Self-Cross Diffusion Guidance for Text-to-Image Synthesis of Similar Subjects

Supplementary Material

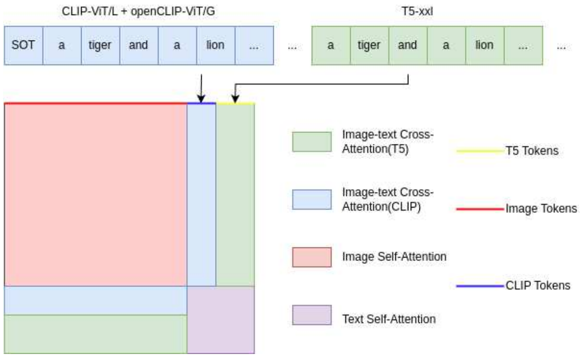


Figure 8. attention maps in multimodal diffusion transformer.

A. More implementation details

Our method is training-free. Following the setting from Attend&Excite [8], we use pseudo-numerical methods [29] and classifier-free guidance [20] to generate images with the original image resolutions of Stable Diffusion models. We apply Self-Cross Diffusion Guidance to the first half(25 steps) of the sampling process(50 steps in total). Empirically, we apply refinements at the 10th and 20th steps of the sampling process with thresholds of 0.2 for the cross-attention response score $S_{\text{cross-attn}}$ and 0.3 for self-cross guidance $S_{\text{self-cross}}$. For each prompt, we generated 65 images with consistent random seeds for each method. Tab. 6 shows the number of prompts and generated images for our experiments.

UNet-based Diffusion models have attention maps of different resolutions including $16 \times 16, 24 \times 24, 32 \times 32$, etc. We chose the attention maps that were found most semantically meaningful. In Stable Diffusion 1, we chose attention maps with a resolution of 16 × 16 [18]. In Stable Diffusion 2, We empirically chose attention maps with a resolution of 24 × 24. Note that for cross-attention maps of sizes larger than 16 × 16, we normalize their sum to 1 so the values in self-attention maps won't be too small in comparison.

For diffusion models based on multimodal diffusion transformers, e.g. Stable Diffusion 3-medium, we replace the conventional cross-attention with the part of attention between text tokens and image tokens(image-text attention) and replace self-attention with the attention between image tokens(image self-attention) [13] [49]. All of these can be extracted from the multimodal diffusion transformer module, as shown in 8. As SD3-medium concatenates the text embeddings from CLIP and T5, we extract the corresponding two image-text attention maps and take the maximum value of the two for each image token (patch) to build the cross-attention map for Self-Cross Diffusion Guidance. Eq.9.

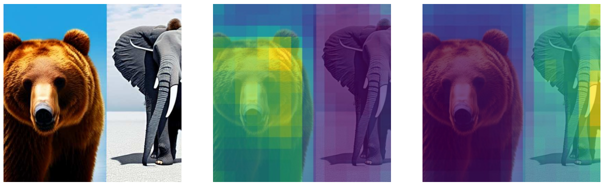


image self-attn "bear" self-attn "elephant"

Figure 9. Artiface of image concatenation by Self-Self guidance between the aggregated self-attention maps.

$$A_i^c[x, y] = \max(A_{i,clip}^c[x, y], A_{i,t5}^c[x, y]) \quad (9)$$

After attention maps were extracted, We averaged attention maps head-wise and layer-wise in our implementation.

Dataset	Animal-Animal	Animal-Obj	Obj-Obj	SSD	TSD
# of prompts	66	144	66	31	21
# of images	4290	9360	4290	2015	1365

Table 6. Number of prompts and images for each dataset.

B. An alternative loss between aggregated self-attention maps

Some readers would suggest an alternative loss to minimize the distance between aggregated self-attention maps(we name it Self-Self guidance in short). Admittedly, this method would achieve comparable results on text-text similarity or TIFA-GPT4o score. However, as shown in Fig. 9, Self-Self guidance easily leads to the artifact of concatenated images. While cross-attention maps correspond to subjects only, aggregated self-attention maps can include background. As Self-Self guidance penalizes any intersection between aggregated self-attention maps, the background is more likely to be separated into two groups resulting in a concatenated image.

