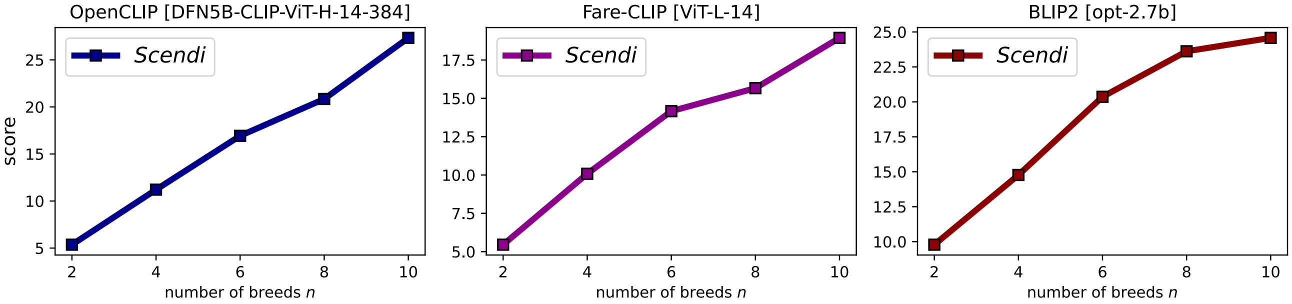


Figure 22. Figure 5's Scendi evaluation of different embeddings: OpenCLIP (left), Robust FAIR-CLIP (middle), BLIP2 (right)

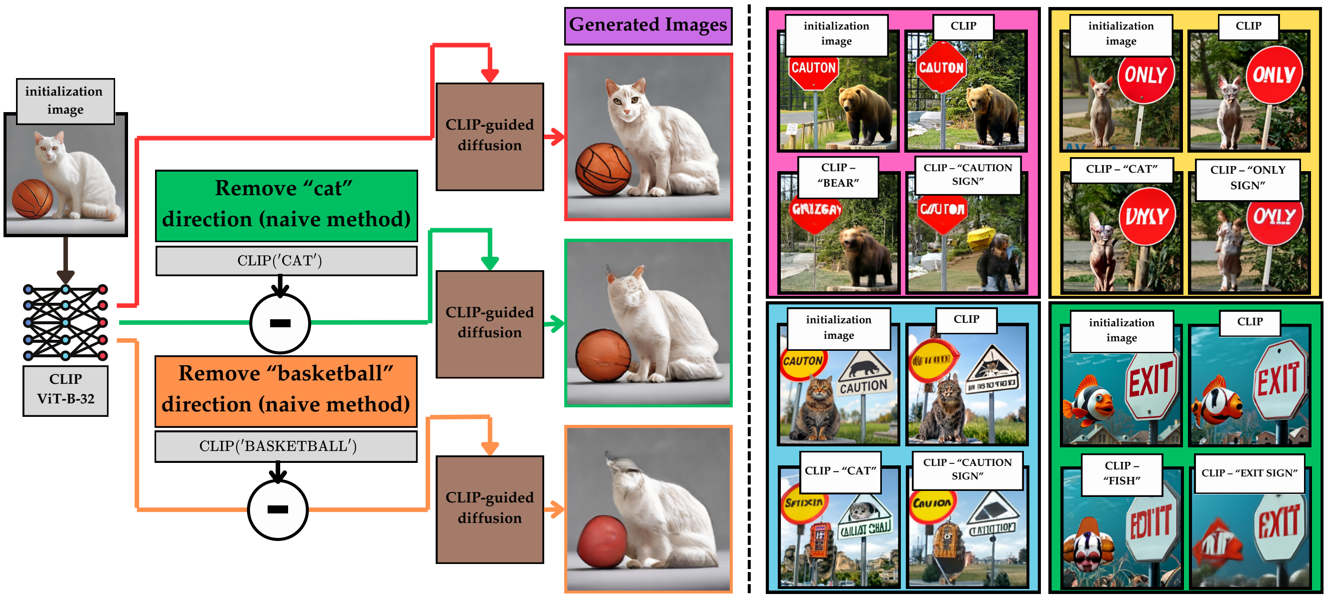
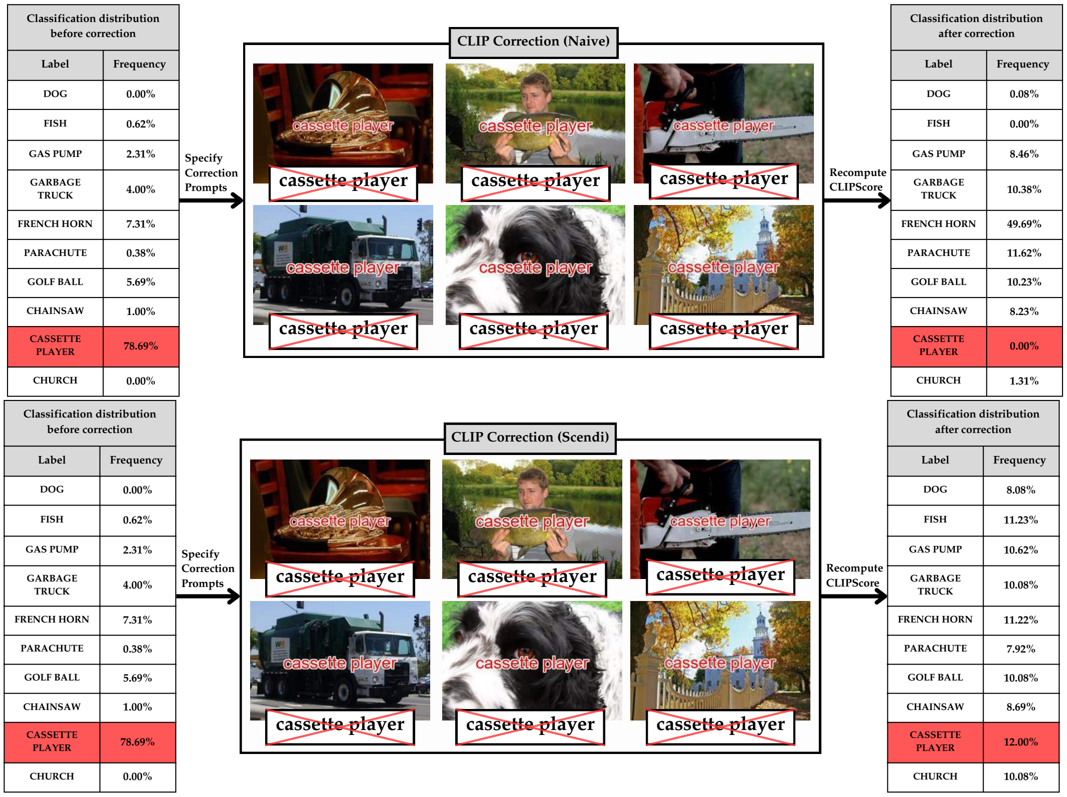


Figure 23. CLIP-guided diffusion process with Naive text embedding cancellation. Starting from an 'initialization image,' generation is guided by CLIP embeddings. The baseline (red arrow) shows unchanged denoising. Naive adjusted CLIP-guided results (green and orange arrows) show image embeddings after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



Classification distribution before correction	
Label	Frequency
DOG	0.00%
FISH	0.62%
GAS PUMP	2.31%
GARBAGE TRUCK	4.00%
FRENCH HORN	7.31%
PARACHUTE	0.38%
GOLF BALL	5.69%
CHAINSAW	1.00%
CASSETTE PLAYER	78.69%
CHURCH	0.00%

Classification distribution after correction	
Label	Frequency
DOG	0.08%
FISH	0.00%
GAS PUMP	8.46%
GARBAGE TRUCK	10.38%
FRENCH HORN	49.69%
PARACHUTE	11.62%
GOLF BALL	10.23%
CHAINSAW	8.23%
CASSETTE PLAYER	0.00%
CHURCH	1.31%

Classification distribution before correction	
Label	Frequency
DOG	0.00%
FISH	0.62%
GAS PUMP	2.31%
GARBAGE TRUCK	4.00%
FRENCH HORN	7.31%
PARACHUTE	0.38%
GOLF BALL	5.69%
CHAINSAW	1.00%
CASSETTE PLAYER	78.69%
CHURCH	0.00%

Classification distribution after correction	
Label	Frequency
DOG	8.08%
FISH	11.23%
GAS PUMP	10.62%
GARBAGE TRUCK	10.08%
FRENCH HORN	11.22%
PARACHUTE	7.92%
GOLF BALL	10.08%
CHAINSAW	8.69%
CASSETTE PLAYER	12.00%
CHURCH	10.08%

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.