C. Question prompt for TIFA-GPT4o

In this section, we detail the implementation of TIFA-GPT4o scores and list the full question prompt used in Fig. 10. GPT4o's answers are translated into True (T) or False (F) values for evaluation.

For Existence (Ext), we calculate the percentage of answers when both Question 1 and Question 3 are True. In other words, the presence of both subjects corresponds to

[2] vanilla SD2.1 usually generates only one subject of the prompt so its 'w/o M' score is high.

You are now an expert to check the faithfulness of the synthesized images. The prompt is ``a {class_A} and a {class_B}''. Based on the image description below, reason and answer the following questions:

1. Is there {class_A} appearing in this image? Give a True/False answer after reasoning.
2. Is the generated {class_A} recognizable and regular (without artifacts) in terms of its shape and semantic structure only? For example, answer False if a two-leg animal has three or more legs, or a two-eye animal has four eyes, or a two-ear animal has one or three ears. Ignore style, object size in comparison to its surroundings. Give a True/False answer after reasoning.
3. Is there {class_B} appearing in this image? Give a True/False answer after reasoning.
4. Is the generated {class_B} recognizable and regular (without artifacts) in terms of its shape and semantic structure only? For example, answer False if a two-leg animal has three or more legs, or a two-eye animal has four eyes, or a two-ear animal has one or three ears. Ignore style, object size in comparison to its surroundings. Give a True/False answer after reasoning.
5. Is the generated content a mixture of {class_A} and {class_B}? An example of mixture is that Sphinx resembles a mixture of a person and a lion. Give a True/False answer after reasoning.

Figure 10. Our Question Prompt for TIFA-GPT4o. Question 1 & 3 ask about the existence of objects; Question 2 & 4 ask about the recognizability of objects; Question 5 asks about whether the generated content resembles some mixture of two categories giving the example of Sphinx as in-context learning.

the intersection of "A appears" and "B appears". Similarly, for Recognizability (Rec), we compute the percentage of answers when both Question 2 and Question 4 are True, ensuring that both subjects are recognizable without artifacts or distortions. For Not a Mixture (w/o M), we compute the percentage of answers where Question 5 is False, reflecting the negation of being a mixture.

D. Unreliability of CLIP scores

The difference in clip scores between INITNO [17], CONFORM [31], and our method is within 1 % as shown in Tab. 8 and 7. However, we found CLIP scores unreliable for evaluating the faithfulness of text prompts and synthetic images for subject mixing. Through experiments, we found that the clip score sometimes can't tell subject mixing, as previous work [22] also pointed out. Fig. 12 shows example images generated by CONFORM [31] and our method with Self-Cross diffusion guidance with the same caption and random seed. For these three pairs of images, Self-Cross diffusion guidance provides visually better images with no subject mixing. However, the corresponding clip scores are much worse than the images generated by CONFORM [31].

Fig. 11 gives a typical example of when the CLIP score is lower for a synthetic image that is more faithful w.r.t. text prompts. Table. 9. Tab. 7 and Tab. 8 show CLIP scores for different methods with multiple datasets respectively. While our method outperforms the original stable diffusion for all datasets, it is on par with or slightly worse than other methods in terms of CLIP scores.

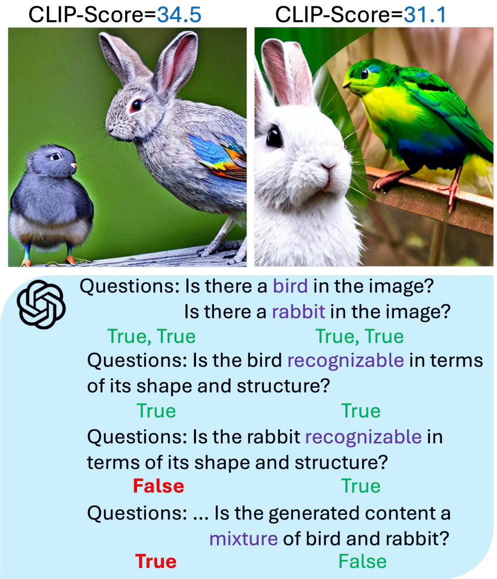
%	Animal-Animal	Animal-Obj	Obj-Obj	SSD-2
SD1.4	31.0	34.3	33.6	31.2
INITNO	33.4	**35.9**	**36.4**	31.7
CONFORM	**33.9**	35.8	35.8	**32.0**
Self-Cross	33.2	35.1	35.9	31.9

Table 7. CLIP Scores with full prompts (↑) for different methods.