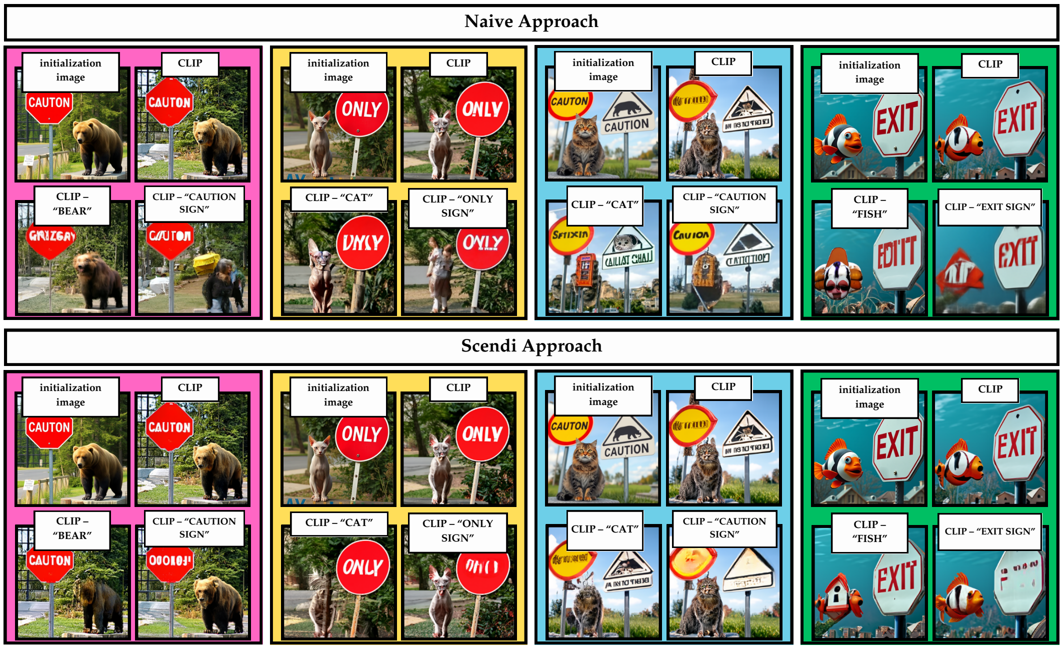


Figure 25. Comparison of samples generated by naive and SC-based decompositions of CLIP.

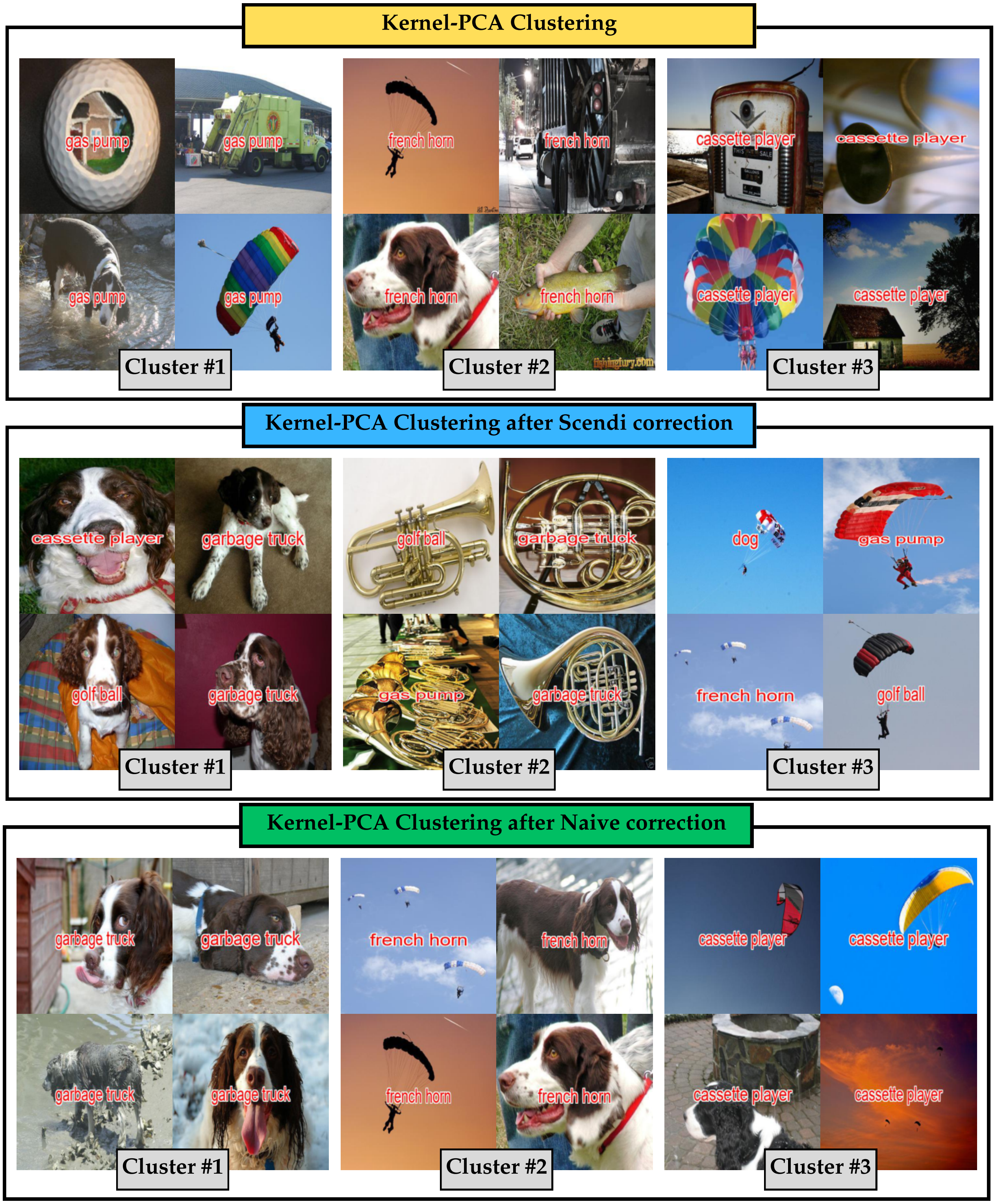


Figure 26. Kernel PCA clusters before and after CLIP correction on the captioned ImageNet dataset, comparing the Scendi decomposition approach with the naive approach.

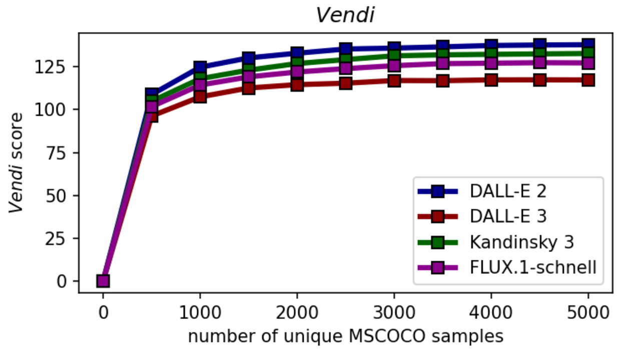


Figure 27. Comparison of different text-to-image models with Vendi-1.0 (generated image diversity) and Scendi (image generator diversity).

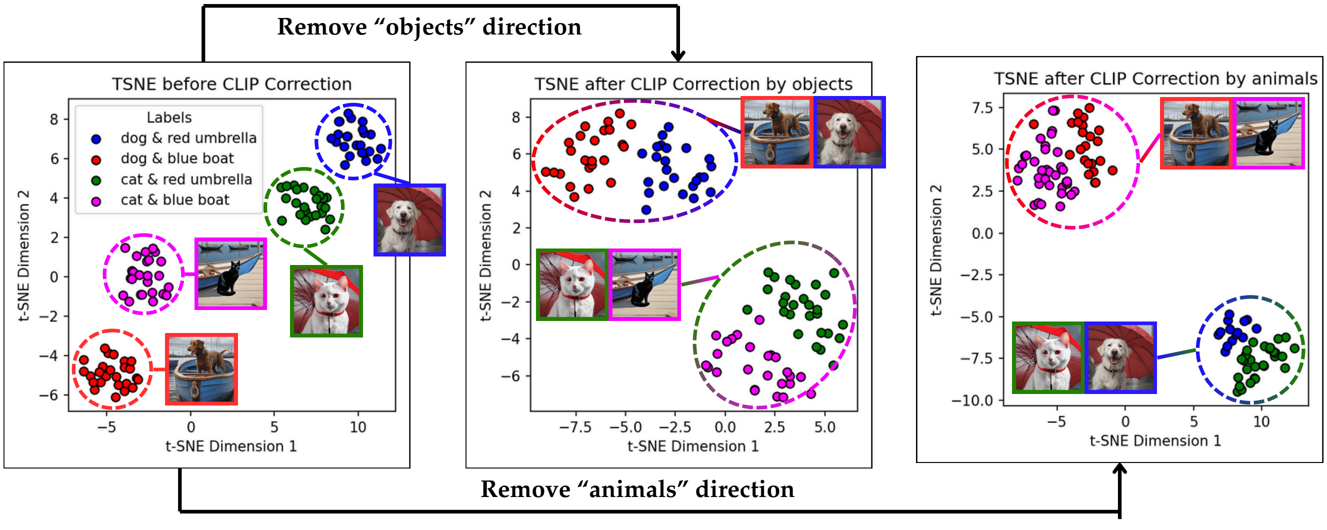


Figure 28. t-SNE plot of animals with objects dataset.