%	Animal-Animal	Animal-Obj	Obj-Obj	SSD-2
SD1.4	21.6	24.8	23.9	25.8
INITNO	24.9	**26.8**	**27.1**	26.2
CONFORM	**25.4**	26.7	26.6	**26.6**
Self-Cross	25.1	26.1	26.7	**26.6**

Table 8. CLIP Scores with minimum object prompts (↑).

Additionally, with the same batch of images, the resulting clip score could be different if we simply swap the order of subjects in the prompt during evaluation, as shown in



Questions: Is there a bird in the image?
Is there a rabbit in the image?
True, True   True, True
Questions: Is the bird recognizable in terms of its shape and structure?
True   True
Questions: Is the rabbit recognizable in terms of its shape and structure?
False   True
Questions: ... Is the generated content a mixture of bird and rabbit?
True   False

Figure 11. (Left): image generated by CONFORM [31]; (right): image generated by our approach under the same seed. Left image shows a higher CLIP score. However, there are obvious content mixing issues in the left image, which GPT4o is able to capture with VQA. This is an example that CLIP score is not as reliable as TIFA for checking subject mixing.

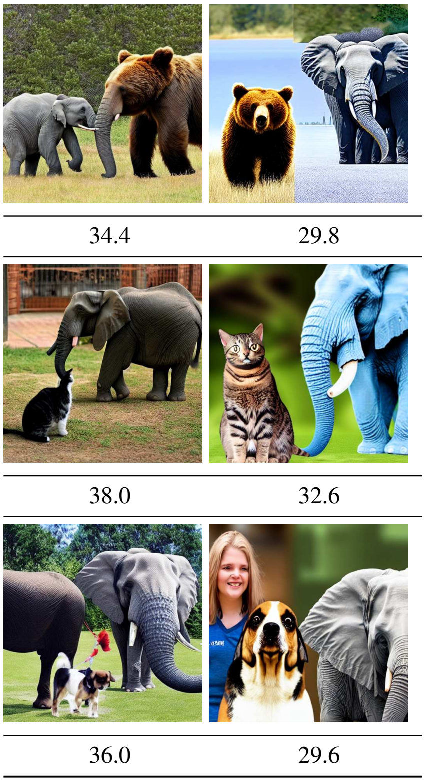


Figure 12. CLIP Scores (↑) for synthetic images generated by CONFORM [31] (left) and our self-cross guidance (right). CLIP scores are unreliable for measuring image quality w.r.t. subject mixing.

Tab. 9. For example, we generated 65 images with the caption "a bear and a turtle". Then we evaluated the clip score with "a bear and a turtle" and "a turtle and a bear" separately. Surprisingly, we found the clip score for the former is 35.1% while the clip score for the latter is only 34.3%.

To conclude, we resort to the more reliable TIFA-GPT4o scores in this paper, which are more correlated with human judgment, as opposed to the popular CLIP scores.

E. More qualitative results

We show more qualitative comparisons in Fig. 14 and Fig. 15. We select four seeds for each prompt and each method to generate.

These samples illustrate that our approach effectively encourages objects to appear as specified in the prompt. For instance, given the prompt "a green backpack and a brown suitcase" in Fig. 14, INITNO [17] sometimes struggles with attribute binding, and CONFORM [31] often fails to include the 'suitcase'. In contrast, our Self-Cross approach successfully addresses these challenges by generating images where both objects are present and correctly aligned with their described attributes. Moreover, our method excels at resolving subject mixing. Images synthesized using our approach typically feature well-disentangled characteristics for each instance. For example, with the prompt "a cat and a rabbit" in Fig. 15, other methods often mix features, such as cat faces with rabbit ears, whereas our Self-Cross method accurately generates distinct and faithful representations of both the cat and the rabbit. Similarly, for the prompt a gray backpack and a green clock, other methods sometimes produce "a green clock-like backpack", blending features improperly. In contrast, our method faithfully adheres to the prompt, producing clear and visually coherent representations of both the backpack and the clock.

%	a bear and a turtle	a bird and a bear	a bird and a rabbit	a bird and a lion
original	35.1	33.9	32.0	33.0
reverse	34.3	34.6	32.7	33.6

Table 9. Inconsistent CLIP Scores ↑ on a set of images with text prompts reversed.

F. Comparison with Attention Refocusing [35]

We further compare our method to Attention Refocusing [35] which depends on external knowledge and model to generate object layout. As shown in Tab. 10, our method demonstrates a significant advantage in Existence (Ext), achieving a 7.56% improvement, and an even more substantial advantage in Recognizability (Rec), with a remarkable 23.34% improvement. These results indicate that our approach more effectively ensures that both subjects appear and are free of artifacts or distortions. Additionally, our method achieves comparable performance in reducing subject mixing (w/o M), demonstrating its robustness in separating distinct features of different subjects within the generated images. Our method also shows an improved text-to-text similarity being 4.5% better, which means our generated images are more faithful to the given prompts.

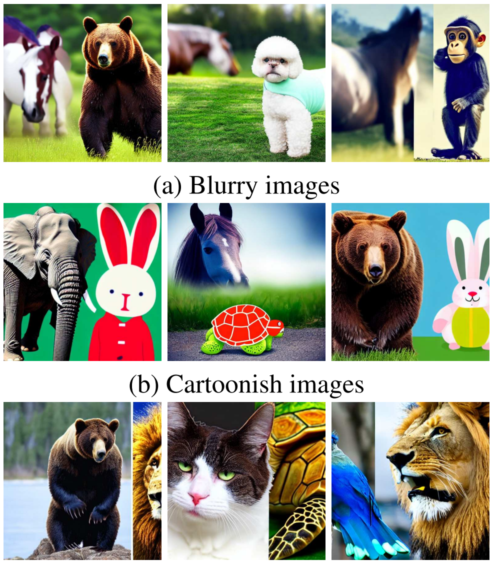
Unlike Attention Refocusing, which relies on a language model to pre-define the layouts, our method operates independently of external knowledge, making it more versatile and applicable to a wider range of scenarios. The superior results in existence and recognizability highlight our approach's ability to generate faithful and high-quality images without relying on external constraints while maintaining competitive performance in mitigating subject mixing.

G. Failure examples and discussion

Except for its success in reducing subject mixing, however, Self-Cross Guidance sometimes generates unsatisfactory images, such as blurry images, cartoons, and images with object-centric problems. These failure cases indicate that the method is not perfect. We show failure examples of our method in Fig. 13. We suspect that the artifact of blur-

Metric (↑)	SD1.4 [41]	Attn-Refocus [35]	Self-Cross (Ours)
Ext	39.51	86.99	**94.55**
Rec	29.70	64.45	**87.79**
w/o M	72.24	**93.80**	92.94
CLIP score	31.0	**33.9**	33.2
Text sim	76.5	79.8	**84.3**

Table 10. Quantitative comparison to Attention Refocusing [35] on Animal-Animal benchmark in terms of TIFA-GPT4o scores [22], CLIP score, and text-to-text similarity (Txt sim) [8]. Attention Refocusing relies on external knowledge by using a language model to pre-define the layout. Our proposed method has a significant advantage for existence (Ext), recognizability (Rec), and text-to-text similarity while reaching a comparable performance on reducing subject mixing (w/o M) and CLIP score.


(a) Blurry images

(b) Cartoonish images

(c) Concatenated subimages.

Figure 13. Our method with self-cross guidance failed in some cases and generated blurry images (a), cartoonish images (b), or concatenated subimages (c).

riness can be addressed by aggregation of attention maps at higher resolution. We also found that previous methods including INITNO [17] and CONFORM [31] may also produce cartoonish or concatenated images.

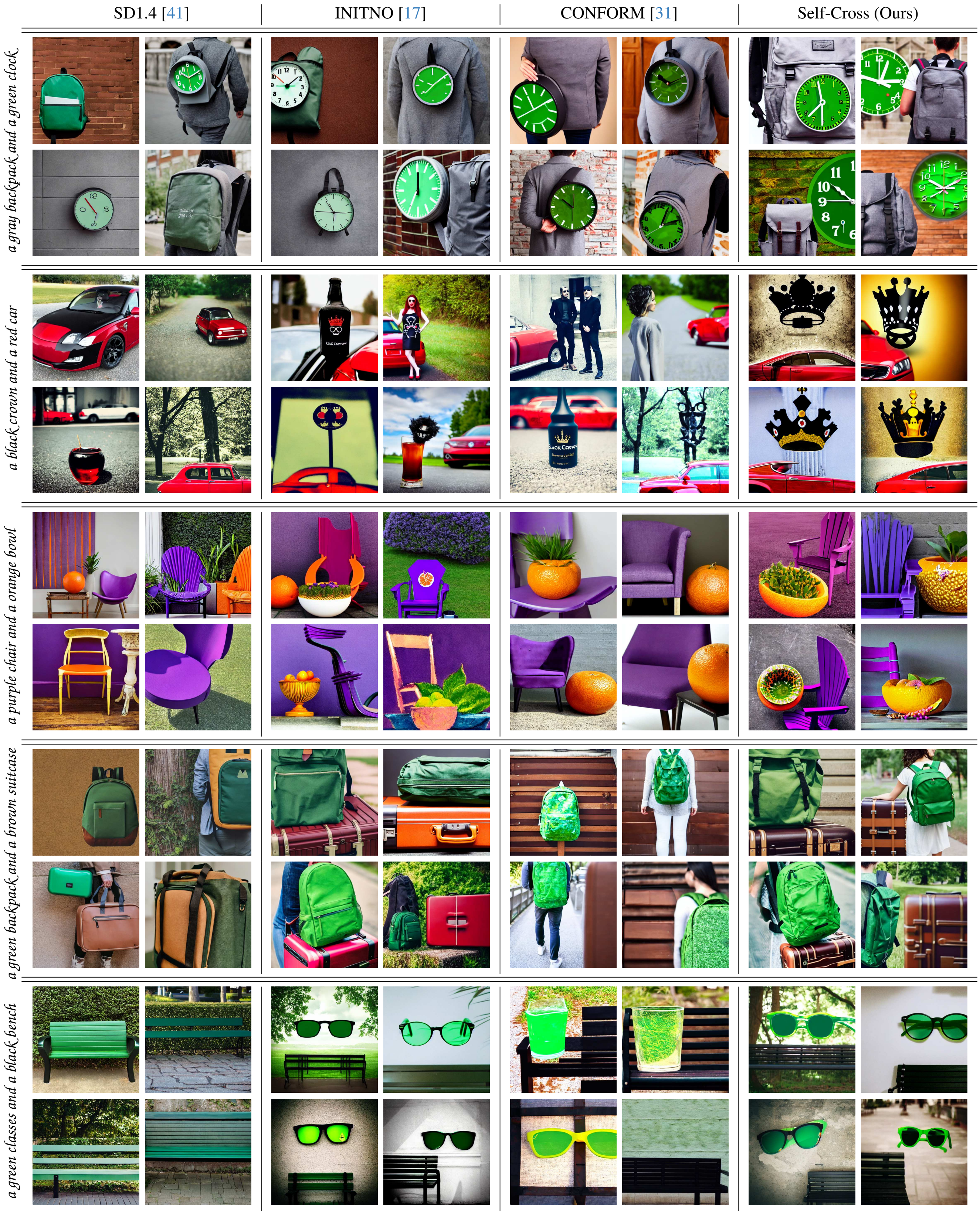

SD1.4 [41]
INITNO [17]
CONFORM [31]
Self-Cross (Ours)
a gray backpack and a green clock
a black crown and a red car
a purple chair and a orange bowl
a green backpack and a brown suitcase
a green classes and a black bench

Figure 14. More qualitative comparisons of Self-Cross (ours) to SD1.4 [41], INITNO [17], CONFORM [31]. For each prompt in the left column, we sample four seeds and show the results of different methods.

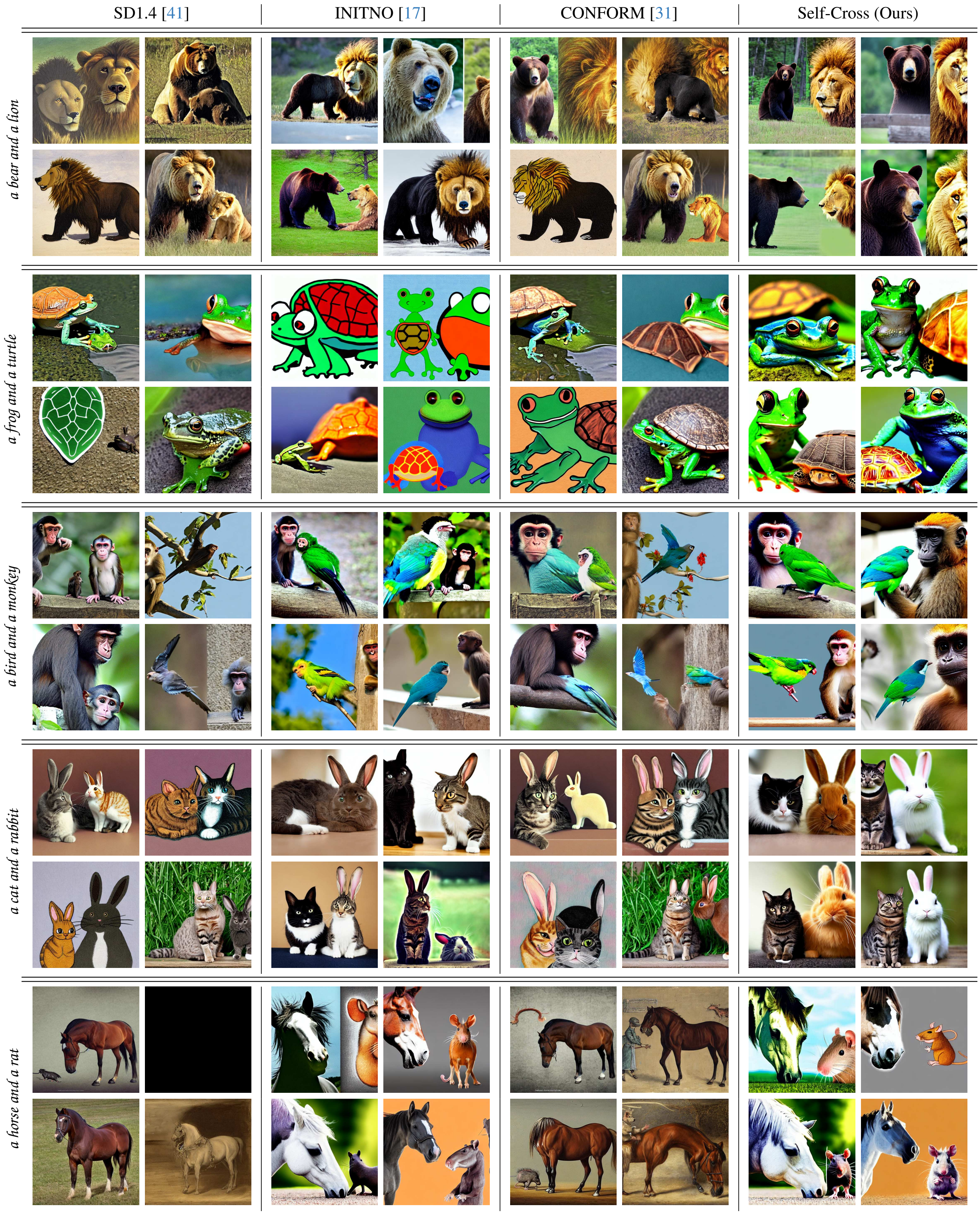


Figure 15. More qualitative comparisons of Self-Cross (ours) to SD1.4 [41], INITNO [17], CONFORM [31]. For each prompt in the left column, we sample four seeds and show the results of different methods.